Untersuchungen zur Diskriminanzanalyse mit hochdimensionalen Daten

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

genehmigt durch die Fakultät für Mathematik

der Otto-von-Guericke-Universität Magdeburg

von Dipl.-Math. Martin Grüning

geb. am 26.04.1969 in Wittenberg

Gutachter:	Prof. Dr. Ludwig A. Hothorn
	Prof. Dr. Norbert Gaffke
	Priv.-Doz. Dr. Siegfried Kropf
Eingereicht am:	04.07.2005
Verteidigung am:	01.12.2005

Danksagung

Mein Dank gilt an dieser Stelle besonders Herrn PD Dr. S. Kropf und Herrn Prof. Dr. N. Gaffke, die diese Arbeit betreut und nach Fertigstellung begutachtet haben. Die Anregungen, Hinweise und Ideen, die ich von ihnen bekam, haben diese Arbeit erst ermöglicht und sehr zu deren Gelingen beigetragen. Ebenso danke ich Herrn Prof. Dr. L. A. Hothorn für die Erstellung des Gutachtens sowie Herrn Prof. Dr. K. Deckelnick als dem Vorsitzenden der Promotionskommission.

Weiterhin danke ich meinen Kollegen vom Institut für Biometrie und Medizinische Informatik für alle Unterstützung; besonders Herrn Dr. F.-W. Röhl und Herrn Dr. S. Mulla-Osman für die Überlassung ihrer Arbeitsplatzrechner für Beispielberechnungen zu dieser Arbeit.

Den Professoren und Mitarbeitern des Instituts für Mathematische Stochastik danke ich für die Möglichkeit, im Rahmen des Oberseminars zu meinem Forschungsthema vortragen zu können. Verschiedene Hinweise und Anregungen in diesem Rahmen haben ebenfalls zur Entstehung der Arbeit beigetragen.

Für die Bereitstellung von Datensätzen danke ich Herrn Dr. M. Eszlinger von der Medizinischen Fakultät der Universität Leipzig (Datensatz „Knoten“), dem Human Genetics Center der University of Texas-Houston (Datensatz „Colon“) sowie der Forschungsgruppe Dr. P. D. Wentzell von der Abteilung Chemie an der Dalhousie University in Halifax, Nova Scotia, Kanada (Datensatz „Gasoil“).

Magdeburg, im Januar 2006 Martin Grüning

Inhaltsverzeichnis

Verwendete Abkürzungen und Symbole

$\mathbb{N}$	Raum der natürlichen Zahlen
$\mathbb{R}$	Raum der reellen Zahlen
$\mathbb{R}^p$	p-dimensionaler Produktraum $\mathbb{R} \times \ldots \times \mathbb{R}$
$\boldsymbol{x}', \boldsymbol{A}'$	transponierter Vektor bzw. transponierte Matrix
$\boldsymbol{I}_p$	p-dimensionale Einheitsmatrix
$\lvert\boldsymbol{A}\rvert$	Determinante der Matrix $\boldsymbol{A}$
$\mathbf{Diag}(\boldsymbol{A})$	aus Diagonalelementen der Matrix $\boldsymbol{A}$ gebildete Diagonalmatrix
$\boldsymbol{A}^{-\frac{1}{2}}$	Inverse $\boldsymbol{X}^{-1}$ der Dreiecksmatrix bei Cholesky-Zerlegung $\boldsymbol{A} = \boldsymbol{X'X}$ ($\boldsymbol{A}$ positiv definit)
$\boldsymbol{S}$	Stichprobenkovarianzmatrix, $\boldsymbol{S} = \frac{1}{n-J} \sum_{j=1}^{J} \sum_{\alpha=1}^{n_j} (\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})(\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})'$
$\mathcal{M}(p,q)$	Raum der reellen $p \times q$-Matrizen
$\mathcal{S}(p)$	Raum der reellen symmetrischen $p \times p$-Matrizen
$\mathrm{PSD}(p)$	Raum der reellen symmetrischen, positiv semidefiniten $p \times p$-Matrizen
$\mathrm{PD}(p)$	Raum der reellen symmetrischen, positiv definiten $p \times p$-Matrizen
$(M, \mathcal{A}, P)$	allgemeiner Wahrscheinlichkeitsraum
$\mathcal{B}^p$	borelsche σ-Algebra auf $\mathbb{R}^p$
u. i. v.	stochastisch unabhängig, identisch verteilt
$\xrightarrow{P}$	stochastische Konvergenz, Konvergenz in Wahrscheinlichkeit
$N(\beta, \sigma^2)$	univariate Normalverteilung
$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	p-dimensionale multivariate Normalverteilung
χ^2_m	χ^2-Verteilung
$W_p(\boldsymbol{\Sigma}, m)$	Wishart-Verteilung
$W_p^{-1}(\boldsymbol{\Psi}, m)$	inverse Wishart-Verteilung
$\square$	Ende eines Beweises
[ ]	Hinweis auf Literaturangaben

Kapitel 1

Einführung

Viele in der Natur oder der menschlichen Gesellschaft beobachtete Vorgänge und Erscheinungen sind in ihren Gesetzmäßigkeiten nicht oder nicht vollständig bekannt. Zu ihrer Beschreibung nutzt man daher oftmals statistische Modelle. Vorgänge und Erscheinungen werden darin durch Zufallsvariable modelliert, deren Verteilungen nicht oder nicht vollständig bekannt sind. Ein Hauptanwendungsgebiet der mathematischen Statistik ist sicher die Analyse von empirischen Daten zum Zweck einer möglichst wirklichkeitsnahen Modellanpassung als Grundlage für das Schlussfolgern verschiedener Aussagen. Durch multivariate statistische Analyse wird das Verhalten mehrerer Größen mit ihren Abhängigkeiten untereinander beschrieben, für das verwendete Modell wird dann entsprechend ein mehrdimensionaler Ereignis- bzw. Zustandsraum zugrundegelegt.

Wurden in den ursprünglichen Untersuchungen zur multivariaten Statistik zunächst nur einige wenige Merkmale betrachtet, so hat man heutzutage oft Situationen mit wesentlich höheren Merkmalszahlen. Ein Beispiel für Aufgabenstellungen mit solchen hochdimensionalen Beobachtungen ist die Analyse von Genexpressionsdaten, wie sie gegenwärtig bei verschiedenen Experimenten mit so genannten Microarrays vorkommen. Typischerweise beträgt dabei die Anzahl der beobachteten Merkmale, also die Dimension des Ereignis- bzw. Zustandsraumes, mehrere tausend bei einer Anzahl der beobachteten Fälle von unter 100. Dies führt zu erheblichen Problemen bei der Datenanalyse.

Bisher beziehen sich die meisten theoretischen Aussagen für solcherlei Probleme auf den Fall, dass die Anzahl der beobachteten Fälle größer als die Anzahl der Merkmale oder mindestens genauso groß ist. Da diese Voraussetzung im genannten Beispiel gewöhnlich nicht erfüllt ist, sind hier die herkömmlichen klassischen Verfahren nicht ohne weiteres anwendbar. Weitergehende theoretische Überlegungen scheinen zunächst notwendig.

Ein Anwendungsgebiet der statistischen Analyse ist die Problematik der Klassifikation. Neben allgemeineren theoretischen Untersuchungen soll diese in dieser Arbeit vorrangig behandelt werden. Die Untersuchungen hier beziehen sich hauptsächlich auf Modelle mit Normalverteilungsannahme.

Nachdem im ersten Teil in die Problematik der Klassifikation eingeführt wird, soll im zweiten Teil der Arbeit das Problem der Parameterschätzung bei hohen Dimensionen, insbesondere der Schätzung der Kovarianzmatrix bei multivariater Normalverteilung, im Vordergrund stehen. Im dritten Teil werden schließlich Ausführungen zur Analyse des Klassifikationsfehlers gemacht. Es werden Untersuchungen zur optimalen Wahl des Ridge-Parameters angestellt. Zum Abschluss wird die Klassifikation von Elementen mittels verschiedener Verfahren an einigen Beispielen demonstriert.

Teil I

Das Problem der Klassifikation in der multivariaten Statistik

In den folgenden Kapiteln wird kurz in grundlegende Prinzipien und Methoden der multivariaten statistischen Analyse eingeführt. Speziell das Problem der Klassifikation wird dann beschrieben und formalisiert.

Für viele Anwendungen der multivariaten Statistik spielt die Normalverteilung $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mit mehrdimensionalem Erwartungswertvektor $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$ eine große Rolle. Für die Analyse von $\boldsymbol{\Sigma}$ als einer symmetrischen Matrix stellt die Spektralzerlegung und die Eigenwert- und Eigenvektoranalyse ein wichtiges Hilfsmittel dar.

Viele Probleme der Statistik lassen sich als statistische Entscheidungsprobleme modellieren. Dies gilt auch für das Problem der Klassifikation.

Kapitel 2

Grundlagen der multivariaten Statistik

2.1 Die multivariate Normalverteilung

Von großer Bedeutung in der multivariaten Statistik ist eine mehrdimensionale Erweiterung der eindimensionalen, auf der Menge der reellen Zahlen definierten, univariaten Normalverteilung $N(\beta, \sigma^2)$ — die mehrdimensionale multivariate Normalverteilung. Für die Parameter β und σ^2 hat man entsprechende mehrdimensionale Erweiterungen.

Die Menge aller positiv definiten reellen $p \times p$-Matrizen wird mit $\mathrm{PD}(p)$ abgekürzt. Die (reguläre) multivariate Normalverteilung $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mit den Parametern $\boldsymbol{\mu} \in \mathbb{R}^p$ und $\boldsymbol{\Sigma} \in \mathrm{PD}(p)$ ist eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^p, \mathcal{B}^p)$ mit der Lebesgue-Dichte

$$(2\pi)^{-\frac{1}{2}p} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right). \tag{2.1}$$

Dabei ist $\boldsymbol{\mu}$ der Erwartungswert(vektor), $\boldsymbol{\Sigma}$ ist die Kovarianzmatrix. Typischerweise hat man hier die Situation, dass die Parameter unbekannt sind. Sind n Beobachtungen von stochastisch unabhängigen, identisch multivariat normalverteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ gegeben, wobei $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ unbekannt seien, hat man das statistische Modell

$$\left(\mathbb{R}^{n \times p}, \mathcal{B}^{n \times p}, \left(\bigotimes_{\alpha=1}^{n} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right)_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathrm{PD}(p)} \right). \tag{2.2}$$

Die symmetrische, positiv definite Kovarianzmatrix $\boldsymbol{\Sigma}$ ist hierbei sicher von besonderem Interesse. Ihre Unkenntnis kann — vor allem bei hohen Dimensionen — zu großen Problemen führen. Symmetrische Matrizen stellen daher einen besonderen Schwerpunkt der weiteren Analysen in dieser Arbeit dar.

2.2 Hauptkomponentenanalyse

2.2.1 Grundlagen

Eine besondere Eigenschaft von symmetrischen Matrizen ist die Möglichkeit ihrer Spektralzerlegung:

Satz 2.1 (Spektralzerlegungssatz, [37], Theorem A.6.4) *Jede reelle symmetrische Matrix $\boldsymbol{A}$ $(p \times p)$ lässt sich darstellen als*

$$\boldsymbol{A} = \boldsymbol{\Gamma\Lambda\Gamma'}, \tag{2.3}$$

wobei $\boldsymbol{\Gamma}$ eine orthogonale Matrix ist und $\boldsymbol{\Lambda}$ Diagonalform hat. Die Hauptdiagonalelemente von $\boldsymbol{\Lambda}$ sind die Eigenwerte von $\boldsymbol{A}$, und die Spaltenvektoren von $\boldsymbol{\Gamma}$ sind die normierten Eigenvektoren.

Bemerkung 2.1 *1. Aus den Eigenschaften orthogonaler Matrizen folgt:* $\boldsymbol{\Lambda = \Gamma' A \Gamma}$

2. *Es gilt: Die Eigenwerte einer symmetrischen Matrix $\boldsymbol{A}$ sind reell. Sie sind alle nichtnegativ genau dann, wenn $\boldsymbol{A}$ positiv semidefinit ist, und alle positiv genau dann, wenn $\boldsymbol{A}$ positiv definit ist. Beweis: z. B. in [47]*

3. *Es gilt auch folgende Umkehrung: Ist $\boldsymbol{\Gamma}$ Orthogonalmatrix und ist $\boldsymbol{\Lambda}$ Diagonalmatrix der gleichen Ordnung, so ist $\boldsymbol{A := \Gamma\Lambda\Gamma'}$ symmetrisch, die Spaltenvektoren von $\boldsymbol{\Gamma}$ sind die Eigenvektoren von $\boldsymbol{A}$, und die Diagonalelemente von $\boldsymbol{\Lambda}$ sind die zugehörigen Eigenwerte. Es lässt sich leicht sehen, dass dann für den i-ten Spaltenvektor $\boldsymbol{\gamma}_{(i)}$ von $\boldsymbol{\Gamma}$ gilt:*

$$\boldsymbol{A\gamma}_{(i)} = \boldsymbol{(\Gamma\Lambda\Gamma')\gamma}_{(i)} = \lambda_i \boldsymbol{\gamma}_{(i)}. \tag{2.4}$$

Damit ist der Spaltenvektor $\boldsymbol{\gamma}_{(i)}$ ein Eigenvektor zum Eigenwert λ_i von $\boldsymbol{A}$.

4. *Ist $\boldsymbol{A}$ eine symmetrische Matrix, so folgt mit der Spektralzerlegung (2.3):*

$$\boldsymbol{A}^k = \boldsymbol{\Gamma\Lambda}^k \boldsymbol{\Gamma'} \; \forall k \in \mathbb{N}. \tag{2.5}$$

Das bedeutet: Für alle $k \in \mathbb{N}$ ist $\boldsymbol{A}^k$ ebenfalls symmetrisch. Sind λ_i $(i = 1, \ldots, p)$ die Eigenwerte von $\boldsymbol{A}$, so sind λ_i^k $(i = 1, \ldots, p)$ die Eigenwerte von $\boldsymbol{A}^k$.

Die Darstellung (2.3) wird als Spektralzerlegung bezeichnet. Da die Kovarianzmatrix einer multivariaten Normalverteilung symmetrisch ist, lässt sie sich wie in (2.3) darstellen. Dies begründet die folgende Definition von Hauptkomponenten.

Definition 2.1 (nach [37]) *Sei* $\boldsymbol{X} : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^p, \mathcal{B}^p)$ *Zufallsvektor mit zugehöriger Verteilung* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Die Matrix* $\boldsymbol{\Gamma}$ *sei nach Satz 2.1 eine Orthogonalmatrix, so dass* $\boldsymbol{\Gamma'\Sigma\Gamma} = \boldsymbol{\Lambda}$ *eine Diagonalmatrix mit* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ *ist. Die Hauptachsentransformation ist ein Zufallsvektor* $\boldsymbol{Y} : (\mathbb{R}^p, \mathcal{B}^p, P^{\boldsymbol{X}}) \to (\mathbb{R}^p, \mathcal{B}^p)$ *mit*

$$\boldsymbol{Y} = \boldsymbol{\Gamma'}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{2.6}$$

Als i-te Hauptkomponente von $\boldsymbol{X}$ *wird das i-te Element des Vektors* $\boldsymbol{Y} = \boldsymbol{\Gamma'}(\boldsymbol{X} - \boldsymbol{\mu})$, *also* $Y_i = \boldsymbol{\gamma}'_{(i)}(\boldsymbol{X} - \boldsymbol{\mu})$, *bezeichnet. Der Vektor* $\boldsymbol{\gamma}_{(i)}$ *ist hierbei der i-te Spaltenvektor von* $\boldsymbol{\Gamma}$.

Mit obigen Bezeichnungen gelten die folgenden Eigenschaften:

1. $\mathrm{E}(Y_i) = 0$
2. $\mathrm{Var}(Y_i) = \lambda_i$
3. $\mathrm{Cov}(Y_i, Y_j) = 0$, falls $i \neq j$
4. $\mathrm{Var}(Y_1) \geq \mathrm{Var}(Y_2) \geq \ldots \geq \mathrm{Var}(Y_p) \geq 0$
5. $\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \mathrm{tr}\boldsymbol{\Sigma}$
6. $\prod_{i=1}^{p} \mathrm{Var}(Y_i) = |\boldsymbol{\Lambda}| = |\boldsymbol{\Sigma}|$

Diese Eigenschaften lassen sich leicht zeigen; es sei dazu wiederum auf [37] verwiesen.

2.2.2 Geometrische Interpretation

Die Hauptachsentransformation lässt sich sehr gut durch ihre geometrische Interpretation veranschaulichen. Dazu soll ein multivariat normalverteilter Zufallsvektor $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{\Sigma}$ sei regulär, d. h. positiv definit) betrachtet werden. Die Bezeichnungen und Definitionen werden hier wie im vorigen Abschnitt verwendet. Durch Subtraktion mit $\boldsymbol{\mu}$ und Multiplikation mit $\boldsymbol{\Gamma}'$ entsteht der neue Zufallsvektor $\boldsymbol{Y} = \boldsymbol{\Gamma'}(\boldsymbol{X} - \boldsymbol{\mu})$ mit $\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{\Gamma'\Sigma\Gamma}) = N(\boldsymbol{0}, \boldsymbol{\Lambda})$. Die Verteilung von $\boldsymbol{X}$ hat die Dichte

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{1}{2}p} |\boldsymbol{\Sigma}| \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{2.7}$$

Betrachtet man die Menge aller Punkte $\boldsymbol{x} \in \mathbb{R}^p$, die die gleiche Dichte haben, d. h., für die gilt $f(\boldsymbol{x}) = \text{konst.}$, so erhält man:

$$(2\pi)^{-\frac{1}{2}p} |\boldsymbol{\Sigma}| \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) = \text{konst.} \tag{2.8}$$

$$\Leftrightarrow \quad (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = \text{konst.} = c > 0 \tag{2.9}$$

$$\Leftrightarrow \quad (\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}')^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c \tag{2.10}$$

$$\Leftrightarrow \quad (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'(\boldsymbol{x}-\boldsymbol{\mu}) = c \tag{2.11}$$

$$\Leftrightarrow \quad \boldsymbol{y}'\boldsymbol{\Lambda}^{-1}\boldsymbol{y} = c \tag{2.12}$$

$$\Leftrightarrow \quad \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i} = c \tag{2.13}$$

$$\Leftrightarrow \quad \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i c} = 1 \tag{2.14}$$

Die letzte Gleichung (2.14) beschreibt dabei ein Ellipsoid im p-dimensionalen Raum $\mathbb{R}^p$ mit Mittelpunkt im Koordinatenursprung. Die Hauptachsen verlaufen entlang der Koordinatenachsen, die Länge der i-ten Halbachse beträgt $\sqrt{\lambda_i c}$. Die allgemeine Gleichung für ein Ellipsoid im $\mathbb{R}^p$ ist durch (2.9) gegeben (mit positiv definiter Matrix $\boldsymbol{\Sigma}$). Der Erwartungswertvektor $\boldsymbol{\mu}$ bildet den Mittelpunkt des Ellipsoids, und es gilt (nach [37, Theorem A.10.2]):

1. Der i-te Spaltenvektor von $\boldsymbol{\Gamma}$, $\boldsymbol{\gamma}_{(i)}$ (d. h. der Eigenvektor zum Eigenwert λ_i) ist der Richtungsvektor der i-ten Hauptachse des Ellipsoids.

2. Die Länge der i-ten Halbachse beträgt $\sqrt{\lambda_i c}$.

Das bedeutet: Die Abbildung $\boldsymbol{Y}$ ist eine Koordinatentransformation im Vektorraum $\mathbb{R}^p$ derart, dass der Koordinatenursprung in den Mittelpunkt und die Koordinatenachsen in Richtung der Hauptachsen des Ellipsoids der Punkte konstanter Dichte gelegt werden. Die p miteinander korrellierten Komponenten des Zufallsvektors werden dabei in p unabhängige Komponenten überführt. (Aus Unkorreliertheit folgt hier wegen Normalverteilungsannahme stochastische Unabhängigkeit, vgl. [3]).

2.3 Statistische Entscheidungstheorie (nach [3] [36] [40])

Viele Probleme der schließenden Statistik lassen sich als statistische Entscheidungsprobleme auffassen. Diese Betrachtungsweise erlaubt die Entwicklung von Optimalitätskonzepten und die Gütebewertung von statistischen Methoden.

Statistische Entscheidungsprobleme lassen sich wie folgt beschreiben: Gegeben sei ein statistisches Modell $(M, \mathcal{A}, \mathcal{P})$ und ein Messraum $(D, \mathcal{D})$. Der Raum D ist hier der Raum von möglichen Entscheidungen. Eine *Entscheidungsfunktion* (auch als Entscheidungsregel oder -verfahren bezeichnet) ist dann eine Zufallsvariable $\delta : (M, \mathcal{A}) \to (D, \mathcal{D})$.

Typischerweise ist die Familie von Wahrscheinlichkeitsverteilungen $\mathcal{P}$ parametrisch, d. h. von der Form $(P_\vartheta)_{\vartheta\in\Theta}$ mit einem gegebenen Parameterraum Θ.

Eine Funktion $L : \Theta \times D \to [0, \infty)$ heißt *Verlustfunktion*, wenn L für jedes $\vartheta \in \Theta$ $\mathcal{D}$-messbar ist. Die Zahl $L(\vartheta, d)$ bezeichnet den *Verlust* einer Entscheidung $d \in D$ bei Vorliegen von $\vartheta \in \Theta$.

Zu einer Entscheidungsfunktion δ ist die *Risikofunktion* $R(\cdot, \delta) : \Theta \to \mathbb{R}$ gegeben durch

$$R(\vartheta, \delta) = E_\vartheta(L(\vartheta, \delta)), \tag{2.15}$$

und $R(\vartheta, \delta)$ bezeichnet jeweils das *Risiko* der Entscheidungsfunktion δ bezüglich ϑ. Dieses Risiko kann nun als Bewertungskriterium für eine Entscheidungsfunktion verwendet werden. Es wird angestrebt, dieses Risiko möglichst klein zu halten.

Das Risiko 2.15 ist jedoch abhängig von dem Parameter ϑ. Zur Bewertung einer Entscheidungsfunktion oder zum Vergleich mehrerer Entscheidungsfunktionen interessiert nun nicht so sehr das Risiko bezüglich eines einzelnen Parameterwertes ϑ, zumal dieser ja gewöhnlich nicht bekannt ist, sondern vielmehr globale Kriterien für den gesamten Parameterraum Θ.

Für zwei Entscheidungsfunktionen δ_1 und δ_2 wird gesagt, die Entscheidungsfunktion δ_1 ist *besser als* δ_2, falls

$$R(\vartheta, \delta_1) \leq R(\vartheta, \delta_2) \ \forall \vartheta \in \Theta \tag{2.16}$$

gilt. In einer (Teil-)Menge Δ von Entscheidungsfunktionen heißt eine Entscheidungsfunktion δ^* *gleichmäßig beste* Entscheidungsfunktion bezüglich Δ, falls

$$R(\vartheta, \delta^*) \leq R(\vartheta, \delta) \ \forall \vartheta \in \Theta, \delta \in \Delta \tag{2.17}$$

gilt. In diesem Zusammenhang wird auch der Begriff der Zulässigkeit gebraucht. Eine Entscheidungsfunktion $\delta \in \Delta$ heißt *zulässig* in der Menge der Entscheidungsfunktionen Δ, falls es keine Entscheidungsfunktion $\delta^* \in \Delta$ gibt, die gleichmäßig besser als δ ist; andernfalls heißt δ *nicht zulässig*.

Neben dieser gleichmäßigen Optimalität gibt es noch zwei weitere, oft verwendete Optimalitätsbegriffe: die Minimax- und die Bayes-Optimalität. Eine Entscheidungsfunktion $\delta^* \in \Delta$ heißt *Minimax-optimal* (oder Minimax-Entscheidungsfunktion) bezüglich Δ, falls

$$\sup_{\vartheta \in \Theta} R(\vartheta, \delta^*) \leq \sup_{\vartheta \in \Theta} R(\vartheta, \delta) \ \forall \delta \in \Delta \tag{2.18}$$

gilt. Eine Minimax-Entscheidungsfunktion ist also eine solche, die das maximale Risiko minimiert.

Für den Begriff der Bayes-Optimalität benötigt man eine „a-priori"-Wahrscheinlichkeitsverteilung Q auf dem Parameterraum Θ. Der Parameterraum wird hierbei als Messraum $(\Theta, \mathcal{T})$ aufgefasst, und die Funktion $R(\cdot, \delta) : \Theta \to \mathbb{R}$ wird als $\mathcal{T}$-messbar vorausgesetzt. Eine Entscheidungsfunktion $\delta^* \in \Delta$ heißt *Bayes-optimal* (oder Bayes-Entscheidungsfunktion) bezüglich Δ, wenn

$$\int_\Theta R(\vartheta, \delta^*) dQ(\vartheta) \leq \int_\Theta R(\vartheta, \delta) dQ(\vartheta) \ \forall \delta \in \Delta \tag{2.19}$$

gilt. Das mittlere Risiko

$$r(Q,\delta) = E_Q R(\vartheta,\delta) = \int_\Theta R(\vartheta,\delta)Q(\vartheta)d\vartheta, \tag{2.20}$$

das hierbei durch δ^* minimiert wird, wird als *Bayes'sches Risiko* von δ bezüglich der a-priori-Verteilung Q bezeichnet. Es konnte nun gezeigt werden, dass solche Bayes-Entscheidungsfunktionen bezüglich irgend einer a-priori-Verteilung Q auf $(\Theta,\mathcal{T})$ auch zulässig sind (z. B. [3]).

2.4 Schätzung von Parametern

Eines der Hauptanwendungsgebiete der statistischen Entscheidungstheorie ist wohl die Schätztheorie der mathematischen Statistik. Ist die „wahre" Verteilung einer Zufallsvariablen nicht oder nur teilweise bekannt, versucht man, aus den erhobenen Daten Rückschlüsse über die vorliegende Verteilung zu ziehen. Durch Schätzungen sollen Parameter einer Wahrscheinlichkeitsverteilung möglichst wirklichkeitsnah bestimmt werden.

Definition 2.2 *Gegeben sei ein statistisches Modell $(M,\mathcal{A},(P_\vartheta)_{\vartheta\in\Theta})$, wobei $(P_\vartheta)_{\vartheta\in\Theta}$ eine Familie von Wahrscheinlichkeitsverteilungen mit Parameterraum Θ ist, und eine nichtleere Menge N. (N ist hier typischerweise ein Vektorraum über $\mathbb{R}$.) Eine Abbildung $T: M\to N$ heißt Statistik. Ein Schätzer für den Parameter ϑ ist eine Statistik $\delta: M\to N$ mit $\Theta\subset N$.*

Solche Schätzer (oder *Schätzfunktionen*) δ lassen sich als spezielle Entscheidungsfunktionen auffassen.

Unter allen solchen möglichen Abbildungen δ ist nun eine nach bestimmten Kriterien besonders „gute" gesucht. Gibt es für $(P_\vartheta)_{\vartheta\in\Theta}$ Dichtefunktionen $(f_\vartheta)_{\vartheta\in\Theta}$, so sind Maximum-Likelihood-Schätzer für ϑ solche, für die die zugehörige Likelihood-Funktion $L(x,\vartheta) = f_\vartheta(x)$ bei gegebenem $x\in M$ den maximalen Wert annimmt.

Definition 2.3 (Maximum-Likelihood-Schätzung) *Sei $(M,\mathcal{A},(P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, μ sei ein Maß auf $\mathcal{A}$, f_ϑ sei μ-Dichte von P_ϑ $\forall\vartheta$. Eine Maximum-Likelihood-Schätzung (ML-Schätzung) für ϑ zu einem gegebenem $x\in M$ ist ein $\hat\vartheta\in\Theta$ mit $f_{\hat\vartheta(x)}(x) = \max\limits_{\vartheta\in\Theta} f_\vartheta(x)$. Sind $\emptyset\neq M_0\subset M, M_0\in\mathcal{A}$ und $\hat\vartheta: M_0\to\Theta$, so dass $\hat\vartheta(x)$ für jedes $x\in M_0$ eine Maximum-Likelihood-Schätzung für ϑ ist, dann ist $\hat\vartheta$ ein auf M_0 definierter Maximum-Likelihood-Schätzer (ML-Schätzer) für ϑ. Gilt $P_\vartheta(M_0)=1$ $\forall\vartheta\in\Theta$, dann ist $\hat\vartheta$ ein (fast überall definierter) Maximum-Likelihood-Schätzer für ϑ.*

Ist das durch Ausdruck (2.2) gegebene Modell mit unbekannten Paramtern zugrundegelegt, hat man $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ geeignet zu schätzen. Für den Fall, dass $n > p$

ist, sind $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{\alpha=1}^{n} \boldsymbol{x}_\alpha$ bzw. $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ Maximum-Likelihood-Schätzer für $\boldsymbol{\mu}$ bzw. $\boldsymbol{\Sigma}$ (siehe z. B. [3], [37]). Oft verwendet man statt $\hat{\boldsymbol{\Sigma}}$ den erwartungstreuen Schätzer

$$\boldsymbol{S} := \frac{1}{n-1}\sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})' \tag{2.21}$$

für $\boldsymbol{\Sigma}$ (d. h. $\boldsymbol{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\boldsymbol{S} = \boldsymbol{\Sigma}$). $\boldsymbol{S}$ wird auch als *Stichprobenkovarianzmatrix* bezeichnet.

Im Mehrstichprobenmodell mit J Stichproben hat man Beobachtungen $\boldsymbol{x}_1^{(j)}, \ldots, \boldsymbol{x}_{n_j}^{(j)}$ zu Zufallsvektoren $\boldsymbol{X}_1^{(j)}, \ldots, \boldsymbol{X}_{n_j}^{(j)}$ $(j = 1, \ldots, J)$ mit J verschiedenen zugrundegelegten Verteilungen in einem gemeinsamen Beobachtungsraum. Sind dies jeweils multivariate Normalverteilungen mit verschiedenen Erwartungswerten und gemeinsamer Kovarianzmatrix, wobei die Parameter alle unbekannt seien, so hat man folgendes Modell (mit $n = \sum_{j=1}^{J} n_j$):

$$\left(\mathbb{R}^{n\times p}, \mathcal{B}^{n\times p}, \left(\bigotimes_{j=1}^{J}\bigotimes_{\alpha=1}^{n_j} N_p(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma})\right)_{(\boldsymbol{\mu}^{(1)},\ldots,\boldsymbol{\mu}^{(J)},\boldsymbol{\Sigma})\in\mathbb{R}^{Jp}\times \mathrm{PD}(p)}\right). \tag{2.22}$$

Die naheliegenden Schätzer für $\boldsymbol{\mu}^{(j)}$ $(j = 1, \ldots, J)$ sind dann jeweils $\bar{\boldsymbol{x}}^{(j)} = \frac{1}{n_j}\sum_{\alpha=1}^{n_j} \boldsymbol{x}_\alpha^{(j)}$. Ein entsprechender erwartungstreuer Schätzer für $\boldsymbol{\Sigma}$ ist die gemeinsame Stichprobenkovarianzmatrix $\boldsymbol{S}$, die dann durch

$$\boldsymbol{S} = \frac{1}{n-J}\sum_{j=1}^{J}\sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})(\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})' \tag{2.23}$$

gegeben ist. Den Einstichprobenfall erhält man hier als Spezialfall mit $J = 1$. In dieser Arbeit wird ansonsten vorrangig der Fall $J = 2$ betrachtet.

Das Mehrstichprobenmodell ist Ausgangspunkt für die Beschreibung des Klassifikationsproblems bei unbekannten Parametern. Mit einer kleinen Modifikation führt es zum Modell, das dem Klassifikationsproblem zugrundeliegt (siehe nächstes Kapitel).

Bemerkung 2.2 *Zur Bezeichnung: Mit n wird die Anzahl der Beobachtungen bezeichnet. Mit $\hat{\vartheta}$ werden hier ausschließlich Maximum-Likelihood-Schätzer für ϑ bezeichnet. D. h., die Statistik $\frac{1}{n}\sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ (im Einstichprobenfall) wird hier nur dann mit $\hat{\boldsymbol{\Sigma}}$ bezeichnet, wenn sie Maximum-Likelihood-Schätzer für $\boldsymbol{\Sigma}$, d. h. regulär ist (vgl. Kapitel 5). $\boldsymbol{S}$ ohne Zusatz bezeichnet die Stichprobenkovarianzmatrix wie oben definiert. Mit $\tilde{\boldsymbol{S}}$ wird allgemein eine Statistik der Form $k\sum_{j=1}^{J}\sum_{\alpha=1}^{n_j} (\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})(\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})', k > 0$ oder ggf. ein Schätzer für $\boldsymbol{\Sigma}$ anderer Art bezeichnet.*

Kapitel 3

Modellierung des Klassifikationsproblems

3.1 Das Problem der Klassifikation als statistisches Entscheidungsproblem

Es soll jetzt auch das Klassifikationsproblem als ein spezielles Entscheidungsproblem beschrieben werden. Ausgangspunkt ist hier zunächst immer eine Definition von endlich vielen Klassen. Aufgrund einer Beobachtung einer Zufallsvariablen ist dann eine Zuordnung zu einer der Klassen zu treffen. Jede der Klassen ist dabei durch eine spezielle Verteilung gekennzeichnet. Sind diese Verteilungen bekannt, so lässt sich das Klassifikationsmodell wie folgt modellieren. Man hat dann ein statistisches Modell

$$(M, \mathcal{A}, \{P_1, \ldots, P_J\}), \tag{3.1}$$

wobei der endliche Parameterraum $\{1, \ldots, J\}$ hier der Entscheidungsraum ist. Aufgrund der Beobachtung $x \in M$ ist bezüglich des Parameters $j \in \{1, \ldots, J\}$ durch eine Funktion $\delta : M \to \{1, \ldots, J\}$ eine geeignete Entscheidung zu treffen; der Parameter entspricht dabei jeweils der zugehörigen Klasse. Gewissermaßen hat man hier ein spezielles Schätzproblem, nämlich das Problem der Schätzung des Parameters $j \in \{1, \ldots, J\}$.

Solch eine Klassifikations-Entscheidung wird gewöhnlich mittels einer Statistik $f : M \to N$ getroffen. Solch eine Statistik f, die eindeutig eine Entscheidung für einen der Parameter zulässt, heißt *Diskriminanzfunktion*. Werden jeweils mehrere Merkmale beobachtet, d. h. ist der gemeinsame Grundraum der Beobachtungen $(M, \mathcal{A})$ mehrdimensional, kommen hier wiederum multivariate Verfahren zum Einsatz.

Ziel ist es nun, für das beschriebene Problem möglichst „geeignete“ Entscheidungsfunktionen zu finden. Als Bewertungskriterium dafür dient wieder das mittels einer Verlustfunktion berechnete Risiko. Eine Verlustfunktion $L : \{1, \ldots, J\} \times \{1, \ldots, J\} \to [0, \infty)$ lässt sich hier mit Hilfe einer nichtnegativen Kostenfunktion

C wie folgt definieren:

$$L(j,i) = \begin{cases} C(j,i), & \text{falls } j \neq i, \\ 0, & \text{falls } j = i. \end{cases} \tag{3.2}$$

Hierbei ist j der wahre und i der geschätzte Parameter. Das Risiko einer Entscheidungsfunktion δ ist dann gegeben durch

$$R(j,\delta) = \mathrm{E}_{P_j} L(j,\delta(x)) = \sum_{i=1}^{J} C(j,i) p_{ji}, \tag{3.3}$$

wobei p_{ji} jeweils durch

$$p_{ji} = P_j(\{x \in M : \delta(x) = i\}) \tag{3.4}$$

gegeben ist. Dieses Risiko soll für jedes j möglichst klein gehalten werden.

Ist außerdem noch eine a-priori-Verteilung Q auf $\{1,\ldots,J\}$ bekannt, so ist das Bayes-Risiko gegeben durch

$$r(Q,\delta) = \mathrm{E}_Q R(j,\delta) = \sum_{j=1}^{J} R(j,\delta) q_j, \tag{3.5}$$

wobei $q_j := Q(\{j\})$ die Einzelwahrscheinlichkeiten für die einzelnen Klassen sind. Ziel ist dann, dieses bayessche Risiko, wenn möglich, zu minimieren, oder aber zumindest möglichst klein zu halten.

3.2 Lineare Diskriminanzanalyse

Wir betrachten jetzt den oft üblichen Fall, dass der Messraum $(\mathbb{R}^p, \mathcal{B}^p)$ zugrundegelegt ist. Oft hat man hierbei die Situation, dass es zwei Klassen gibt (d. h. $J = 2$). Haben in diesem Fall die Wahrscheinlichkeitsverteilungen P_j $(j = 1,2)$ jeweils p-dimensionale Lebesgue-Dichten p_j $(j = 1,2)$, so lässt sich die Entscheidungsregel angeben, die für gegebene a-priori-Wahrscheinlichkeiten q_1 und q_2 das Risiko (3.5) minimiert:

Satz 3.1 ([3], Theorem 6.3.1) *Gegeben sei das statistische Modell*

$$(\mathbb{R}^p, \mathcal{B}^p, (P_j)_{j\in\{1,2\}}), \tag{3.6}$$

die Verteilungen P_j haben p-dimensionale Lebesgue-Dichten p_j $(j = 1,2)$. Ist Q a-priori-Verteilung auf $\{1,2\}$ mit $q_j = Q(\{j\})$ $(j = 1,2)$ und C eine nichtnegative Kostenfunktion auf $\{1,2\}$, so wird das Risiko (3.5) minimiert, wenn die Entscheidung nach folgender Bedingung getroffen wird:

$$\delta(\boldsymbol{x}) = \begin{cases} 1, & \textit{falls } \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} \geq \frac{q_2 C(2,1)}{q_1 C(1,2)} \\ 2, & \textit{falls } \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} < \frac{q_2 C(2,1)}{q_1 C(1,2)} \end{cases} \tag{3.7}$$

Geht man hierbei davon aus, dass die Beobachtungen jeweils multivariat normalverteilt sind, dann gilt im Modell (3.6) jeweils $P_j = N_p(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)})$ $(j = 1, 2)$. Oft betrachtet man hierbei die Situation, dass die Varianzen und Kovarianzen für beide Klassen gleich sind (d. h. $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \boldsymbol{\Sigma}$). (Die Kovarianzmatrix $\boldsymbol{\Sigma}$ sei hier wieder als regulär, d. h. positiv definit, angenommen.) Die Verteilungen P_1 und P_2 unterscheiden sich dann nur durch ihre Erwartungswerte $\boldsymbol{\mu}^{(1)}$ und $\boldsymbol{\mu}^{(2)}$. Die Dichten sind dann von der Form

$$p_j(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{1}{2}p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\Big[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^{(j)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(j)})\Big] \ (j = 1, 2). \tag{3.8}$$

Als eine Folgerung aus Satz 3.1 lässt sich für diese Situation wieder die optimale Entscheidungsregel angeben:

Satz 3.2 ([3], Theorem 6.4.1) *Gegeben sei das statistische Modell (3.6), die Verteilungen P_j haben jeweils die durch (3.8) gegebenen Lebesgue-Dichten p_j $(j = 1, 2)$. Ist Q a-priori-Verteilung auf $\{1, 2\}$ mit $q_j = Q(\{j\})$ $(j = 1, 2)$ und C eine nichtnegative Kostenfunktion auf $\{1, 2\}$, so wird das Risiko (3.5) minimiert, wenn die Entscheidung nach folgender Bedingung getroffen wird:*

$$\delta(\boldsymbol{x}) = \begin{cases} 1, & \textit{falls } \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) - \frac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \geq \ln(k) \\ 2, & \textit{falls } \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) - \frac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) < \ln(k), \end{cases} \tag{3.9}$$

wobei k durch

$$k = \frac{q_2 C(2, 1)}{q_1 C(1, 2)} \tag{3.10}$$

gegeben ist.

In der in (3.9) angegebenen Bedingung steht auf der linken Seite die wohlbekannte Diskriminanzfunktion der linearen Diskriminanzanalyse:

$$f(\boldsymbol{x}) = \Big(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})\Big)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}). \tag{3.11}$$

Gilt hierbei $q_1 = q_2 = \frac{1}{2}$ und $C(1, 2) = C(2, 1)$, so ist $\ln(k) = 0$. Dann wird für Klasse 1 entschieden, falls $f(\boldsymbol{x}) \geq 0$ gilt, und für Klasse 2, falls $f(\boldsymbol{x}) < 0$ gilt. (Der neutrale Fall $f(\boldsymbol{x}) = 0$ stellt hier eine Menge vom Maß Null dar. Oft wird diese Nullmenge auch formal der Klasse 2 zugeordnet; dies stellt keine wesentliche Änderung dar.)

Die Gleichung $f(\boldsymbol{x}) = 0$ teilt hier den gesamten Raum der Beobachtungen $\mathbb{R}^p$ in zwei Regionen der Zuordnung zu jeweils einer der Klassen 1 oder 2. Diese Gleichung ist eine lineare Gleichung von $\boldsymbol{x}$, daher spricht man hier auch von *linearer Diskriminanzanalyse*.

3.3 Klassifikation bei unbekannten Parametern

Gewöhnlich hat man aber die Situation, dass die Parameter der zugrundeliegenden Verteilungen unbekannt sind. Hat man dann zu jeder Klasse eine gewisse Anzahl von Beobachtungen, von denen die Klassenzugehörigkeit bekannt ist („Lernstichproben"), so lässt sich daraus die Information über die Verteilung bzw. über gewisse Verteilungsparameter ableiten. Eine Schätzung dieser Parameter der Verteilungen der einzelnen Klassen ist hierbei aber nicht vordergründig von Interesse; gesucht ist letztendlich wieder eine Entscheidung für den zu einer weiteren Zufallsvariablen gehörigen Klassenparameter.

Allgemein lässt sich dieses Problem wie folgt als Entscheidungsproblem beschreiben: Eine Beobachtung $x \in M$ ist die Realisierung einer vom jeweiligen Klassenparameter $j \in \{1, \dots, J\}$ abhängigen Zufallsvariablen $X^{(j)}$, deren Verteilung P_j, zumindest teilweise, unbekannt ist. Zu jedem $j = 1, \dots, J$ seien nun jeweils n_j Beobachtungen $x_1^{(j)}, \dots, x_{n_j}^{(j)}$ von unabhängigen, identisch verteilten Zufallsvariablen $X_1^{(j)}, \dots, X_{n_j}^{(j)}$ gegeben, außerdem eine Beobachtung x einer weiteren Zufallsvariablen $X^{(j_X)}$ mit unbekanntem Klassenparameter $j_X \in \{1, \dots, J\}$, wobei nun eine Schätzung des zu X gehörigen Klassenparameters j_X gesucht ist. Allgemein hat man dann das folgende statistische Modell:

$$\left(M^{n+1}, \bigotimes_{\alpha=1}^{n+1} \mathcal{A}, \left(\bigotimes_{j=1}^{J} \bigotimes_{\alpha=1}^{n_j} P_j \otimes P_{j_X} \right)_{(P_1,\dots,P_J,j_X) \in \mathcal{P}^J \times \{1,\dots,J\}} \right). \tag{3.12}$$

Hierbei ist $\mathcal{P}$ eine gegebene Familie von Wahrscheinlichkeisverteilungen. Es ist nun wieder eine Entscheidung über den Parameter j zu treffen. Gesucht ist hierfür eine geeignete Entscheidungsfunktion. Gewissermaßen hat man auch hier wieder das Problem der Schätzung des Klassenparameters j aus dem Parameterraum $\{1, \dots, J\}$.

Gewöhnlich sind die Verteilungen von der Form $P_j = P_{\vartheta(j)}\ \forall j$ mit unbekannten Parametern $\vartheta_j = \vartheta(j)$, die die zur jeweiligen Klasse j ($j = 1, \dots, J$) gehörige Verteilung bestimmen. Ist Θ_0 der gemeinsame Parameterraum für die Verteilungsparameter $\vartheta_j\ \forall j$, so verwendet man für bayessche Betrachtungen hier a-priori-Verteilungen Q auf $\Theta = \{1, \dots, J\} \times \Theta_0$. Ist $R(j, \delta)$ das Risiko der Entscheidungsregel δ zu gegebener Verlustfunktion, so ist das bayessche Risiko dann gegeben durch

$$r(Q, \delta) = \mathrm{E}_Q R(j, \delta). \tag{3.13}$$

Dieses Risiko soll wieder möglichst klein gehalten werden.

Ein naheliegender Ansatz wäre nun, die Parameter $\vartheta_j\ \forall j$ zunächst durch Schätzungen aus den Lernstichproben $x_1^{(1)}, \dots, x_{n_1}^{(1)}, \dots, x_1^{(J)}, \dots, x_{n_J}^{(J)}$ zu ersetzen und dann weiter wie in Abschnitt 3.1 vorzugehen. Solche heuristischen Entscheidungsregeln, bei denen die wahren Parameter einfach durch ihre Schätzungen ersetzt werden, werden auch als *Plug-in*-Regeln bezeichnet (vgl. [8]). Hat man für

alle $j = 1, \ldots, J$ einen konsistenten Schätzer δ_j für $\vartheta(j)$, so streben die Schätzungen $\delta_j(x_1^{(j)}, \ldots, x_{n_j}^{(j)})$ für $n_j \to \infty$ gegen ϑ_j $(j = 1, \ldots, J)$. Für hinreichend große Stichprobenumfänge $n_1, \ldots, n_J$ liegen die Schätzungen dann nahe genug an den wahren Parametern, um so vorgehen zu können. Ist dies aber nicht der Fall, so erreicht man auf diese Weise nicht unbedingt eine Minimierung des Risikos (3.3) bzw. (3.13). Es sind hier also Entscheidungsregeln δ gesucht, die, basierend auf den Beobachtungen $x_1^{(1)}, \ldots, x_{n_1}^{(1)}, \ldots, x_1^{(J)}, \ldots, x_{n_J}^{(J)}$, ein möglichst kleines Risiko bei der Entscheidung für j_X erreichen.

Bei der Klassifikation mit unbekannten Verteilungsparametern sind viele der verwendeten Verfahren heuristisch entstanden. Für manche dieser Verfahren konnten im Nachhinein auch theoretisch gute Eigenschaften, wie zum Beispiel die Zulässigkeit in einer gewissen Klasse von Entscheidungsregeln, nachgewiesen werden [33] [8].

Wir betrachten wieder das Problem der linearen Diskriminanzanalyse, jetzt aber mit unbekannten Parametern. Gegeben seien n Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ von stochastisch unabhängigen Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, die sich in Zufallsvektoren $\boldsymbol{X}_1^{(1)}, \ldots, \boldsymbol{X}_{n_1}^{(1)}$ mit zugehöriger Verteilung $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ und $\boldsymbol{X}_1^{(2)}, \ldots, \boldsymbol{X}_{n_2}^{(2)}$ mit zugehöriger Verteilung $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ $(n = n_1 + n_2)$ aufteilen. Die Parameter $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ und $\boldsymbol{\Sigma}$ seien unbekannt. Man hat dann das Modell

$$\left(\mathbb{R}^{(n+1)\times p}, \mathcal{B}^{(n+1)\times p}, \left(\bigotimes_{\alpha=1}^{n_1} N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}) \otimes \bigotimes_{\alpha=1}^{n_2} N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}) \otimes N(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}) \right)_{\boldsymbol{\vartheta}\in\boldsymbol{\Theta}} \right) \tag{3.14}$$

mit

$$\boldsymbol{\vartheta} = (\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}, j), \tag{3.15}$$

$$\boldsymbol{\Theta} = \mathbb{R}^p \times \mathbb{R}^p \times \mathrm{PD}(p) \times \{1, 2\}. \tag{3.16}$$

Die Statistiken

$$\bar{\boldsymbol{x}}^{(1)} = \frac{1}{n_1} \sum_{\alpha=1}^{n_1} \boldsymbol{x}_\alpha^{(1)}, \bar{\boldsymbol{x}}^{(2)} = \frac{1}{n_2} \sum_{\alpha=1}^{n_2} \boldsymbol{x}_\alpha^{(2)} \tag{3.17}$$

und

$$\boldsymbol{S} = \frac{1}{n-2} \sum_{j=1}^{2} \sum_{\alpha=1}^{n_j} (\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})(\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}}^{(j)})' \tag{3.18}$$

sind konsistente Schätzer für $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ und $\boldsymbol{\Sigma}$. Bei hinreichend großen Stichprobenumfängen n_1 und n_2 können die Parameter in der Diskriminanzfunktion (3.11) einfach durch die entsprechenden Schätzungen ersetzt werden; man erhält dann die Diskriminanzfunktion

$$f(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)}) \right)' \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}), \tag{3.19}$$

wobei wieder für $j = 1$ entschieden wird, falls $f(\boldsymbol{x}) \geq 0$ ist, andernfalls für $j = 2$.

Die Diskriminanzfunktion (3.19) wurde bereits 1936 von R. A. Fisher vorgeschlagen [17]; sie ist daher auch als *Fisher's lineare Diskriminanzregel* bekannt. Dieses Vorgehen wird allgemein auch als Lineare Diskriminanzanalyse (LDA) bezeichnet.

Hier wird jedoch vor allem der Fall betrachtet, dass nicht von „hinreichend großen" Stichprobenumfängen ausgegangen werden kann. Es sind dann wieder Schätzer δ (bzw. Schätzungen $\delta(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}, \boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}; \boldsymbol{x})$) für den zu $\boldsymbol{X}$ gehörendenden Klassenparameter $j_{\boldsymbol{X}}$ gesucht; dabei soll ebenfalls wieder das Risiko (3.3) für alle j (bzw. das bayessche Risiko (3.13)) möglichst klein gehalten werden.

Kapitel 4

Die Problematik hoher Dimensionen in der statistischen Analyse

Zur Lösung des Klassifikationsproblems von hochdimensionalen Beobachtungen gibt es gegenwärtig verschiedene Ansätze. Auf der einen Seite sind dies verschiedene, auf dem Normalverteilungsmodell beruhende Verfahren — zum Teil Modifikationen der Fisher'schen linearen Diskriminanzanalyse. Daneben gibt es auch andere Ansätze — etwa das Prinzip der k *nächsten Nachbarn* (k-nearest neighbour, k-NN). Dort werden für eine zu klassifizierende Beobachtung $\boldsymbol{x}$ jeweils die k Lernstichprobenelemente bestimmt, die die kleinsten Abstände zu $\boldsymbol{x}$ haben. Die Zuordnung erfolgt dann zu der Klasse, die unter diesen Elementen am häufigsten vorkommt [39] [49].

Eine neuere Entwicklung ist die Methode der „Support Vector Machines“ (SVM), die besonders bei hochdimensionalen Daten Verwendung findet. Bei dieser Methode wird der hochdimensionale Raum der Beobachtungen durch eine oder mehrere trennende Hyperebene(n) aufgeteilt. Diese Hyperebenen werden nach geeigneten Kriterien so ausgewählt, dass dadurch eine möglichst optimale Zuordnung der Beobachtungen zu den verschiedenen Klassen erreicht wird [38].

Eine ganze Klasse von zumeist neueren Methoden bilden die so genannten Ensemble-Methoden. Solche Verfahren wenden mehrere Klassifikationsalgorithmen parallel an und generieren dann aus den Einzel-Klassifikationsergebnissen schließlich eine resultierende Klassifikation [12]. Eine neuere Arbeit zu diesem Thema ist zum Beispiel die Dissertationsschrift [25], die sich mit den so genannten *Bundling*-Klassifikatoren, einem speziellen Typ von solchen Ensemble-Methoden, beschäftigt.

In dieser Arbeit sollen aber die auf dem Normalverteilungsmodell beruhenden Verfahren und daraus resultierende Probleme der Analyse im Vordergrund stehen.

4.1 Problem der Parameterschätzung bei hohen Dimensionen

Die Verwendung von $\boldsymbol{S}$ in der Diskriminanzfunktion (3.19) ist ausgeschlossen, wenn die Regularitätsbedingung

$$n - 2 \geq p, \tag{4.1}$$

die eine Voraussetzung für die Invertierbarkeit der Matrix darstellt, nicht erfüllt ist. In der Literatur gibt es verschiedene Vorschläge, $\boldsymbol{S}$ in Gleichung (3.19) durch anders berechnete Statistiken (gewöhnlich Modifikationen von $\boldsymbol{S}$) zu ersetzen. Dazu sei hier besonders auf die Arbeiten [33] und [29] verwiesen, die zu einem Teil am Institut für Biometrie und Medizinische Informatik des Magdeburger Universitätsklinikums entstanden sind.

Im Einstichproben-Normalverteilungsmodell bei unbekannten Parametern lautet Bedingung (4.1)

$$n - 1 \geq p. \tag{4.2}$$

Ist nun Bedingung (4.2) nicht erfüllt, d. h., gilt $n - 1 < p$, sind die Matrizen $\frac{1}{n}\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ und $\boldsymbol{S} = \frac{1}{n-1}\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ — dies gilt allgemein für Matrizen der Form $\boldsymbol{A} = k\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ — singulär. Die Likelihood-Funktion hat hier die Gestalt

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}pn}|\boldsymbol{\Sigma}|^{\frac{1}{2}n}} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_\alpha - \boldsymbol{\mu})\right). \tag{4.3}$$

Für singuläre Matrizen $\boldsymbol{A}$ ist (4.3) für die Wahl $\boldsymbol{\Sigma} = \boldsymbol{A}$ gar nicht definiert. Es stellt sich nun die Frage, inwieweit in diesem Fall überhaupt Maximum-Likelihood-Schätzer existieren oder welche sinnvollen Alternativen es gibt. Es ist zu untersuchen, was in dieser Situation allgemein zur Parameterschätzung, insbesondere zur Schätzung der Kovarianzmatrix $\boldsymbol{\Sigma}$, gesagt werden kann.

4.2 Methoden mit Dimensionsreduzierung

Eine Möglichkeit, das Problem der hohen Dimension in den Griff zu bekommen, ist eine geeignete Dimensionsreduzierung. Die Daten werden auf geeignete Weise in einen Raum mit niedrigerer Dimension transformiert, in welchem die weitere Analyse vorgenommen wird. Dies kann beispielsweise auf Basis der Hauptkomponenten geschehen, wie es bei der Methode der *Hauptkomponentenanalyse* (engl. principal component analysis, PCA), geschieht. Die $q \leq n - J$ Hauptkomponenten mit den größten Varianzen bilden die neuen Variablen, die für die weitere Analyse verwendet werden. Die Größe von q kann dabei entweder fest vorgegeben sein oder ergibt sich durch eine Abbruchbedingung, z. B. $l_i \geq c$ $(i = 1, \ldots, q)$ und $l_i < c$ $(i = q + 1, \ldots, p)$ (vgl. z. B. [29]).

4.2.1 Partielle kleinste Quadrate

Eine andere Methode, die auch mit Dimensionsreduzierung verbunden ist, ist die Methode der *partiellen kleinsten Quadrate* (engl. partial least squares, PLS). Diese Methode wurde ursprünglich von H. Wold eingeführt [51] und wird in verschiedenen Anwendungsgebieten zur Regressions- oder Diskriminanzanalyse eingesetzt, insbesondere für die Analyse hochdimensionaler Daten. Besonders verbreitet ist die Anwendung von PLS in der Chemometrie zur spektrometrischen Kalibrierung. Prädiktoren sind dabei die Emissionsintensitäten zu den verschiedenen Frequenzen eines gewissen Spektrums, abhängige Variablen sind die Anteile von bestimmten Stoffen in einem Stoffgemisch [44] [30].

Die PLS-Analyse basiert auf der Annahme, dass die Beziehung zwischen abhängigen Variablen $\boldsymbol{X} = X_1, \ldots, X_p$ und einer oder mehrerer unabhängiger Variabler $\boldsymbol{Y} = Y_1, \ldots, Y_k$ im linearen Regressionsmodell

$$Y_{ij} = \beta_{0,j} + \sum_{i=1}^{p} X_i \beta_{ij} \ (j = 1, \ldots, k) \tag{4.4}$$

auf einer gewissen Anzahl von $m \leq p$ unbekannten Faktoren beruht. Diese werden als zusätzliche unbekannte Variablen $Z_1, \ldots, Z_m$ modelliert. Die angenommenen kausalen Zusammenhänge zwischen allen diesen bekannten und unbekannten Variablen lassen sich in einem gerichteten Graphen G darstellen. Die Variablen $X_1, \ldots, X_p, Y_1, \ldots, Y_k, Z_1, \ldots, Z_m$ sind die Knoten des Graphen. Nach den angenommenen Abhängigkeiten existieren nun gerichtete Kanten (Bögen) (X_i, Z_l) für gewisse $(i, l) \in \{1, \ldots, p\} \times \{1, \ldots, m\}$, (Z_i, Z_l) für gewisse $(i, l) \in \{1, \ldots, m\}^2$, sowie (Z_l, Y_j) für gewisse $(l, j) \in \{1, \ldots, m\} \times \{1, \ldots, k\}$. Für jeden Bogen $(u, v) \in G$ wird eine lineare Abhängigkeit angenommen (d. h., es gibt eine Darstellung $v = \beta_0 + \beta_1 u$).

Der PLS-Algorithmus läuft nun in drei Phasen ab. In der ersten Phase werden in einem iterativen Prozess die einzelnen Variablen Z_l $(l = 1, \ldots, m)$, die „PLS-Variablen“, als Linearkombinationen aus den Variablen $X_i \in \{X_1, \ldots, X_p\}$, die im Abhängigkeitsgraphen Vorgänger von Z_l sind, d. h., für die es einen Bogen (X_i, Z_l) gibt, bestimmt. Jeder Iterationsschritt beinhaltet die Bestimmung der Kleinste-Quadrate-Schätzung eines gewissen linearen Regressionsproblems und andere Operationen. In der zweiten und dritten Phase werden — ebenfalls als Kleinste-Quadrate-Schätzungen entsprechender linearer Regressionsprobleme — die Koeffizienten für die übrigen Abhängigkeiten des Graphen bestimmt (nach [51]).

Inzwischen gibt es eine ganze Reihe verschiedener Algorithmen, die zu derselben PLS-Lösung führen. Waren die ursprünglichen Ideen, die zur Entwicklung der PLS-Methode geführt haben, auch vorwiegend heuristischer Art, so konnten doch später einige wichtige Eigenschaften der PLS-Methode bewiesen werden. Es konnte gezeigt werden, dass hier ein Maximierungsproblem gelöst wird: Der Koeffizientenvektor (Richtungsvektor) $\boldsymbol{w}_i$ für die i-te PLS-Variable ist (bei $k = 1$)

die Lösung γ von

$$\max_{\|\boldsymbol{\gamma}\|=1, \boldsymbol{w}_l' \boldsymbol{S} \boldsymbol{\gamma}=0 (l=1,\ldots,i-1)} \sum_{j=1}^{p} \mathrm{Corr}^2(Y, X_j\gamma_j)\mathrm{Var}(X_j\gamma_j) \tag{4.5}$$

(nach [50]). Die PLS-Methode hat somit gewisse Ähnlichkeiten mit der Hauptkomponentenanalyse (PCA). Bei der PCA-Methode werden ebenfalls zueinander orthogonale neue Variablen als Linearkombination aus den ursprünglichen Variablen bestimmt. Im Unterschied zu PLS werden bei PCA nicht die Korrelationen, sondern nur die Varianzen maximiert (vgl. [32] [21]).

Die PLS-Methode ist gegenwärtig in vielen statistischen Programmpaketen enthalten, z. B. als Prozedur PLS im Programmpaket SAS/STAT. Für die Anwendung in der Klassifikation wird hierbei eine binäre Klassenvariable als abhängige Variable Y verwendet.

4.2.2 Variablen-Korrelations-Analyse

In einer neueren Arbeit [34] schlägt J. Läuter eine multiple Testprozedur vor, bei der zunächst Cluster korrelierter Variablen („Modelle“) gebildet und diese dann unter strenger Kontrolle des multiplen Testniveaus im engeren Sinne zwischen den Gruppen verglichen werden.

Es wird zunächst der Einstichprobenfall betrachtet. Die Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ von unabhängigen, identisch $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ seien gegeben. Wir betrachten das Testproblem

$$\begin{aligned} H_0 &: \ \boldsymbol{\mu} = \boldsymbol{0} \\ H_1 &: \ \boldsymbol{\mu} \neq \boldsymbol{0}. \end{aligned} \tag{4.6}$$

Auf die einzelnen Variablen X_i $(i = 1, \ldots, p)$ wird zunächst eine geeignete univariate Statistik $F(x_{1,i}, \ldots, x_{n,i})'$ angewendet, etwa die Beta-Statistik B, die bei Gültigkeit der Nullhypothese der Beta-Verteilung $\mathrm{B}(\frac{1}{2}, \frac{n-1}{2})$ unterliegt. Unter der Nullhypothese $\boldsymbol{\mu} = \boldsymbol{0}$ hat jede Zufallsvariable $X_{i,\alpha}$ jeweils die gleiche Verteilung wie die mit -1 multiplizierte Zufallsvariable $-X_{i,\alpha}$ $(i = 1, \ldots, p, \alpha = 1, \ldots, n)$. Dann hat auch die Beta-Statistik die gleiche Verteilung, wenn dabei einige der Beobachtungen $x_{\alpha,i}$ durch $-x_{\alpha,i}$ ersetzt werden.

Jeder der n Beobachtungsvektoren $\boldsymbol{x}_i$ kann also entweder mit $+1$ oder mit -1 multipliziert werden, dadurch ergeben sich 2^n Variationsmöglichkeiten $(\varepsilon_1\boldsymbol{x}_1, \ldots, \varepsilon_n\boldsymbol{x}_n$ mit $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n$. Für jede dieser 2^n Variationen wird dann aus den univariaten Statistiken F jeweils der Wert

$$F^*_{\boldsymbol{x}}(\varepsilon) = \max_{i=1,\ldots,p} F(\varepsilon_1 x_{1,i}, \ldots, \varepsilon_n x_{n,i})' \tag{4.7}$$

berechnet. Ein exakter Test zum Niveau α für das Testproblem (4.6) ist dann mit der Testfunktion

$$\varphi(\boldsymbol{x}) = \left\{ \begin{array}{ll} 1, & \\ 0, & \text{falls } \#(F_{\boldsymbol{x}} \leq F^*_{\boldsymbol{x}}) \begin{array}{l} \leq \\ > \end{array} 2^n\alpha \end{array} \right. \tag{4.8}$$

(mit $F_{\boldsymbol{x}} = F^*_{\boldsymbol{x}}(1,\ldots,1)$) gegeben.

Dies lässt sich nun auf Testprobleme von einzelnen Variablen oder Variablenteilmengen — Modellen — übertragen. Für alle Modelle $M \subset \{1,\ldots,p\}$ betrachten wir die Testprobleme $T^{(M)}$:

$$\begin{array}{ll} H_0^{(M)}: & \mu_i = 0 \\ H_1^{(M)}: & \mu_i \neq 0 \end{array} \quad \forall i \in M. \tag{4.9}$$

Aus der geordneten Folge der Werte $F^*_{\boldsymbol{x}}$

$$F^{(1)}, F^{(2)}, \ldots, F^{(2^n)} \tag{4.10}$$

lässt sich auch das $(1-\alpha)$-Quantil $F^*_{1-\alpha}$ bestimmen. Die Bedingung

$$F_M := \max_{i \in M} F_i(x_i) > F^*_{1-\alpha} \tag{4.11}$$

liefert nun ein Kriterium, dessen Anwendung das Niveau α im multiplen Sinn exakt einhält. Das heißt, die Testfunktion

$$\varphi(\boldsymbol{x}) = \left\{ \begin{array}{ll} 1, & \text{falls } F_M \begin{array}{l} > \\ \leq \end{array} F^*_{1-\alpha} \\ 0, & \end{array} \right. \tag{4.12}$$

liefert zugleich für alle Testprobleme $T^{(M_0)}$ mit $M_0 \subset M$ einen exakten α-Niveau-Test.

In der Arbeit [34] wird nun die folgende Möglichkeit angegeben, wie solche Variablenteilmengen — Modelle — sinnvollerweise gebildet werden können. Gegeben seien die Beobachtungsvektoren $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, $\bar{\boldsymbol{x}}$ ist der Vektor der Mittelwerte. Ausgehend von der Datenmatrix $\boldsymbol{X}$ $(n \times p)$, deren Zeilen jeweils die transponierten Vektoren $\boldsymbol{x'}_1, \ldots, \boldsymbol{x'}_n$ sind, wird zunächst die Matrix der empirischen Korrelationen

$$\boldsymbol{R} = [\mathbf{Diag}(\boldsymbol{X'X})]^{-1/2} \boldsymbol{X'X} [\mathbf{Diag}(\boldsymbol{X'X})]^{-1/2} \tag{4.13}$$

berechnet. Für jede Variable i werden dann Variablenmengen M_i gebildet: Zur Menge M_i gehören alle Variablen j $(j = 1,\ldots,p)$, die für ein gegebenes c $(0 \leq c \leq 1)$ die Bedingung

$$r^2_{ij} \geq c \tag{4.14}$$

erfüllen.

Im Zweistichprobenfall wird zur Berechnung von $\boldsymbol{R}$ in (4.13) die Matrix $\boldsymbol{X}$ durch die Matrix $\boldsymbol{Y} := \boldsymbol{X} - \bar{\boldsymbol{X}}$ $(n \times p)$, deren Zeilen jeweils die transponierten Vektoren $(\boldsymbol{x}^{(1)}_1 - \bar{\boldsymbol{x}})', \ldots, (\boldsymbol{x}^{(1)}_{n_1} - \bar{\boldsymbol{x}})', (\boldsymbol{x}^{(2)}_1 - \bar{\boldsymbol{x}})', \ldots, (\boldsymbol{x}^{(2)}_{n_2} - \bar{\boldsymbol{x}})'$ sind, ersetzt. (Es gilt $n = n_1 + n_2$, $\bar{\boldsymbol{x}}$ ist hier der Vektor der Gesamtmittelwerte aus beiden Stichproben.)

Diese Idee der Bildung der Variablenmodelle wird hier nun wie folgt auf das Zwei-Stichproben-Klassifikationsproblem angewandt. Gegeben seien die Lernstichprobenvektoren $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}, \boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$; die Matrix $\boldsymbol{R}$ wird wie beschrieben berechnet. Es werden wieder zu jedem i $(i = 1, \ldots, p)$ mit Anwendung von Bedingung (4.14) Variablenmengen M_i bestimmt.

Mit $\boldsymbol{Y}_i$ wird jetzt die Teilmatrix von $\boldsymbol{Y}$ bezeichnet, die die in M_i enthaltenen Variablen enthält. Der Eigenvektor $\boldsymbol{g}$ zum größten Eigenwert von $\boldsymbol{Y'}_i\boldsymbol{Y}_i$ wird nun als Gewichtsvektor für den neuen Datenvektor
$\boldsymbol{z}_i = (\, z_{1,i}^{(1)}, \ldots, z_{n_1,i}^{(1)}, z_{1,i}^{(2)}, \ldots, z_{n_2,i}^{(2)},)'$ verwendet. Aus den Ausgangsdatenvektoren $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}, \boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$ erhält man $z_{\alpha,i}^{(j)}$ einfach durch Multiplikation mit $\boldsymbol{g}$:

$$z_{\alpha,i}^{(j)} := \boldsymbol{x'}_{\alpha}^{(j)}\boldsymbol{g} \ (\alpha = 1, \ldots, n_j, j = 1, 2). \tag{4.15}$$

Die Anzahl der neuen Datenvektoren kann gegebenenfalls durch geeignete zusätzliche Kriterien weiter eingeschränkt werden. Auf die neu gebildeten Datenvektoren wird schließlich die klassische lineare Diskriminanzregel (3.19) angewandt.

Das beschriebene Vorgehen liefert eine neue Klassifikationsmethode. Da die neuen Datenvektoren hier auf Grundlage der Korrelationen zwischen den einzelnen Variablen gebildet werden, bezeichnen wir diese Methode als *Variablen-Korrelations-Analyse* (VCA).

4.3 Die einparametrige Ridge-Methode

In unseren weiteren Betrachtungen soll die im Folgenden vorgestellte einparametrige Ridge-Methode im Mittelpunkt stehen. Bei singulärer Stichprobenkovarianzmatrix $\boldsymbol{S}$ stellt die Ridge-Methode eine mögliche Alternative ohne Dimensionsreduzierung dar. D. h., das ursprüngliche Modell des p-dimensionalen Raumes wird dabei beibehalten. Eine Ridge-Schätzung für $\boldsymbol{\Sigma}$ ist eine Statistik der Form $\tilde{\boldsymbol{S}} + \boldsymbol{M}$, wobei $\tilde{\boldsymbol{S}}$ die Form $k\boldsymbol{A}, k > 0$ hat und $\boldsymbol{M}$ positiv definit ist.

Für eine spezielle Klasse von Ridge-Regeln konnte die Eigenschaft der Zulässigkeit für eine gewisse Klasse von Entscheidungsregeln nachgewiesen werden. Es konnte eine a-priori-Verteilung Q auf $\{1, 2\} \times \mathbb{R} \times \mathbb{R} \times \mathrm{PD}(p)$ angegeben werden, für die die angegebene Ridge-Regel das bayessche Risiko (3.13) minimiert [33] [19].

Im einfachsten Fall ist $\boldsymbol{M}$ hierbei ein positives Vielfaches der Einheitsmatrix $\boldsymbol{I}_p$. Die Ridge-Schätzung hat dann die Form

$$\boldsymbol{S}_{\text{ridge}} = \boldsymbol{S} + \lambda\boldsymbol{I}. \tag{4.16}$$

Die Verwendung dieser Form der Ridge-Methode ist sicher besonders angebracht, wenn bekannt ist, dass die Varianzen der einzelnen Variablen etwa die gleiche Größenordnung haben. Diese Voraussetzung ist aber nicht zwingend notwendig.

Bei der Diskriminanzanlayse wird in der Diskriminanzfunktion (3.19) die Ridge-Schätzung (4.16) statt der Stichprobenkovarianzmatrix $\boldsymbol{S}$ eingesetzt; wir

erhalten dann die Diskriminanzfunktion

$$f(\boldsymbol{x}) = \Big(\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\Big)' (\boldsymbol{S} + \lambda \boldsymbol{I})^{-1} (\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}). \tag{4.17}$$

Diese Methode ist als *regularisierte Diskriminanzanalyse* (RDA) bekannt. Sie wird z. B. von Š. Raudys in [42] als Klassifikationsmethode beschrieben.

Man hat bei dieser Methode einen zusätzlichen Parameter λ, der geeignet zu wählen ist. Kriterium dafür ist gewöhnlich das Risiko (3.5), das nach Möglichkeit minimiert werden soll. Man geht dabei zumeist so vor, dass man in mehreren Durchläufen das Klassifikationsverfahren mit verschiedenen Werten von λ an gegebenen Beobachtungen anwendet und jeweils mittels Kreuzvalidierung das zugehörige Risiko schätzt. Für das spezielle Problem erhält man dann näherungsweise den optimalen Wert für λ. [1] Dieses Vorgehen hat unter anderem den Nachteil eines relativ hohen Rechenaufwandes. Gefragt wäre hier eine Methode, direkt anhand der Beobachtungen (d. h. ohne Durchführung der Kreuzvalidierungsberechnungen) einen geeigneten Wert für den Parameter λ zu bestimmen. Dieser Wert kann entweder ein konstanter Wert sein (d. h. unabhängig von den Daten), oder aber als Funktion $\lambda(\bar{\boldsymbol{x}}^{(1)}, \bar{\boldsymbol{x}}^{(2)}, \boldsymbol{S})$ von den beobachteten Daten abhängen.

In dieser Arbeit wird hierbei neben der Minimierung des Klassifikationsrisikos auch das Kriterium der Determinanten-erwartungstreuen Schätzung für die Kovarianzmatrix $\boldsymbol{\Sigma}$ oder deren Inverse verwendet.

[1] Soll für das Verfahren außerdem die Fehlerrate durch Kreuzvalidierung geschätzt werden, so ist dabei doppelte Kreuzvalidierung durchzuführen, d. h., die Kreuzvalidierung für λ ist bei jedem Validierungslauf separat und jeweils nur innerhalb der Lernstichproben vorzunehmen.

Teil II

Parameterschätzung bei hohen Dimensionen

Die meisten theoretischen Aussagen zur multivariaten Statistik in der Literatur (bei multivariater Normalverteilung) beziehen sich auf den Fall einer regulären Stichprobenkovarianzmatrix. Viele Methoden zur Klassifikation sind aus Methoden zur Parameterschätzung abgeleitet oder aber mit solchen sehr verwandt. Daher sollen in diesem zweiten Teil zunächst allgemeiner Aussagen zur Parameterschätzung, insbesondere auch bei singulärer Stichprobenkovarianzmatrix, gemacht werden. Im Mittelpunkt der Betrachtungen steht dabei die Schätzung der Kovarianzmatrix $\mathbf{\Sigma}$.

Das Maximum-Likelihood-Prinzip ist ein sehr häufig angewandtes Prinzip der statistischen Analyse überhaupt. Aufgrund seiner guten Eigenschaften liefert dieses besonders bei hinreichend hohen Beobachtungsumfängen gute Ergebnisse. Daher wird die Frage der Existenz von Maximum-Likelihood-Schätzern hier zunächst näher untersucht.

Die anschließenden Stabilitätsbetrachtungen sollen vor allem dazu dienen, die Besonderheit der Situation einer singulären Stichprobenkovarianzmatrix zu charakterisieren. Es wird eine Aussage über den Abstand der Schätzungen vom wahren Parameter hergeleitet.

Der Einfachheit halber wird in diesem Teil der Arbeit hauptsächlich das Einstichprobenmodell zugrundegelegt. Die Betrachtungen lassen sich aber leicht auf das Mehrstichprobenmodell mit J Stichproben übertragen. Statt $n-1$ hat die Stichprobenkovarianzmatrix $\boldsymbol{S}$ dann jeweils den Rang $n-J$. Die meisten Aussagen gelten dann analog, wenn $n-1$ an entsprechender Stelle jeweils durch $n-J$ ersetzt wird.

Kapitel 5

Maximum-Likelihood-Schätzung

Die Theorie der Hauptkomponenten soll jetzt für Aussagen zur Maximum-Likelihood-Schätzung angewandt werden. im Folgenden geht es zunächst um die Schätzung der Eigenwerte und Eigenvektoren der Kovarianzmatrix einer p-dimensionalen multivariaten Normalverteilung. Dazu sollen Aussagen zur Existenz von Maximum-Likelihood-Schätzern bei regulärer sowie auch bei singulärer Stichprobenkovarianzmatrix gemacht werden.

5.1 Schätzung bei regulärer Stichprobenkovarianzmatrix

Sind die Eigenwerte einer symmetrischen Matrix — wie der Kovarianzmatrix einer multivariaten Normalverteilung — alle paarweise verschieden, so sind die zugehörigen normierten Eigenvektoren bis auf das Vorzeichen eindeutig bestimmt. Bei eindeutig festgelegter Reihenfolge der Komponenten $x_1, \ldots, x_p$ von Vektoren $\boldsymbol{x} \in \mathbb{R}^p$ lässt sich eine eindeutige Auswahl vornehmen, etwa durch die Forderung, dass die erste von Null verschieden Komponente positiv sein soll. Lässt man bei der Spektralzerlegung nur auf solche Weise eindeutig bestimmte Orthogonalmatrizen zu, so hat man hier eine bijektive Abbildung. Für die Maximum-Likelihood-Schätzung lässt sich dann der folgende Satz anwenden.

Satz 5.1 (z. B. [3], Corollary 3.2.1) *Sei $(M, \mathcal{A}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\varphi : \Theta \to \Theta^*$ eine Bijektion. Ist $\hat{\vartheta}$ Maximum-Likelihood-Schätzer für ϑ, so ist $\varphi(\hat{\vartheta})$ Maximum-Likelihood-Schätzer für $\varphi(\vartheta)$.*

Abweichend von (2.2) betrachten wir jetzt das folgende statistische Modell:

$$\left(\mathbb{R}^{n\times p}, \mathcal{B}^{n\times p}, \left(\bigotimes_{\alpha=1}^{n} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right)_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\boldsymbol{\Theta}}\right) \tag{5.1}$$

mit $\boldsymbol{\Theta} = \mathbb{R}^p \times \{\boldsymbol{A} \in \mathrm{PD}(p) \text{ mit EW } \lambda_1 > \ldots > \lambda_p > 0\}$. Die Vektoren $\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(p)}$ seien die zugehörigen normierten Eigenvektoren von $\boldsymbol{\Sigma}$ (mit eindeutig bestimmtem Vorzeichen). Wir definieren eine Abbildung φ auf $\boldsymbol{\Theta}$ gemäß

$$\varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}, \lambda_1, \ldots, \lambda_p, \boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(p)}). \tag{5.2}$$

Es sei $(\bar{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}})$ Maximum-Likelihood-Schätzer für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Mit Wahrscheinlichkeit 1 sind die Eigenwerte von $\hat{\boldsymbol{\Sigma}}$ alle voneinander verschieden, und nach Satz 5.1 ist $\varphi(\bar{\boldsymbol{x}}, \hat{\Sigma}) = (\bar{\boldsymbol{x}}, l_1, \ldots, l_p, \boldsymbol{g}_{(1)}, \ldots, \boldsymbol{g}_{(p)})$ Maximum-Likelihood-Schätzer für $\varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}, \lambda_1, \ldots, \lambda_p, \boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(p)})$, wobei hier $l_1 > \ldots > l_p > 0$ die Eigenwerte und $\boldsymbol{g}_{(1)}, \ldots, \boldsymbol{g}_{(p)}$ die zugehörigen normierten Eigenvektoren (mit entsprechend eindeutig festgelegtem Vorzeichen) seien.

Dies gilt allerdings nur, falls die Eigenwerte von $\boldsymbol{\Sigma}$ alle voneinander verschieden sind. Für den Fall, dass Eigenwerte von $\boldsymbol{\Sigma}$ mehrfach vorkommen, hat T. W. Anderson in [2] ein ähnliches Resultat bewiesen. Im nächsten Abschnitt soll dies noch weiter verallgemeinert werden. Es soll dann die Situation einer möglicherweise singulären Stichprobenkovarianzmatrix betrachtet werden.

5.2 Schätzung bei nicht notwendig regulärer Stichprobenkovarianzmatrix

T. W. Anderson [2, Theorem 2] hat die Maximum-Likelihood-Schätzer für die Eigenwerte und die Eigenvektoren der Kovarianzmatrix $\boldsymbol{\Sigma}$ auch für den Fall von mehrfachen Eigenwerten von $\boldsymbol{\Sigma}$ angegeben. In diesem Abschnitt soll im Wesentlichen die Aussage dieses Satzes noch erweitert werden. Dazu wird die Idee des Beweises von [2] übernommen; der Beweis wurde aufgearbeitet und wird hier für die erweiterte Aussage wesentlich ausführlicher präsentiert.

Lemma 5.1 *Es seien*

$$a_1 \geq \ldots \geq a_p \geq 0 \tag{5.3}$$

und

$$0 \leq b_1 \leq \ldots \leq b_p \tag{5.4}$$

gegeben. Eine Matrix $\boldsymbol{P} = [p_{ij}]$ minimiert genau dann den Wert von

$$\sum_{i=1}^{p} \sum_{j=1}^{p} a_i b_j p_{ij}^2 \tag{5.5}$$

bezüglich aller $p \times p$-Matrizen mit folgenden Eigenschaften:

$$\sum_{j=1}^{p} p_{ij}^2 = 1 \ \forall i = 1, \ldots, p \tag{5.6}$$

$$\sum_{i=1}^{p} p_{ij}^2 = 1 \ \forall j = 1, \ldots, p, \tag{5.7}$$

wenn (mindestens) eine der folgenden Bedingungen erfüllt ist:

$$\sum_{j:b_i=b_j} p_{ij}^2 = 1 \ \forall i = 1, \ldots, p \tag{5.8}$$

$$\sum_{i:a_i=a_j} p_{ij}^2 = 1 \ \forall j = 1, \ldots, p. \tag{5.9}$$

Der minimale Wert von (5.5) beträgt $\sum_{i=1}^{p} a_i b_i$.

Beweis: Zunächst soll gezeigt werden, dass das Erfülltsein von Bedingung (5.8) oder (5.9) für das Erreichen des Minimums notwendig ist. Dies soll indirekt erfolgen.

Sei $\boldsymbol{P}$ eine Matrix mit den Eigenschaften (5.6) und (5.7), es seien aber (5.8) und (5.9) nicht erfüllt. Dann gibt es eine Spalte j mit $\sum_{i:a_i=a_j} p_{ij}^2 < 1$. Es gibt also (mindestens) ein i, so dass

$$p_{ij} \neq 0, a_i \neq a_j \tag{5.10}$$

gilt. Sei k der kleinste Index i, für den (5.10) für ein j gilt:

$$k := \min\{i : \exists j : a_i \neq a_j \text{ und } p_{ij} \neq 0\}. \tag{5.11}$$

Das heißt, es gilt

$$p_{ij} = 0, \text{falls } a_i \neq a_j \ \forall i < k \ \forall j. \tag{5.12}$$

Da (5.8) nicht erfüllt ist, gibt es auch eine Zeile i mit $\sum_{j:b_i=b_j} p_{ij}^2 < 1$. Es gibt also ein j, so dass

$$p_{ij} \neq 0, b_i \neq b_j \tag{5.13}$$

gilt. Sei l der kleinste Index j, so dass (5.13) für ein i gilt:

$$l := \min\{j : \exists i : b_i \neq b_j \text{ und } p_{ij} \neq 0\}. \tag{5.14}$$

Damit gilt

$$p_{ij} = 0, \text{falls } b_i \neq b_j \ \forall i, j < l. \tag{5.15}$$

Mit (5.7) gilt für die Summe der Normen von r Spaltenvektoren von $\boldsymbol{P}$

$$\sum_{\alpha=1}^{r} \sum_{i=1}^{p} p_{i_\alpha j}^2 = r \tag{5.16}$$

und damit für die Spalten j mit $a_j = a_k$:

$$\sum_{j:a_j=a_k} \sum_{i=1}^{p} p_{ij}^2 = |\{j : a_j = a_k\}| \tag{5.17}$$

($|\{\ldots\}|$ bezeichnet hier die Mächtigkeit einer Menge $\{\ldots\}$). Mit $p^2_{kj_k} > 0$ für ein j_k mit $a_{j_k} \neq a_k$ folgt dann wegen $\sum\limits_{j:a_j=a_k} p^2_{kj} \leq 1 - p^2_{kj_k}$

$$\sum_{j:a_j=a_k} \sum_{i:a_i=a_k} p^2_{ij} \leq \sum_{j:a_j=a_k} \sum_{i=1}^{p} p^2_{ij} - p^2_{kj_k}, \tag{5.18}$$

d. h., es gilt

$$\sum_{j:a_j=a_k} \sum_{i:a_i=a_k} p^2_{ij} < \sum_{j:a_j=a_k} \sum_{i=1}^{p} p^2_{ij}. \tag{5.19}$$

Das heißt, es gibt eine Spalte j mit $a_j = a_k$ und $\sum\limits_{i:a_i=a_k} p^2_{ij} < 1$, es gibt also eine Zeile c mit

$$a_c \neq a_k, \tag{5.20}$$

damit gilt auch $a_c \neq a_j$, und $q_c := p^2_{cj} > 0$. Nach (5.12) gilt $c \geq k$; wegen (5.20) folgt $c > k$.

Für die Zeilenvektoren von $\boldsymbol{P}$ gilt analog Bedingung (5.6). Damit erhält man auf gleiche Weise: Es gibt eine Zeile i mit $b_i = b_l$ und $\sum\limits_{j:b_j=b_l} p^2_{ij} < 1$. D. h., es gibt eine Spalte d mit

$$b_d \neq b_l, \tag{5.21}$$

damit gilt wieder $b_i \neq b_d$, und $q_d := p^2_{id} > 0$. Nach (5.15) gilt $d \geq l$; mit (5.21) folgt $d > l$.

Sei nun $q := \min\{q_c, q_d\}$. Ersetzt man p^2_{kl} durch $p'^2_{kl} := p^2_{kl} + q$, p^2_{kd} durch $p'^2_{kd} := p^2_{kd} - q$, p^2_{cl} durch $p'^2_{cl} := p^2_{cl} - q$ und p^2_{cd} durch $p'^2_{cd} := p^2_{cd} + q$, erhält man für (5.5)

$$\sum_{i=1}^{p}\sum_{j=1}^{p} a_i b_j p^2_{ij} + q(a_k b_l - a_k b_d - a_c b_l + a_c b_d) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i b_j p^2_{ij} + q(a_k - a_c)(b_l - b_d). \tag{5.22}$$

Da $c > k$ ist, folgt aus (5.20) und Voraussetzung (5.3) $a_k < a_c$. Weiterhin ist $d > l$, damit gilt wegen (5.21) und (5.4) $b_l > b_d$. Da $q > 0$ ist, ist Ausdruck (5.22) kleiner als Ausdruck (5.5). Dies ist ein Widerspruch zur Minimalität von (5.5). Das zeigt, dass mindestens eine der Bedingungen (5.8) und (5.9) zur Erreichung des Minimums erfüllt sein muss.

Sei nun umgekehrt angenommen, dass (5.8) gilt, d. h. $\sum\limits_{j:b_i=b_j} p^2_{ij} = 1 \ \forall i$. Für (5.5) gilt dann:

$$\sum_{i=1}^{p}\sum_{j=1}^{p} a_i b_j p^2_{ij} \;=\; \sum_{i=1}^{p} \sum_{j:b_i=b_j} a_i b_j p^2_{ij} \tag{5.23}$$

$$= \sum_{i=1}^{p} a_i b_i \underbrace{\sum_{j:b_i=b_j} p_{ij}^2}_{=1} \tag{5.24}$$

$$= \sum_{i=1}^{p} a_i b_i. \tag{5.25}$$

Für den Fall, dass (5.9) gilt, erhält man in analoger Weise ebenfalls Ausdruck (5.25). Dieser Wert ist jeweils unabhängig von weiteren Eigenschaften von $\boldsymbol{P}$. Damit ist gezeigt, dass das Erfülltsein einer der Bedingungen (5.8) und (5.9) für das Erreichen des Minimums ausreichend ist. □

Korollar 5.1 *Seien (5.3) und (5.4) vorausgesetzt. Das Erfülltsein einer der Bedingungen (5.8) und (5.9) ist auch notwendig und hinreichend für die Minimierung von Ausdruck (5.5) bezüglich aller orthogonaler $p \times p$-Matrizen; der minimale Wert ist hierbei ebenfalls $\sum_{i=1}^{p} a_i b_i$.*

Beweis: Eine orthogonale $p \times p$-Matrix erfüllt auch die Bedingungen (5.6) und (5.7). Die Menge aller orthogonaler $p \times p$-Matrizen ist also eine Teilmenge der Menge der $p \times p$-Matrizen mit den Eigenschaften (5.6) und (5.7). Die Einheitsmatrix ist eine orthogonale Matrix, die trivialerweise (5.8) und (5.9) erfüllt und damit (5.5) minimiert. Damit wird das Minimum auch von orthogonalen Matrizen angenommen, und die Aussage gilt nach obigem Lemma. □

Bei singulärer Stichprobenkovarianzmatrix lässt sich bei multivariater Normalverteilung kein Maximum-Likelihood-Schätzer für den gesamten Parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ angeben, wie wir später noch sehen werden. Es wird im Folgenden die Frage untersucht, inwieweit für $\boldsymbol{\mu}$ und die Eigenvektoren von $\boldsymbol{\Sigma}$ Maximum-Likelihood-Schätzer existieren — unter der speziellen Annahme, dass die Eigenwerte von $\boldsymbol{\Sigma}$ bekannt seien.

Satz 5.2 *Seien $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ Beobachtungen von stochastisch unabhängigen, identisch $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $\boldsymbol{\Sigma}$ sei positiv definit. Die Eigenwerte $\lambda_1 \geq \ldots \geq \lambda_p > 0$ von $\boldsymbol{\Sigma}$ seien bekannt, $\boldsymbol{\Gamma}$ sei die orthogonale Matrix, deren Spalten $\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(p)}$ die zugehörigen normierten Eigenvektoren sind. Die Eigenwerte von $\tilde{\boldsymbol{S}} = \frac{1}{n} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ seien mit $l_1 \geq \ldots \geq l_p \geq 0$ bezeichnet, $\boldsymbol{G}$ sei die orthogonale Matrix, deren Spalten die zugehörigen normierten Eigenvektoren $\boldsymbol{g}_{(1)}, \ldots, \boldsymbol{g}_{(p)}$ sind. Eine Statistik $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Gamma}})$ ist genau dann Maximum-Likelihood-Schätzer für $(\boldsymbol{\mu}, \boldsymbol{\Gamma})$, wenn gilt:*

1. $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$
2. $\hat{\boldsymbol{\Gamma}}$ *ist von der Form $\boldsymbol{GP}$, wobei $\boldsymbol{P}$ eine orthogonale $p \times p$-Matrix ist, die mindestens eine der folgenden Bedingungen (5.26) bzw. (5.27) erfüllt.*

$$\sum_{j:\lambda_i=\lambda_j} p_{ij}^2 = 1 \ \forall i = 1, \ldots, p \tag{5.26}$$

$$\sum_{i:l_i=l_j} p_{ij}^2 \;=\; 1 \;\forall j = 1, \ldots, p. \tag{5.27}$$

Beweis: Die gemeinsame Verteilung von $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ hat die Likelihood-Funktion

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}pn} |\boldsymbol{\Sigma}|^{\frac{1}{2}n}} \exp\left(-\frac{1}{2} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) \right). \tag{5.28}$$

Zur Maximierung von (5.28) ist es äquivalent, die Log-Likelihood-Funktion

$$l = \ln L = -\frac{1}{2} pn \ln(2\pi) - \frac{1}{2} n \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) \tag{5.29}$$

zu maximieren.

Es gilt die folgende Identität:

$$\begin{aligned} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})(\boldsymbol{x}_\alpha - \boldsymbol{\mu})' &= \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})' + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \\ &= \boldsymbol{A} + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \end{aligned} \tag{5.30}$$

mit $\boldsymbol{A} := \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$. (Beweis in [3, 3.2])

Für die Spur des Produktes zweier Matrizen $\boldsymbol{B}(k \times r)$ und $\boldsymbol{C}(r \times k)$ gilt folgende Eigenschaft:

$$\mathrm{tr}\boldsymbol{BC} = \sum_{i=1}^{k} \sum_{j=1}^{r} b_{ij} c_{ji} = \mathrm{tr}\boldsymbol{CB}. \tag{5.31}$$

Dies ist leicht zu zeigen (zum Beweis siehe z. B. [47, 2.8]). Außerdem gilt $\mathrm{tr}c = c$ für jede Zahl c und $\mathrm{tr} \sum_{\alpha=1}^{n} \boldsymbol{B}_\alpha = \sum_{\alpha=1}^{n} \mathrm{tr}\boldsymbol{B}_\alpha$. Da $(\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})$ eine Zahl ist, erhält man mit (5.30) und (5.31)

$$\begin{aligned} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) &= \sum_{\alpha=1}^{n} \mathrm{tr}(\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) && (5.32) \\ &= \mathrm{tr}\boldsymbol{\Sigma}^{-1} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})(\boldsymbol{x}_\alpha - \boldsymbol{\mu})' && (5.33) \\ &= \mathrm{tr}\boldsymbol{\Sigma}^{-1} A + \mathrm{tr}\boldsymbol{\Sigma}^{-1} n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' && (5.34) \\ &= \mathrm{tr}\boldsymbol{\Sigma}^{-1} A + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}). && (5.35) \end{aligned}$$

Für (5.29) erhält man dann

$$l = -\frac{1}{2} pn \ln(2\pi) - \frac{1}{2} n \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \mathrm{tr}\boldsymbol{\Sigma}^{-1} \boldsymbol{A} - \frac{1}{2} n (\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}). \tag{5.36}$$

Die Matrix $\boldsymbol{\Sigma}$ soll positiv definit sein, dies gilt dann auch für $\boldsymbol{\Sigma}^{-1}$. Damit gilt

$$(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \geq 0, \tag{5.37}$$

der Term erreicht mit $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ seinen minimalen Wert 0. Die Wahl von $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ ist somit notwendige Voraussetzung für die Maximierung der Likelihood-Funktion — dies gilt unabhängig vom Rang von $\boldsymbol{A}$.

Da $-\frac{1}{2}pn\ln(2\pi)$ von $\boldsymbol{\Sigma}$ unabhängig ist, bleibt nun, das Maximum von

$$-\frac{1}{2}n\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{A} \tag{5.38}$$

zu bestimmen. Mit $\boldsymbol{L}$ (bzw. $\boldsymbol{\Lambda}$) sei hier wieder die Diagonalmatrix bezeichnet, deren Diagonalelemente die Eigenwerte von $\boldsymbol{S}$ (bzw. von $\boldsymbol{\Sigma}$) sind. Wird nun $\boldsymbol{A}$ durch $n\tilde{\boldsymbol{S}}$, $\tilde{\boldsymbol{S}}$ wiederum durch $\boldsymbol{GLG'}$ sowie $\boldsymbol{\Sigma}$ durch $\boldsymbol{\Gamma\Lambda\Gamma'}$ ersetzt, erhält man

$$-\frac{1}{2}n\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{A} = -\frac{1}{2}n\ln|\boldsymbol{\Gamma\Lambda\Gamma'}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma\Lambda\Gamma'})^{-1}n\boldsymbol{GLG'}. \tag{5.39}$$

Die Matrix $\hat{\boldsymbol{\Gamma}}$ ist nun so zu bestimmen, dass deren Spaltenvektoren die gesuchten Maximum-Likelihood-Schätzer für die Eigenvektoren von $\boldsymbol{\Sigma}$ sind. $\boldsymbol{\Gamma}$ ist damit orthogonal, und es gilt nach Abschnitt 2.2.1 $|\boldsymbol{\Gamma\Lambda\Gamma'}| = |\boldsymbol{\Lambda}|$ und $(\boldsymbol{\Gamma\Lambda\Gamma'})^{-1} = \boldsymbol{\Gamma\Lambda}^{-1}\boldsymbol{\Gamma'}$. Damit erhält man für (5.39)

$$-\frac{1}{2}n\ln|\boldsymbol{\Gamma\Lambda\Gamma'}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma\Lambda\Gamma'})^{-1}n\boldsymbol{GLG'} = -\frac{1}{2}n\ln|\boldsymbol{\Lambda}| - \frac{1}{2}\mathrm{tr}n\boldsymbol{\Gamma\Lambda}^{-1}\boldsymbol{\Gamma'GLG'}. \tag{5.40}$$

Der erste Term ist also von der Wahl von $\boldsymbol{\Gamma}$ unabhängig; es bleibt,

$$-\frac{1}{2}\mathrm{tr}n\boldsymbol{\Gamma\Lambda}^{-1}\boldsymbol{\Gamma'GLG'} \tag{5.41}$$

zu maximieren bzw.

$$\mathrm{tr}\boldsymbol{\Gamma\Lambda}^{-1}\boldsymbol{\Gamma'GLG'} \tag{5.42}$$

zu minimieren. Mit (5.31) und $\boldsymbol{P} := \boldsymbol{G'\Gamma}$ erhält man für (5.42)

$$\mathrm{tr}\boldsymbol{\Gamma\Lambda}^{-1}\,\boldsymbol{\Gamma'GLG'} = \mathrm{tr}\boldsymbol{\Lambda}^{-1}\,\boldsymbol{\Gamma'GLG'\Gamma} = \mathrm{tr}\boldsymbol{\Lambda}^{-1}\boldsymbol{P'LP}. \tag{5.43}$$

Da $\boldsymbol{\Lambda}$ und $\boldsymbol{L}$ Diagonalmatrizen sind, erhält man für (5.43) (mit (5.31))

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{P'})(\boldsymbol{LP}) &= \mathrm{tr}(\boldsymbol{LP})(\boldsymbol{\Lambda}^{-1}\boldsymbol{P'}) && (5.44)\\
&= \mathrm{tr}(\boldsymbol{LP})(\boldsymbol{P\Lambda}^{-1})' && (5.45)\\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}(l_i p_{ij})(p_{ij}\lambda_j^{-1}) && (5.46)\\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{l_i}{\lambda_j}p_{ij}^2. && (5.47)
\end{aligned}$$

Als ein Produkt zweier orthogonaler Matrizen ist auch $\boldsymbol{P}$ orthogonal. Setzt man nun $a_i := l_i$ und $b_j := \frac{1}{\lambda_j}$ $(i, j = 1, \ldots, p)$, so erhält man nach Lemma 5.1 und Korollar 5.1 das Minimum von (5.47) für alle orthogonalen Matrizen $\boldsymbol{P}$, die eine der Bedingungen (5.26) und (5.27) erfüllen. Mit $\boldsymbol{P} = \boldsymbol{G'}\hat{\boldsymbol{\Gamma}}$ erhält man $\hat{\boldsymbol{\Gamma}} = \boldsymbol{G}\boldsymbol{G'}\hat{\boldsymbol{\Gamma}} = \boldsymbol{G}\boldsymbol{P}$. □

Korollar 5.2 *Es gelten die obigen Bezeichnungen und Voraussetzungen. Dann ist stets* $(\bar{\boldsymbol{x}}, \boldsymbol{G})$ *ein Maximum-Likelihood-Schätzer für* $(\boldsymbol{\mu}, \boldsymbol{\Gamma})$.

Beweis: Wähle für $\boldsymbol{P}$ die Einheitsmatrix. □

Bemerkung 5.1

1. *Der Satz gilt auch für den Fall, dass ein oder mehrere Eigenwerte von* $\tilde{\boldsymbol{S}}$ *Null sind, d. h. für den Fall, dass* $\tilde{\boldsymbol{S}}$ *singulär ist, also für* $n \leq p$. *Insbesondere existieren für diesen Fall nicht notwendig eindeutige Maximum-Likelihood-Schätzer für die Eigenvektoren von* $\boldsymbol{\Sigma}$.

2. *Sind mehrere Eigenwerte von* $\tilde{\boldsymbol{S}}$ *Null, so lassen sich die zugehörigen Eigenvektoren nicht mehr eindeutig zuordnen. Bis auf das Vorzeichen eindeutig lassen sich aber mit dem Satz die Maximum-Likelihood-Schätzer zu den übrigen Eigenvektoren bestimmen. Die noch fehlenden Vektoren sind dann unter Beachtung der Orthogonalität untereinander frei wählbar.*

Satz 5.3 *Seien* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *Beobachtungen von stochastisch unabhängigen,* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*-verteilten Zufallsvektoren* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $\boldsymbol{\Sigma}$ *sei positiv definit. Die Matrix* $\tilde{\boldsymbol{S}} = \frac{1}{n}\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ *sei singulär. Die Matrizen* $\boldsymbol{G}$, $\boldsymbol{L}$, $\boldsymbol{\Gamma}$ *und* $\boldsymbol{\Lambda}$ *seien definiert wie oben. Dann gibt es unter allen Diagonalmatrizen vom Format* $p \times p$ *mit positiven Diagonalelementen kein* $\hat{\boldsymbol{\Lambda}}$, *so dass* $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Lambda}})$ *Maximum-Likelihood-Schätzer für* $(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda})$ *ist.*

Beweis: Da $\tilde{\boldsymbol{S}}$ singulär ist, gilt $|\tilde{\boldsymbol{S}}| = 0$ und wegen $\tilde{\boldsymbol{S}} = \boldsymbol{G}\boldsymbol{L}\boldsymbol{G'}$ auch $|\boldsymbol{L}| = 0$. Es gibt also einen Eigenwert l_i von $\tilde{\boldsymbol{S}}$ mit $l_i = 0$; da die Eigenwerte als geordnet vorausgesetzt seien, gilt $l_p = 0$. Hier ist nun (5.39) bezüglich $\boldsymbol{\Gamma}$ und $\boldsymbol{\Lambda}$ zu maximieren. Mit gleicher Argumentation wie oben erhält man den Ausdruck

$$\ln|\boldsymbol{\Lambda}| + \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{l_i}{\lambda_j}p_{ij}^2, \tag{5.48}$$

der bezüglich $\boldsymbol{\Gamma}$ und $\boldsymbol{\Lambda}$ zu minimieren ist.

Es sei nun angenommen, dass ein Maximum-Likelihood-Schätzer $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Lambda}})$ mit $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, $\hat{\lambda}_i > 0\ \forall i$ existiert. $\hat{\boldsymbol{\Gamma}}$ muss dann die in Satz 5.2 genannten Eigenschaften erfüllen. Definiere weitere Diagonalmatrix $\boldsymbol{\Lambda}^*$ mit $\lambda_i^* := \hat{\lambda}_i$ $(i = 1, \ldots, p-1)$, und $\lambda_p^* := \hat{\lambda}_p/2$. Nach Lemma 5.1 und Korollar 5.1 hat der zweite Term von (5.48) den minimalen Wert $\sum_{i=1}^{p} l_i/\hat{\lambda}_i$. Da $l_p = 0$ ist, bleibt dieser Wert auch mit λ_p^* unverändert. Für den ersten Term gilt:

$$\ln|\hat{\boldsymbol{\Lambda}}| = \ln(\hat{\lambda}_1 \cdot \ldots \cdot \hat{\lambda}_p) > \ln(\lambda_1^* \cdot \ldots \cdot \lambda_p^*) = \ln|\boldsymbol{\Lambda}^*|. \tag{5.49}$$

Damit ist $(\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Lambda}})$ nicht Minimallösung von (5.48). Ein solcher Maximum-Likelihood-Schätzer für $(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda})$ kann somit nicht existieren. □

Als direkte Folgerung erhält man hier mit Anwendung des Satzes 5.1 eine bekannte Existenzaussage für Maximum-Likelihood-Schätzer für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (vgl. [4, Korollar 9.6 und Beispiel 9.11]):

Korollar 5.3 *Seien $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ Beobachtungen von stochastisch unabhängigen, identisch $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $\boldsymbol{\Sigma}$ sei positiv definit. Die Matrix $\tilde{\boldsymbol{S}} = \frac{1}{n} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ sei singulär. Dann gibt es unter allen positiv definiten $p \times p$-Matrizen kein $\hat{\boldsymbol{\Sigma}}$, so dass $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ Maximum-Likelihood-Schätzer für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ist.*

Bemerkung 5.2

1. *Die Ergebnisse zeigen: Das eigentliche Problem bei der Schätzung der Matrix $\boldsymbol{\Sigma}$ im Fall $n - 1 < p$ ist die Schätzung ihrer Eigenwerte. Für die Eigenvektoren können — wenn auch nicht eindeutig — ML-Schätzungen angegeben werden.*

2. *Sei mit $\boldsymbol{A}$ die Matrix $\sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ bezeichnet, so dass $\tilde{\boldsymbol{S}} = \frac{1}{n}\boldsymbol{A}$ und $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{A}$ gilt. Ist $\tilde{\boldsymbol{S}} = \boldsymbol{GLG'}$ die Spektralzerlegung von $\tilde{\boldsymbol{S}}$, so gilt mit $\boldsymbol{A} = n\tilde{\boldsymbol{S}}$*

$$\boldsymbol{A} = n\boldsymbol{GLG'} \tag{5.50}$$

und mit $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{A}$

$$\boldsymbol{S} = \frac{1}{n-1} n\, \boldsymbol{GLG'} = \frac{n}{n-1}\boldsymbol{GLG'} = \boldsymbol{G}\Big(\frac{n}{n-1}\boldsymbol{L}\Big)\boldsymbol{G'}. \tag{5.51}$$

So erhält man die Spektralzerlegung von $\boldsymbol{S}$. Das bedeutet, dass $\boldsymbol{S}$ und $\tilde{\boldsymbol{S}}$ dieselben Eigenvektoren haben und sich nur in ihren Eigenwerten unterscheiden. (Damit können insbesondere die Eigenvektoren von $\boldsymbol{S}$ ebenso zur Konstruktion von Maximum-Likelihood-Schätzern verwendet werden.)

5.3 Induzierte Likelihood-Funktionen

5.3.1 Einführung

Bei bijektiven Parametertransformationen lässt sich mit Satz 5.1 das Prinzip der Maximum-Likelihood-Schätzung einfach auf den Bildraum übertragen. Für beliebige (d. h. nicht notwendig bijektive) Funktionen betrachtet man stattdessen *induzierte Likelihood-Funktionen* (vgl. z. B. [16] [40]).

Definition 5.1 ([16], Definition 4.1.1, [40], 2.1) *Sei $(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, μ sei ein Maß auf $\mathcal{A}$, f_ϑ sei μ-Dichte $\forall\vartheta \in \Theta$. Die Funktion $L(\cdot, x) : \Theta \to [0, \infty)$ mit $L(\vartheta, x) = f_\vartheta(x)$ bezeichne hierbei die Likelihood-Funktion. Weiterhin sei $g : \Theta \to \mathcal{G}$ eine Funktion auf dem Parameterraum Θ.*

Sei $\Theta^* := g(\Theta)$*, für ein* $\vartheta^* \in \Theta^*$ *sei* $C(\vartheta^*) := \{\vartheta \in \Theta | g(\vartheta) = \vartheta^*\}$*. Für* $x \in \Omega$ *und* $\vartheta^* \in \Theta^*$ *setze*

$$M(\vartheta^*, x) = \sup_{\vartheta \in C(\vartheta^*)} L(\vartheta, x). \tag{5.52}$$

Dann heißt $M(\cdot, x) : \Theta^* \to [0, \infty]$ *die durch* g *induzierte Likelihood-Funktion,* $M(\vartheta^*, x)$ *heißt durch* g *induzierte Likelihood.*

Gilt für ein $\hat{\vartheta}^* \in \Theta^*$ *zu einem gegebenen* $x \in \Omega$ $M(\hat{\vartheta}^*, x) \geq M(\vartheta^*, x)$ *für alle* $\vartheta^* \in \Theta^*$*, so heißt* $\hat{\vartheta}^*$ *Maximum-Likelihood-Schätzung (ML-Schätzung) für* $g(\vartheta)$*. Sind* $\emptyset \neq \Omega_0 \subset \Omega, \Omega_0 \in \mathcal{A}$ *und* $\hat{\vartheta}^* : \Omega_0 \to \Theta^*$*, so dass* $\hat{\vartheta}^*(x)$ *für jedes* $x \in \Omega_0$ *eine ML-Schätzung für* $g(\vartheta)$ *ist, dann ist* $\hat{\vartheta}^*$ *ein auf* Ω_0 *definierter Maximum-Likelihood-Schätzer (ML-Schätzer) für* $g(\vartheta)$*. Gilt* $P_\vartheta(\Omega_0) = 1$ $\forall \vartheta \in \Theta$*, dann ist* $\hat{\vartheta}^*$ *ein (fast überall definierter) ML-Schätzer für* $g(\vartheta)$*.*

Die Eigenschaft der Invarianz der Maximum-Likelihood-Methode, die in Satz 5.1 für bijektive Abbildungen ausgedrückt ist, lässt sich nun in ähnlicher Weise auch allgemeiner für beliebige Abbildungen formulieren. Dies ist in dem folgenden Lemma ausgedrückt.

Lemma 5.2 ([16], Lemma 4.1.5, [40], 2.1.) *Sei* Θ *der Parameterraum in einem statistischen Modell und sei* $g : \Theta \to \Theta^*$ *eine beliebige Abbildung. Ist* $\hat{\vartheta}$ *ML-Schätzer für* $\vartheta \in \Theta$*, so ist* $g(\hat{\vartheta})$ *ML-Schätzer für* $g(\vartheta)$*.*

Wir betrachten jetzt bei dem Modell von n Beobachtungen von unabhängigen, identisch multivariat normalverteilten Zufallsvektoren $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(i = 1, \ldots, n)$ die Abbildungen $\boldsymbol{f}$ mit $\boldsymbol{f}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\Lambda}$ und $\boldsymbol{g}$ mit $\boldsymbol{g}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\Gamma}$, wobei $\boldsymbol{\Lambda}$ hier wieder die Diagonalmatrix der geordneten Eigenwerte von $\boldsymbol{\Sigma}$ und $\boldsymbol{\Gamma}$ die Matrix sei, deren Spaltenvektoren die normierten Eigenvektoren von $\boldsymbol{\Sigma}$ sind. Existiert ein ML-Schätzer $(\bar{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}})$ für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (bei $n - 1 \geq p$), so ist nach Lemma 5.2 $\boldsymbol{f}(\bar{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}})$ und $\boldsymbol{g}(\bar{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}})$ ML-Schätzer für $\boldsymbol{\Lambda}$ und $\boldsymbol{\Gamma}$. Dies sind gerade die aus den Eigenwerten bzw. Eigenvektoren von $\hat{\boldsymbol{\Sigma}}$ gebildeten Matrizen $\boldsymbol{L}$ bzw. $\boldsymbol{G}$.

5.3.2 ML-Schätzung bei singulärer Stichprobenkovarianzmatrix

Wir betrachten weiterhin das Modell von n Beobachtungen von unabhängigen, identisch verteilten (u. i. v.) Zufallsvektoren $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(i = 1, \ldots, n)$. Im Fall $n - 1 < p$ existiert kein ML-Schätzer für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Mit $\lambda_1 \geq \ldots \geq \lambda_p > 0$ seien die Eigenwerte von $\boldsymbol{\Sigma}$ bezeichnet. Durch eine Vorzeichenregelung seien die Eigenvektoren zu einfach vorkommenden Eigenwerten jeweils eindeutig bestimmt. Bei mehrfachen Eigenwerten sei durch zusätzliche Bedingungen stets eine eindeutige Zuordnung der zugehörigen Eigenvektoren sichergestellt[1]. Mit $\boldsymbol{\Gamma}$ sei jeweils die

[1]Eine eindeutige Wahl lässt sich etwa dadurch erreichen, dass die frei wählbaren Komponenten einfach $1/p$ gesetzt werden. Die erhaltenen Vektoren lassen sich dann untereinander lexikographisch ordnen.

(eindeutig bestimmte) Matrix bezeichnet, deren Spaltenvektoren diese normierten Eigenvektoren von $\boldsymbol{\Sigma}$ seien; $\mathcal{G}$ sei die Menge aller möglichen, aus normierten Eigenvektoren gebildeten Matrizen $\boldsymbol{\Gamma}$, $\mathcal{L}$ sei die Menge aller Diagonalmatrizen $\boldsymbol{\Lambda}$ mit geordneten Diagonalelementen $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Analog seien jeweils $l_1 \geq \ldots \geq l_p \geq 0$ die Eigenwerte und entsprechend $\boldsymbol{G}$ die Matrix der auf gleiche Weise eindeutig bestimmten zugehörigen normierten Eigenvektoren von $\tilde{\boldsymbol{S}} = \frac{1}{n}\boldsymbol{A}$.

Es kann nun gezeigt werden, dass im Falle der Nichtexistenz von ML-Schätzern für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ auch kein ML-Schätzer für $\boldsymbol{\Lambda}$ existiert. Dies ist die Aussage des folgenden Satzes.

Satz 5.4 *Es gelten die gleichen Voraussetzungen und Bezeichnungen wie bisher. Ist die Anzahl der Beobachtungen n nicht größer als die Anzahl der Variablen p, dann gibt es in der Menge $\mathcal{L}$ keinen ML-Schätzer für $\boldsymbol{\Lambda}$.*

Beweis: Die Abbildung $\boldsymbol{f} : \mathbb{R}^p \times \mathrm{PD}(p) \to \mathcal{L}$ sei wieder durch $\boldsymbol{f(\mu, \Sigma) = \Lambda}$ gegeben. Sei $\boldsymbol{\Lambda} \in \mathcal{L}$. Die Menge $C(\boldsymbol{\Lambda})$ ist dann gegeben durch

$$C(\boldsymbol{\Lambda}) = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathrm{PD}(p) | \boldsymbol{f(\mu, \Sigma) = \Lambda}\}. \tag{5.53}$$

Mit $\boldsymbol{\Sigma = \Gamma \Lambda \Gamma'}$ ist die induzierte Likelihood gegeben durch
$M(\boldsymbol{\Lambda}) = \sup_{(\boldsymbol{\mu}, \boldsymbol{\Gamma})} L(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda})$. Zu gegebenem $\boldsymbol{\Lambda}$ ist (mit $\boldsymbol{x} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$) dazu die Likelihood-Funktion

$$L((\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{1}{2}pn} |\boldsymbol{\Sigma}|^{\frac{1}{2}n}} \exp\left(-\frac{1}{2} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) \right) \tag{5.54}$$

bezüglich aller $(\boldsymbol{\mu}, \boldsymbol{\Gamma}) \in \mathbb{R}^p \times \mathcal{G}$ zu maximieren. Nach Satz 5.2 wird das Maximum von L bei $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ und $\boldsymbol{\Gamma} = \boldsymbol{GP}$ mit den dort angegebenen Eigenschaften angenommen. Wie im Beweis von Satz 5.2 gezeigt, ist es (bei $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$) für die Maximierung der Log-Likelihood-Funktion äquivalent, den Ausdruck

$$\ln |\boldsymbol{\Lambda}| + \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{l_i}{\lambda_j} p_{ij}^2 \tag{5.55}$$

zu minimieren. Bei optimaler Wahl von $\boldsymbol{\Gamma}$ erhält man (mit Lemma 5.1) für Ausdruck (5.55) den Wert

$$\ln |\boldsymbol{\Lambda}| + \sum_{i=1}^{p} \frac{l_i}{\lambda_i}, \tag{5.56}$$

der dann unabhängig von $\boldsymbol{\Gamma}$ ist. Ein ML-Schätzer für $\boldsymbol{\Lambda}$ müsste also bezüglich aller $\boldsymbol{\Lambda} \in \mathcal{L}$ Ausdruck (5.56) minimieren. Nach der Voraussetzung $n \leq p$ ist mindestens ein l_i gleich Null. Nach dem Beweis von Satz 5.3 existiert solch ein optimales $\boldsymbol{\Lambda}$ nicht, womit die Aussage des Satzes gezeigt ist. □

Es folgt nun eine Aussage über ML-Schätzer für $\boldsymbol{\Gamma}$. Solche existieren auch bei singulärer Stichprobenkovarianzmatrix.

Satz 5.5 *Es gelten die gleichen Voraussetzungen und Bezeichnungen wie bisher. Die Matrix* $\tilde{\boldsymbol{S}} = \frac{1}{n}\boldsymbol{A}$ *sei singulär, damit sei mindestens ein Eigenwert von* $\tilde{\boldsymbol{S}}$ *Null. Die ML-Schätzer sind dann von der Form* $\boldsymbol{GP}$*, wobei* $\boldsymbol{P}$ *eine orthogonale* $p \times p$*-Matrix ist, die die folgende Bedingung erfüllt.*

$$\exists k : \sum_{i:l_i=0} p_{ik}^2 = 1. \tag{5.57}$$

Beweis: Die Abbildung $\boldsymbol{g} : \mathbb{R}^p \times \mathrm{PD}(p) \to \mathcal{G}$ sei wieder durch $\boldsymbol{g(\mu, \Sigma) = \Gamma}$ gegeben. Für ein $\boldsymbol{\Gamma} \in \mathcal{G}$ ist die Menge $C(\boldsymbol{\Gamma})$ dann gegeben durch

$$C(\boldsymbol{\Gamma}) = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathrm{PD}(p) | \boldsymbol{g(\mu, \Sigma) = \Gamma}\}. \tag{5.58}$$

Mit $\boldsymbol{x} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ ist die Likelihood-Funktion wieder gegeben durch

$$L((\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{1}{2}pn} |\boldsymbol{\Sigma}|^{\frac{1}{2}n}} \exp\left(-\frac{1}{2} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) \right). \tag{5.59}$$

Für ein $\boldsymbol{\Gamma} \in \mathcal{G}$ ist damit die induzierte Likelihood $M(\boldsymbol{\Gamma}, \boldsymbol{x})$ wie folgt definiert:

$$M(\boldsymbol{\Gamma}, \boldsymbol{x}) = \sup_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in C(\boldsymbol{\Gamma})} L((\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{x}). \tag{5.60}$$

Statt der Likelihood-Funktion wird wieder die Log-Likelihood-Funktion betrachtet. Mit der optimalen Wahl $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ erhält man — mit $\boldsymbol{P := G'\Gamma}$ — mit gleicher Argumentation den Ausdruck

$$-\ln|\boldsymbol{\Lambda}| - \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{l_i}{\lambda_j} p_{ij}^2, \tag{5.61}$$

der nun zu maximieren ist. Ausdruck (5.61) ist eine Funktion der Eigenwerte $\lambda_1, \ldots, \lambda_p$:

$$\tilde{l}(\lambda_1, \ldots, \lambda_p) = -\sum_{j=1}^{p} \ln(\lambda_j) - \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{l_i}{\lambda_j} p_{ij}^2. \tag{5.62}$$

Sei nun eine Orthogonalmatrix $\boldsymbol{P}$ so bestimmt, dass für den Index k die Bedingung (5.57) erfüllt sei. Bedingung (5.57) ist dann äquivalent zu

$$\sum_{i:l_i>0} p_{ik}^2 = 0. \tag{5.63}$$

Läuft nun λ_k gegen Null, wobei die restlichen λ_j größer als 0 seien, so gilt

$$\lim_{\lambda_k \to 0} \sum_{i=1}^{p} \frac{l_i}{\lambda_j} p_{ij}^2 = \lim_{\lambda_k \to 0} \sum_{i:l_i=0} \frac{l_i}{\lambda_j} p_{ij}^2 = 0 \tag{5.64}$$

sowie

$$\lim_{\lambda_k \to 0} \sum_{j=1}^{p} \ln(\lambda_j) = -\infty. \tag{5.65}$$

Damit gilt für $\tilde{l}$

$$\lim_{\lambda_k \to 0} \tilde{l}(\lambda_1, \ldots, \lambda_p) = \infty. \tag{5.66}$$

Somit hat für $\hat{\mathbf{\Gamma}} = \boldsymbol{GP}$ das Supremum $\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in C(\hat{\mathbf{\Gamma}})} L((\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{x})$ hier den Wert ∞, damit ist $\hat{\mathbf{\Gamma}}$ ML-Schätzer für $\mathbf{\Gamma}$.

Sei nun die Bedingung (5.57) für $\boldsymbol{P}$ nicht erfüllt. Wir betrachten jetzt den Ausdruck

$$\tilde{l}(\lambda_1, \ldots, \lambda_p) = \sum_{j=1}^{p} \ln(\lambda_j) \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{l_i}{\lambda_j} p_{ij}^2, \tag{5.67}$$

der nun bezüglich der λ_j zu minimieren ist. Die Funktion $\tilde{l}$ ist für $\lambda_j > 0$ differenzierbar, die partiellen Ableitungen sind jeweils

$$\frac{\partial \tilde{l}}{\partial \lambda_j} = \frac{1}{\lambda_j} - \frac{1}{\lambda_j^2} \sum_{i=1}^{p} l_i p_{ij}^2. \tag{5.68}$$

Setzt man nun $\frac{\partial \tilde{l}}{\partial \lambda_j^*} = 0$, so erhält man

$$\lambda_j^* = \sum_{i=1}^{p} l_i p_{ij}^2. \tag{5.69}$$

Nach Voraussetzung, dass (5.57) nicht gilt, gilt hier stets $\lambda_j^* > 0$. Für die zweite partielle Ableitung an der Stelle λ_j^* gilt

$$\begin{aligned}
\left.\frac{\partial^2 \tilde{l}}{\partial \lambda_j^2}\right|_{\lambda_j = \lambda_j^*} &= -\frac{1}{\lambda_j^{*2}} + 2\frac{1}{\lambda_j^{*3}} \sum_{i=1}^{p} l_i p_{ij}^2 && (5.70)\\
&= -\frac{1}{\left(\sum_{i=1}^{p} l_i p_{ij}^2\right)^2} + \frac{2}{\left(\sum_{i=1}^{p} l_i p_{ij}^2\right)^2} && (5.71)\\
&= \frac{1}{\left(\sum_{i=1}^{p} l_i p_{ij}^2\right)^2} && (5.72)\\
&> 0. && (5.73)
\end{aligned}$$

Somit hat nach der Extremwerttheorie der Differentialrechnung $\tilde{l}$ an der Stelle λ_j^* ein Minimum, d. h., λ_j^* maximiert die Likelihood-Funktion.

Sei $\mathbf{\Lambda^*}$ die Diagonalmatrix, deren Diagonalelemente jeweils nach Gleichung (5.69) gebildet werden. Da alle Diagonalelemente positiv sind, ist der erhaltene Wert der Likelihood-Funktion endlich, d. h. kleiner als im ersten Fall bei $\lambda_k \to 0$ und damit nicht maximal. □

5.4 Schlussfolgerungen

Aus den Ergebnissen der Betrachtungen zur Maximum-Likelihood-Schätzung in diesem Kapitel lässt sich nun als Fazit formulieren:

1. Bei unbekanntem $\boldsymbol{\mu}$ maximiert in jedem Fall die Wahl $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ die Likelihood-Funktion (bzw. induzierte Likelihood).

2. Im Fall $n \leq p$ gibt es (bei unbekanntem $\boldsymbol{\Lambda}$) keine optimale Diagonalmatrix $\hat{\boldsymbol{\Lambda}}$ mit ausschließlich positiven Diagonalelementen[2], die die Likelihood- bzw. induzierte Likelihoodfunktion maximieren würde. Damit ist hier gezeigt, dass bei singulärer Stichprobenkovarianzmatrix die Schätzung der Eigenwerte das eigentliche Problem darstellt.

3. In jedem Fall führt (bei unbekanntem $\boldsymbol{\Gamma}$) die Wahl $\hat{\boldsymbol{\Gamma}} = \boldsymbol{G}$ zu einer Maximierung der Likelihood- bzw. induzierten Likelihoodfunktion (bei Nichteindeutigkeit ist dies unabhängig von der Wahl von $\boldsymbol{G}$).

Das zeigt, dass der Maximum-Likelihood-Ansatz im Fall $n - 1 < p$ nur für die Schätzung von $\boldsymbol{\mu}$ und $\boldsymbol{\Gamma}$ anwendbar ist. Für die Schätzung von $\boldsymbol{\Lambda}$ ist man da auf andere Ansätze angewiesen. Beides lässt sich dann aber miteinander kombinieren: Man bestimmt zunächst eine „geeignete“ Schätzung für $\boldsymbol{\Lambda}$. Zu gegebenem $\boldsymbol{\Lambda^*}$ mit positiven Diagonalelementen wird in jedem Fall der Wert der induzierten Likelihood $M(\boldsymbol{\Lambda^*}) = \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in C(\boldsymbol{\Lambda}^*)} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ erreicht durch $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\bar{\boldsymbol{x}}, \boldsymbol{G\Lambda^* G'})$, was insgesamt dann die Schätzung für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ liefert.

[2]Eine Lebesgue-Dichte im $\mathbb{R}^p$ ist nur im Fall von Diagonalmatrizen $\hat{\boldsymbol{\Lambda}}$ mit ausschließlich positiven Diagonalelementen definiert; die Maximum-Likelihood-Betrachtung kann sich also auch nur auf diesen Fall beziehen.

Kapitel 6

Stabilität in der Schätztheorie

Für das Multinormalverteilungsmodell sind verschiedene Möglichkeiten zur Schätzung für $\boldsymbol{\Sigma}$ bekannt. Die Schätzer $\tilde{\boldsymbol{S}}$ sind dabei gewöhnlich von einer Matrix der Form $k\boldsymbol{A}$, wobei k eine beliebige positive reelle Zahl ist und $\boldsymbol{A}$ durch $\sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ gegeben ist, abgeleitet, d. h. $\tilde{\boldsymbol{S}} = \boldsymbol{f}(k\boldsymbol{A})$. (Ridge-Schätzungen sind z. B. von der Form $\boldsymbol{S}_{\text{ridge}} = \frac{1}{n-1}\boldsymbol{A} + \boldsymbol{M}$.) Falls $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ nicht bekannt sind, sind die Statistiken der Form $\boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (\bar{\boldsymbol{x}}, k\boldsymbol{A})$ suffizient für $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (d. h., die bedingte Verteilung von $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ für gegebenes $\boldsymbol{T} = (\bar{\boldsymbol{x}}, k\boldsymbol{A})$ ist von $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ unabhängig; vgl. dazu z. B. [3]). Damit enthalten diese Statistiken gewissermaßen die gesamte in der Beobachtung $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ enthaltene Information über den Parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Daher ist es naheliegend, Statistiken der Form $k\boldsymbol{A}$ — in Verbindung mit $\bar{\boldsymbol{x}}$ als Schätzung für $\boldsymbol{\mu}$ — als Grundlage für die Schätzung von $\boldsymbol{\Sigma}$ zu verwenden. Es ist schwierig, allgemein für die Schätzer $\tilde{\boldsymbol{S}} = \boldsymbol{f}(k\boldsymbol{A})$ genaue Aussagen zur Güte der Schätzung zu machen. Die folgenden Untersuchungen sollen sich daher zunächst auf Schätzer der Form $k\boldsymbol{A}$ konzentrieren.

Für die Güte eines Schätzers wäre unter anderem eine Quantifizierung des Schätzfehlers von Bedeutung. Gewöhnlich wird der Schätzfehler mit zunehmender Anzahl der Beobachtungen immer kleiner. Ist ein Schätzer konsistent, so strebt er bei festem Parameter für $n \to \infty$ gegen den zu schätzenden Parameter; der Schätzfehler strebt dann gegen 0. Um solcherlei asymptotische Betrachtungen soll es hier aber nicht gehen, da hier in besonderer Weise die Situation, dass n im Verhältnis zur Merkmalsanzahl p klein ist, betrachtet wird. Oft ist der Fehler der Schätzung auch von der Varianz der zugrundeliegenden Verteilung der Beobachtungen selbst abhängig. In vielen Fällen kann er bei fester Anzahl der Beobachtungen beliebig klein gehalten werden, wenn diese hinreichend klein wird.

Die Schätzer für die Kovarianzmatrix $\boldsymbol{\Sigma}$ von der Form $k\boldsymbol{A}$ sind im Fall $n - 1 < p$ singuläre Matrizen. Was lässt sich in diesem Fall über den Schätzfehler aussagen? Kann er — bei festem n — auch beliebig klein werden? D. h., kann der Abstand einer singulären Matrix zum wahren Parameter, der regulären Matrix $\boldsymbol{\Sigma}$, beliebig klein werden, oder gibt es eine untere Schranke für den Mindestabstand?

Diese Frage soll in diesem Kapitel im Mittelpunkt der Untersuchungen stehen.

6.1 Der Stabilitätsbegriff in der Schätztheorie

Für die nachfolgenden Betrachtungen sollen im Folgenden einige Begriffe und Bezeichnungen neu eingeführt oder aber präziser gefasst, gegebenenfalls auch etwas erweitert oder modifiziert werden.

Die folgenden Betrachtungen zur Stabilität, insbesondere die Begriffsdefinitionen der verschiedenen Arten von Stabilität und der Stabilitätszahl, sind von Betrachtungen aus dem Bereich der Numerischen Mathematik abgeleitet. Um die Verwendbarkeit von numerischen Verfahren zur Lösung von gewöhnlichen Differentialgleichungen beurteilen zu können, werden dort in ähnlicher Weise verschiedene Arten von Stabilität verwendet [11].

Varianzstabilität charakterisiert hier zunächst eine Abhängigkeit des Fehlers von der Varianz der Verteilung einer Einzelbeobachtung P_ϑ. Bei hinreichend kleiner Varianz kann der mittlere Fehler — das Risiko — des Schätzers auch beliebig klein werden. Es ist damit aber noch nicht gesagt, wie stark der mittlere Fehler mit größer werdender Varianz ansteigt; er kann auch beliebig groß werden. Existiert eine Stabilitätszahl, so liegt ein linearer Zusammenhang vor. Die Stabilitätszahl gibt an, wie stark die Änderung des mittleren Fehlers des Schätzers bezüglich der Änderung der Varianz $\mathrm{Var}[P_\vartheta]$ ist.

Das Konzept der P-Stabilität hat Ähnlichkeit mit der Betrachtung von Konfidenzbereichen, doch geht es hier um einen anderen Sachverhalt. Für einen Konfidenzbereich kann gesagt werden, dass er mit einer gewissen Wahrscheinlichkeit den zu schätzenden Wert $g(\vartheta)$ enthält. Im Falle der P-Stabilität hingegen ist abgesichert, dass der Fehler der Schätzung mit positiver Wahrscheinlichkeit unter einem beliebig gewählten positiven Wert ε liegt, also mit positiver Wahrscheinlichkeit beliebig klein werden kann.

Gegeben sei nun ein statistisches Modell $(M, \mathcal{A}, (P_\vartheta)_{\vartheta\in\Theta})$ und eine Abbildung $g : \Theta \to \Gamma$, $\delta : M \to \Gamma$ sei ein Schätzer für $g(\vartheta) \in \Gamma$. Eine Funktion $L : \Theta \times \Gamma \to [0, \infty)$ wird hier als Verlustfunktion bezeichnet, wenn

$$L(\vartheta, g(\vartheta)) = 0 \tag{6.1}$$

gilt. Hierbei ist $L(\vartheta, \delta(x))$ der Fehler der Schätzung von $g(\vartheta)$. Zu gegebener Verlustfunktion L ist die Risikofunktion $R(\vartheta, \delta)$ wieder gegeben durch

$$R(\vartheta, \delta) = \mathrm{E}_\vartheta L(\vartheta, \delta(x)). \tag{6.2}$$

Für die folgende Definition betrachten wir die Situation von Beobachtungen von n u. i. v. Zufallsvariablen, d. h. statistische Modelle der Form

$$\left(M^n, \mathcal{A}^n, \left(\bigotimes_{\alpha=1}^{n} P_\vartheta\right)_{\vartheta\in\Theta}\right). \tag{6.3}$$

Mit $\mathrm{E}[P_\vartheta]$ sei der Erwartungswert und mit $\mathrm{Var}[P_\vartheta]$ (bzw. $\mathbf{Cov}[P_\vartheta]$) die Varianz (bzw. Kovarianzmatrix) der Verteilung P_ϑ einer Einzelbeobachtung, jeweils mit Parameter $\vartheta \in \Theta$, bezeichnet.

Definition 6.1 (Varianz-Stabilität) *Gegeben sei das durch Ausdruck (6.3) definierte statistische Modell, eine Abbildung $g : \Theta \to \Gamma$ und ein (messbarer) Schätzer $\delta : (M^n, \mathcal{A}^n) \to (\Gamma, \mathcal{C})$. Mit $L : \Theta \times \Gamma \to [0, \infty)$ sei eine Verlustfunktion mit zugehöriger Risikofunktion $R(\vartheta, \delta) = \mathrm{E}_\vartheta L(\vartheta, \delta(x))$ bezeichnet. Es sei $(M, \mathcal{A}) \subset (\mathbb{R}, \mathcal{B})$, der Erwartungswert $\mathrm{E}[P_\vartheta]$ möge existieren und sei endlich für alle $\vartheta \in \Theta$. Der Schätzer δ heißt für $g(\vartheta)$ varianzstabil oder V-stabil bezüglich R, wenn es zu jedem $\varepsilon > 0$ ein $\kappa > 0$ gibt, so dass für alle $\vartheta \in \Theta$, für die $\mathrm{Var}[P_\vartheta] < \kappa$ gilt, die Bedingung $R(\vartheta, \delta) < \varepsilon$ erfüllt ist. Allgemeiner wird für den mehrdimensionalen Fall $(M, \mathcal{A}) \subset (\mathbb{R}^k, \mathcal{B}^k)$ mit existierender Kovarianzmatrix $\mathbf{Cov}[P_\vartheta]$ gesagt, $\boldsymbol{\delta}$ heißt für $g(\vartheta)$ varianzstabil oder V-stabil bezüglich R, wenn es zu jedem $\varepsilon > 0$ ein $\kappa > 0$ gibt, so dass für alle $\vartheta \in \Theta$ aus $\|\mathbf{Cov}[P_\vartheta]\| < \kappa$ die Bedingung $R(\vartheta, \boldsymbol{\delta}) < \varepsilon$ folgt.*

Diese Begriffsdefinition erlaubt zunächst eine qualitative Unterscheidung, ob überhaupt Stabilität vorliegt oder nicht. Um das Maß an Stabilität quantifizieren zu können, soll zusätzlich eine Stabilitätszahl definiert werden.

Definition 6.2 (Stabilitätszahl) *Es gelten die gleichen Bezeichnungen und Voraussetzungen wie bisher. Mit $s(\delta, R)$ sei das Infimum aller Konstanten $c(\delta, R)$ bezeichnet, für die*

$$R(\vartheta, \delta) \leq c(\delta, R)\mathrm{Var}[P_\vartheta] \tag{6.4}$$

bzw. für den mehrdimensionalen Fall

$$R(\vartheta, \boldsymbol{\delta}) \leq c(\boldsymbol{\delta}, R)\|\mathbf{Cov}[P_\vartheta]\| \tag{6.5}$$

für alle $\vartheta \in \Theta$ gilt. Die Zahl $s(\delta, R)$ heißt Stabilitätszahl von δ.

Es folgt ein weiterer Stabilitätsbegriff; hier werden wieder ganz allgemeine statistische Modelle betrachtet.

Definition 6.3 (P-Stabilität) *Gegeben sei ein statistisches Modell $(M, \mathcal{A}, (P_\vartheta)_{\vartheta \in \Theta})$, eine Abbildung $g : \Theta \to \Gamma$ und ein (messbarer) Schätzer $\delta : (M, \mathcal{A}) \to (\Gamma, \mathcal{C})$. Mit $L : \Theta \times \Gamma \to [0, \infty)$ sei eine Verlustfunktion bezeichnet. Der Schätzer δ für $g(\vartheta)$ heißt P-stabil bezüglich L, wenn für alle $\vartheta \in \Theta$ und zu jedem $\varepsilon > 0$*

$$P_\vartheta(\{L(\vartheta, \delta) < \varepsilon\}) > 0 \tag{6.6}$$

gilt.

Beispiel 6.1 (Univariate Normalverteilung) *Seien $x_1, \ldots, x_n$ Beobachtungen von unabhängigen, identisch univariat normalverteilten Zufallsvariablen, so wird das statistische Modell*

$$\left(\mathbb{R}^n, \mathcal{B}^n, \left(\bigotimes_{\alpha=1}^{n} N(\beta, \sigma^2)\right)_{(\beta,\sigma^2)\in\mathbb{R}\times(0,\infty)}\right)$$

verwendet. Als Schätzer für $g_1(\beta, \sigma^2) = \beta$ und $g_2(\beta, \sigma^2) = \sigma^2$ sollen $\delta_1(x) = \bar{x} = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha$ bzw. $\delta_2(x) = \frac{1}{n-1}\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2$ für $n > 1$ betrachtet werden. Verwendet man als Risikofunktion den mittleren quadratischen Fehler, d. h. $R(\vartheta, \delta) = \mathrm{E}_\vartheta(\delta(x) - g(\vartheta))^2$, so gilt für δ_1:

$$R((\beta, \sigma^2), \delta_1) = \mathrm{E}_{\beta,\sigma^2}(\bar{x} - \beta)^2 = \mathrm{Var}_{\beta,\sigma^2}\bar{x} = \frac{1}{n}\sigma^2.$$

Zu jedem $\varepsilon > 0$ lässt sich κ durch $\kappa := n\varepsilon$ wählen, so dass $\mathrm{Var}_{\beta,\sigma^2}\bar{x} < \varepsilon$ gilt, falls $\mathrm{Var}[N(\beta, \sigma^2)] < \kappa = n\varepsilon$ ist. Damit ist $\delta_1(x) = \bar{x}$ varianzstabiler Schätzer für β. Da hier $R((\beta, \sigma^2), \delta_1) = \frac{1}{n}\mathrm{Var}[P_\vartheta]$ gilt, hat δ_1 die Stabilitätszahl $\frac{1}{n}$.

Die Statistik $\frac{1}{\sigma^2}\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2$ hat eine χ^2_{n-1}-Verteilung [28]. Verwendet man auch für δ_2 als Risikofunktion den mittleren quadratischen Fehler, so erhält man mit $\mathrm{Var}[\chi^2_m] = 2m$:

$$\begin{aligned} R((\beta, \sigma^2), \delta_2) &= \mathrm{E}_{\beta,\sigma^2}\left(\frac{1}{n-1}\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2 - \sigma^2\right)^2 \\ &= \mathrm{Var}(\delta_2) \\ &= \left(\frac{\sigma^2}{n-1}\right)^2 \mathrm{Var}(\chi^2_{n-1}) \\ &= \frac{2\sigma^4}{n-1}. \end{aligned}$$

Für jedes $\varepsilon > 0$ lässt sich dann jeweils ein $\kappa > 0$ gemäß $\kappa := \sqrt{\frac{n-1}{2}\varepsilon}$ wählen. Für $\mathrm{Var}[N(\beta, \sigma^2)] < \kappa = \sqrt{\frac{n-1}{2}\varepsilon}$ gilt dann $\sigma^2 < \sqrt{\frac{n-1}{2}\varepsilon}$, damit $\sigma^4 < \frac{n-1}{2}\varepsilon$, d. h. $R((\beta, \sigma^2), \delta_2) = \frac{2\sigma^4}{n-1} < \varepsilon$. Es lässt sich aber keine reelle Zahl $c(\delta_2, R) > 0$ angeben, so dass für alle $\sigma^2 > 0$ $\frac{2\sigma^4}{n-1} \le c(\delta_2, R)\sigma^2$ gilt. In diesem Fall ist δ_2 also varianzstabil, hat aber keine Stabilitätszahl.

Es soll jetzt der besondere Fall $n = 1$ mit dem Schätzer $\delta_2' = \frac{1}{n}\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2$ für $g_2 = \sigma^2$ betrachtet werden. Dann gilt bei unbekanntem β $\delta_2'(x) = (x - \bar{x})^2 = (x - x)^2 = 0$. Mit gleicher Risikofunktion erhält man:

$$R((\beta, \sigma^2), \delta_2') = \mathrm{E}_{\beta,\sigma^2}(0 - \sigma^2)^2 = \sigma^4.$$

In diesem Fall lässt sich zu jedem $\varepsilon > 0$ κ gemäß $\kappa := \sqrt{\varepsilon}$ wählen. Falls $\mathrm{Var}[N(\beta, \sigma^2)] < \kappa = \sqrt{\varepsilon}$ gilt, folgt $(\sigma^2)^2 < \varepsilon$, d. h. $R((\beta, \sigma^2), \delta_2') < \varepsilon$. Damit

ist δ_2' für diesen Fall ebenfalls varianzstabil. Der Schätzfehler beträgt bei Verwendung von quadratischer Verlustfunktion

$$L((\beta, \sigma^2), \delta_2'(x)) = (\delta_2'(x) - \sigma^2)^2 = \sigma^4 \ \forall x \in \mathbb{R},$$

d. h.,

$$P(\{L((\beta, \sigma^2), \delta_2') = \sigma^4\}) = 1.$$

Wird ein $\varepsilon' > 0$ mit $\varepsilon' \leq \sigma^4$ gewählt, so ist die Wahrscheinlichkeit für $\{L((\beta, \sigma^2), \delta_2') < \varepsilon'\}$ gleich Null. Damit ist δ_2' für den Fall $n = 1$ nicht P-stabil.

Das Beispiel zeigt, dass ein varianzstabiler Schätzer nicht notwendig P-stabil sein muss. Die P-Stabilität erweist sich hier als die stärkere Bedingung. Für den Fall $n = 1$ liefert der Schätzer δ_2' praktisch keine Information über σ^2 und ist somit eigentlich unbrauchbar, obwohl er varianzstabil ist.

6.2 Grundlagen aus der Matrizentheorie

Die folgenden Betrachtungen sollen sich auch wieder auf das Modell einer multivariaten Normalverteilung beziehen. Von besonderem Interesse sind hierbei sicher Stabilitätsaussagen für den Schätzer der Kovarianzmatrix. Zunächst sollen im Folgenden einige wichtige Aussagen aus der Matrizentheorie angegeben werden, die für die Herleitung solcher Aussagen benötigt werden.

6.2.1 Der Raum der reellen $p \times q$-Matrizen als normierter Raum

Der Raum $\mathcal{M}(p, q)$ der reellen $p \times q$-Matrizen ist ein Vektorraum über $\mathbb{R}$. Sind $\|.\|_1$ und $\|.\|_2$ Normen auf $\mathbb{R}^q$ bzw. $\mathbb{R}^p$, so lässt sich auf $\mathcal{M}(p, q)$ eine Matrixnorm definieren durch

$$\|\boldsymbol{A}\| := \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_1} = \sup_{\|\boldsymbol{x}\|_1 = 1} \|\boldsymbol{A}\boldsymbol{x}\|_2. \tag{6.7}$$

Damit kann man $(\mathcal{M}(p, q), \|.\|)$ als normierten Raum auffassen.

Es lässt sich nun leicht zeigen, dass für so definierte Matrixnormen die Abschätzung

$$\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\| \tag{6.8}$$

für alle $\boldsymbol{A} \in \mathcal{M}(p, q)$ und $\boldsymbol{x} \in \mathbb{R}^q$ gilt (vgl. z. B. [20]). Für $\boldsymbol{x} = \boldsymbol{0}$ ist dies klar. Ist $\boldsymbol{x} \neq \boldsymbol{0}$, so gilt nach der Definition (6.7)

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|\boldsymbol{A}\|, \tag{6.9}$$

daraus folgt sofort (6.8).

Weiterhin gilt für solche Matrixnormen auch die Abschätzung

$$\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\| \tag{6.10}$$

(vgl. z. B. [20]).

Verwendet man speziell die euklidische Norm als Vektornorm, so erhält man für quadratische Matrizen ($p \times p$) die Norm

$$\begin{aligned} \|\boldsymbol{A}\| &= \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} && (6.11)\\ &= \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\sqrt{(\boldsymbol{Ax})'\boldsymbol{Ax}}}{\sqrt{\boldsymbol{x'x}}} && (6.12)\\ &= \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \sqrt{\frac{\boldsymbol{x'A'Ax}}{\boldsymbol{x'x}}}. && (6.13) \end{aligned}$$

Es kann gezeigt werden, dass in diesem Fall

$$\|\boldsymbol{A}\| = \sqrt{\mu_1} \tag{6.14}$$

gilt, wobei μ_1 der größte Eigenwert von $\boldsymbol{A'A}$ ist (vgl. z. B. [46]). Diese Matrixnorm heißt *Spektralnorm*. Für symmetrische Matrizen gilt in diesem Fall speziell

$$\|\boldsymbol{A}\| = |\lambda_1|, \tag{6.15}$$

wobei λ_1 der betragsmäßig größte Eigenwert von $\boldsymbol{A}$ ist.

6.2.2 Die Norm als Verlustfunktion

Hat man nun ein statistisches Modell $(M, \mathcal{A}, (P_\vartheta)_{\vartheta\in\Theta})$, einen normierten Raum $(\Omega, \|.\|)$ und eine Abbildung $g : \Theta \to \Gamma$ mit $\Gamma \subset \Omega$, so lässt sich für einen Schätzer $\delta : M \to \Gamma$ mit der Norm auf Ω eine Verlustfunktion durch

$$L(\vartheta, \delta(x)) := \|\delta(x) - g(\vartheta)\| \tag{6.16}$$

definieren. Ausdruck (6.16) genügt tatsächlich Bedingung (6.1), denn mit den Normeigenschaften gilt $L(\vartheta, g(\vartheta)) = \|g(\vartheta) - g(\vartheta)\| = 0$.

6.2.3 Positiv semidefinite Matrizen

Mit PSD(p) sei jetzt die Menge der positiv semidefiniten reellen $p \times p$-Matrizen bezeichnet. (Eine Matrix $\boldsymbol{A}$ ($p \times p$, symmetrisch) wird hier als *positiv semidefinit* bezeichnet, wenn $\boldsymbol{x'Ax} \geq 0$ für alle $\boldsymbol{x} \in \mathbb{R}^p$ gilt.) Die Definitionen von positiv definiten und positiv semidefiniten Matrizen ließen sich auf beliebige quadratische reelle Matrizen anwenden. Es ist aber üblich, sie allein für symmetrische

Matrizen zu verwenden. Wenn hier eine Matrix als positiv definit bzw. als positiv semidefinit bezeichnet wird, sei stets vorausgesetzt, dass eine symmetrische Matrix gemeint ist.

Die Menge der symmetrischen $p \times p$-Matrizen bildet einen Vektorraum und ist damit ein Unterraum von $\mathcal{M}(p,p)$. Dieser Vektorraum sei hier mit $\mathcal{S}(p)$ bezeichnet. Im Raum $\mathbb{R}^p$ wird hier stets das übliche Skalarprodukt

$$(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x'y} = \sum_{i=1}^{p} x_i y_i \tag{6.17}$$

und die damit gegebene euklidische Norm

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} = \sqrt{\sum_{i=1}^{p} x_i^2} \tag{6.18}$$

verwendet. Mit dieser Norm auf $\mathbb{R}^p$ ist zugleich durch Gleichung (6.7) die Spektralnorm (6.15) gegeben. Zusammen mit dieser Norm wird hier im Folgenden auch der Vektorraum $\mathcal{S}(p)$ als normierter Raum betrachtet.

Die Menge der positiv semidefiniten reellen $p \times p$-Matrizen $\mathrm{PSD}(p)$ wird in den folgenden Betrachtungen stets als eine Teilmenge des auf diese Weise gegebenen normierten Raumes $\mathcal{S}(p)$ angesehen. Aus der Definition ist sofort zu sehen, dass $\mathrm{PSD}(p)$ die Menge der positiv definiten Matrizen $\mathrm{PD}(p)$ enthält. Darüber hinaus kann gezeigt werden, dass $\mathrm{PSD}(p)$ der Abschluss von $\mathrm{PD}(p)$ ist:

Satz 6.1 $\overline{\mathrm{PD}(p)} = \mathrm{PSD}(p)$

Beweis: Es ist zunächst zu zeigen, dass alle Elemente aus $\mathrm{PSD}(p)$ auch Berührungspunkte von $\mathrm{PD}(p)$ sind. Jede beliebige Matrix $\boldsymbol{A} \in \mathrm{PSD}(p)$ lässt sich darstellen als

$$\boldsymbol{A} = \lim_{\varepsilon \to 0, \varepsilon > 0} (\boldsymbol{A} + \varepsilon \boldsymbol{I}) \tag{6.19}$$

mit Einheitsmatrix $\boldsymbol{I}$. Die Matrix $\boldsymbol{A} + \varepsilon \boldsymbol{I}$ ist für jedes $\varepsilon > 0$ positiv definit. Damit ist in jeder Umgebung von $\boldsymbol{A}$ ein Element aus $\mathrm{PD}(p)$ enthalten, also sind alle Elemente von $\mathrm{PSD}(p)$ Berührungspunkte von $\mathrm{PD}(p)$.

Es bleibt zu zeigen: $\overline{\mathrm{PD}(p)} \subset \mathrm{PSD}(p)$. Ist $\boldsymbol{A}$ ein Berührungspunkt von $\mathrm{PD}(p)$, so gibt es eine Folge $(\boldsymbol{A}_n)_{n \in \mathbb{N}} \in \mathrm{PD}(p)$, so dass

$$\lim_{n \to \infty} \boldsymbol{A}_n = \boldsymbol{A} \tag{6.20}$$

gilt. Für ein beliebiges $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x} \neq \boldsymbol{0}$ gilt dann für jedes $n \in \mathbb{N}$ $\boldsymbol{x'A}_n\boldsymbol{x} > 0$. Damit folgt

$$\boldsymbol{x'Ax} = \lim_{n \to \infty} \boldsymbol{x'A}_n\boldsymbol{x} \geq 0. \tag{6.21}$$

Das bedeutet, $\boldsymbol{A}$ ist positiv semidefinit. Dies gilt für beliebige Berührungspunkte von $\mathrm{PD}(p)$. Damit folgt $\overline{\mathrm{PD}(p)} \subset \mathrm{PSD}(p)$. $\square$

6.3 Stabilität der Schätzer bei multivariater Normalverteilung

Wir betrachten jetzt wieder folgende Situation: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ seien n Beobachtungen von unabhängigen, identisch $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Sind $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ unbekannt, hat man wieder das durch Ausdruck (2.2) gegebene statistische Modell. Für den Schätzer $\bar{\boldsymbol{x}}$ für $\boldsymbol{\mu}$ lassen sich die Aussagen von Beispiel 6.1 auf den mehrdimensionalen Fall übertragen. Auch hier gibt es eine quadratische Verlustfunktion, die durch

$$L(\boldsymbol{\vartheta}, \boldsymbol{\delta}(\boldsymbol{x})) = (\boldsymbol{\delta}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\vartheta}))'(\boldsymbol{\delta}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\vartheta})) \tag{6.22}$$

gegeben ist. Verwendet man diese Funktion, so erhält man für $\boldsymbol{\delta}(\boldsymbol{x}) = \bar{\boldsymbol{x}}$ als Schätzer für $\boldsymbol{g}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\mu}$:

$$\begin{aligned}
R(\boldsymbol{\mu}, \bar{\boldsymbol{x}}) &= \mathrm{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) && (6.23)\\
&= \mathrm{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\sum_{i=1}^{p}(\bar{x}_i - \mu_i)^2\right) && (6.24)\\
&= \mathrm{tr}(\boldsymbol{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})') && (6.25)\\
&= \mathrm{tr}(\mathbf{Cov}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\bar{\boldsymbol{x}})) && (6.26)\\
&= \mathrm{tr}\frac{1}{n}\boldsymbol{\Sigma} && (6.27)\\
&= \frac{1}{n}\mathrm{tr}\boldsymbol{\Sigma}. && (6.28)
\end{aligned}$$

Für die Spur einer reellwertigen $p \times p$-Matrix $\boldsymbol{A}$ mit den Eigenwerten $\lambda_1, \ldots, \lambda_p$ gilt allgemein folgende Identität:

$$\sum_{i=1}^{p}\lambda_i = \mathrm{tr}\boldsymbol{A} \tag{6.29}$$

(vgl. dazu z. B. [37], [47]). Ist λ_1 der größte Eigenwert von $\boldsymbol{\Sigma}$, so gilt damit

$$\mathrm{tr}\boldsymbol{\Sigma} \leq p\lambda_1. \tag{6.30}$$

Ist $\boldsymbol{x}_1$ der zu λ_1 gehörige normierte Eigenvektor, so gilt wegen $\|\boldsymbol{x}_1\| = 1$ und $\lambda_1\boldsymbol{x}_1 = \boldsymbol{\Sigma}\boldsymbol{x}_1$

$$\lambda_1 = \|\lambda_1\boldsymbol{x}_1\| = \|\boldsymbol{\Sigma}\boldsymbol{x}_1\| \leq \|\boldsymbol{\Sigma}\|. \tag{6.31}$$

Für die Abschätzung der Risikofunktion erhält man dann

$$R(\boldsymbol{\mu}, \bar{\boldsymbol{x}}) = \frac{1}{n}\mathrm{tr}\boldsymbol{\Sigma} \leq \frac{1}{n}p\lambda_1 \leq \frac{p}{n}\|\boldsymbol{\Sigma}\| = \frac{p}{n}\|\mathbf{Cov}[N(\boldsymbol{\mu}, \boldsymbol{\Sigma})]\|. \tag{6.32}$$

Damit ist zumindest eine obere Schranke für die Stabilitätszahl gegeben; es kann hier festgestellt werden, dass

$$c(\boldsymbol{\delta}, R) \leq \frac{p}{n} \tag{6.33}$$

gilt.

Bemerkung 6.1 *Der in Ungleichung (6.33) ausgedrückte Zusammenhang ließe sich nun in drei verschiedene Richtungen interpretieren:*

1. *Bei festem p und fester Varianz der Einzelbeobachtungen $\mathbf{Cov}[N(\boldsymbol{\mu}, \boldsymbol{\Sigma})]$ sinkt das Risiko der Schätzung mit steigender Anzahl der Beobachtungen n. Dies ist eine wohlbekannte Tatsache, die sofort aus der Konsistenz des Schätzers folgt. Aber um diesen Sachverhalt soll es hier nicht vordergründig gehen.*

2. *Man könnte Ungleichung (6.33) auch dahingehend interpretieren, dass man folgert, dass sich bei konstanter Anzahl der Beobachtungen durch die Erhöhung der Anzahl der beobachteten Merkmale p der Verlust L, der Fehler der Schätzung, erhöhe. Nur sind solcherlei Folgerungen sehr mit Vorsicht zu betrachten. Eine Verlustfunktion, wie auch eine Abstandsfunktion in einem metrischen Raum, ist zunächst nur für einen bestimmten (metrischen) Raum definiert, d. h., Abstandsfunktionen in zwei verschiedenen metrischen Räumen sind zunächst zwei verschiedene Funktionen. Daher lassen sich Abstände von Elementen des einen Raumes nicht ohne weiteres mit Abständen von Elementen eines anderen Raumes vergleichen.*

   Betrachtet man nun etwa den Vektorraum $\mathbb{R}^p$ als Teilmenge des Vektorraumes $\mathbb{R}^{p+k}$, indem man alle $(x_1, \ldots, x_p) \in \mathbb{R}^p$ mit den Vektoren $(x_1, \ldots, x_p, 0, \ldots, 0) \in \mathbb{R}^{p+k}$ identifiziert, so ließe sich hier eine $p+k$-dimensionale Abstandsfunktion, etwa der euklidische Abstand $d_{p+k}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{\alpha=1}^{p+k}(x_\alpha - y_\alpha)^2}$, auf alle Paare von Elementen anwenden. Der $p+k$-dimensionale Abstand $d_{p+k}((x_1, \ldots, x_p, 0, \ldots, 0), (y_1, \ldots, y_p, 0, \ldots, 0))$ wäre dann gleich dem p-dimensionalen euklidischen Abstand $d_p((x_1, \ldots, x_p), (y_1, \ldots, y_p))$ auf $\mathbb{R}^p \times \mathbb{R}^p$. Betrachtet man eine auf diese Weise gegebene Dimensionsänderung, so gilt sicherlich für zwei Vektoren $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{p+k}$ $d_p(\boldsymbol{x}, \boldsymbol{y}) \leq d_{p+k}(\boldsymbol{x}, \boldsymbol{y})$. Durch die Erniedrigung bzw. Erhöhung der Dimension wird der so berechnete Abstand automatisch kleiner bzw. größer. Gleiches gilt ebenso für in ähnlicher Weise definierte mehrdimensionale Abstandsmaße, wie auch die Verlustfunktion (6.22).

3. *Es sei jetzt vorausgesetzt, dass die Merkmalsanzahl p und die Anzahl der Beobachtungen n fest seien. Von Interesse ist nun, wie sich das Risiko der Schätzung von $\boldsymbol{\mu}$ bei Änderung von $\|\boldsymbol{\Sigma}\|$ ändert. Die Abschätzung (6.33) gibt an, dass das Risiko R in Abhängigkeit von $\|\boldsymbol{\Sigma}\|$ durch eine lineare Funktion*

mit Anstieg $\frac{p}{n}$ beschränkt ist. Es liegt eine lineare Abhängigkeit vor, die durch den Faktor $\frac{p}{n}$ quantifiziert werden kann. Diese Eigenschaft ist durch die Stabilitätszahl ausgedrückt.

Jetzt soll für die Schätzung von $\mathbf{\Sigma}$ besonders wieder die P-Stabilität betrachtet werden. Es soll gezeigt werden, dass eine positiv semidefinite Matrix, die nicht positiv definit ist, als Schätzung für $\mathbf{\Sigma}$ nicht P-stabil ist. Es gibt einen Mindestabstand, den eine solche Matrix zu einer gegebenen Matrix $\mathbf{\Sigma}$ mit Wahrscheinlichkeit 1 nicht unterschreiten kann. Dafür wird eine Abschätzung angegeben.

Satz 6.2 *Sei $\mathbf{\Sigma} \in \mathrm{PD}(p)$, λ_p sei der kleinste Eigenwert von $\mathbf{\Sigma}$. Für jede Matrix $\boldsymbol{A} \in \mathrm{PSD}(p) \setminus \mathrm{PD}(p)$ gilt*

$$\|\boldsymbol{A} - \mathbf{\Sigma}\| \geq \lambda_p. \tag{6.34}$$

Beweis: Sei $\boldsymbol{A} \in \mathrm{PSD}(p) \setminus \mathrm{PD}(p)$. Dann gibt es mindestens einen Eigenwert l_p von $\boldsymbol{A}$ mit $l_p = 0$. Sei $\boldsymbol{y}_p$ der zugehörige normierte Eigenvektor, so dass $\boldsymbol{A}\boldsymbol{y}_p = l_p\boldsymbol{y}_p = 0$ und $\|\boldsymbol{y}_p\| = 1$ gilt. Mit $\lambda_1, \ldots, \lambda_p$ seien die Eigenwerte von $\mathbf{\Sigma}$ und mit $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ die zugehörigen normierten Eigenvektoren bezeichnet. Diese sind orthogonal zueinander und bilden eine Basis des Vektorraumes $\mathbb{R}^p$. Der Vektor $\boldsymbol{y}_p$ lässt sich somit als Linearkombination von $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$,

$$\boldsymbol{y}_p = \sum_{i=1}^{p} \alpha_i \boldsymbol{x}_i, \tag{6.35}$$

darstellen. Mit $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ sei der Koeffizientenvektor von $\boldsymbol{y}_p$ bezeichnet. Da die Vektoren $\boldsymbol{x}_i$ $(i = 1, \ldots, p)$ zueinander orthogonal sind, gilt dies auch für die Vektoren $\alpha_i\boldsymbol{x}_i$. Wegen $\|\boldsymbol{y}_p\| = 1$ und $\|\boldsymbol{x}_i\| = 1$ $(i = 1, \ldots, p)$ erhält man mit dem Satz des Pythagoras:

$$1 = \|\boldsymbol{y}_p\|^2 = \left\| \sum_{i=1}^{p} \alpha_i \boldsymbol{x}_i \right\|^2 = \sum_{i=1}^{p} \|\alpha_i \boldsymbol{x}_i\|^2 = \sum_{i=1}^{p} \alpha_i^2 \|\boldsymbol{x}_i\|^2 = \sum_{i=1}^{p} \alpha_i^2 = (\boldsymbol{\alpha}, \boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|^2. \tag{6.36}$$

Das heißt,

$$\|\boldsymbol{\alpha}\| = 1. \tag{6.37}$$

Mit dem Satz des Pythagoras gilt

$$\left\| \sum_{i=1}^{p} \alpha_i \lambda_i \boldsymbol{x}_i \right\| = \sqrt{\sum_{i=1}^{p} \|\alpha_i \lambda_i \boldsymbol{x}_i\|^2}. \tag{6.38}$$

Damit lässt sich nun die folgende Abschätzung für den Abstand $\|\boldsymbol{A} - \mathbf{\Sigma}\|$ angeben:

$$\|\boldsymbol{A} - \mathbf{\Sigma}\| \quad = \quad \sup_{\|\boldsymbol{x}\|=1} \|(\boldsymbol{A} - \mathbf{\Sigma})\boldsymbol{x}\| \tag{6.39}$$

$$
\begin{aligned}
&\geq \|(\boldsymbol{A}-\boldsymbol{\Sigma})\boldsymbol{y}_p\| && (6.40)\\
&= \|\boldsymbol{\Sigma}\boldsymbol{y}_p\| && (6.41)\\
&= \left\|\boldsymbol{\Sigma}\left(\sum_{i=1}^{p}\alpha_i\boldsymbol{x}_i\right)\right\| && (6.42)\\
&= \left\|\sum_{i=1}^{p}\alpha_i\lambda_i\boldsymbol{x}_i\right\| && (6.43)\\
&= \sqrt{\sum_{i=1}^{p}\alpha_i^2\lambda_i^2} && (6.44)\\
&\geq \sqrt{\sum_{i=1}^{p}\alpha_i^2\lambda_p^2} && (6.45)\\
&= \lambda_p. && (6.46)
\end{aligned}
$$

Damit ist die Behauptung des Satzes gezeigt. □

Die Aussage lässt sich nun auf die Schätzung der Kovarianzmatrix einer multivariaten Normalverteilung übertragen.

Korollar 6.1 *Gegeben sei das durch Ausdruck (2.2) definierte statistische Modell, es gelte $n \leq p$. Die Matrix $\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ sei mit $\boldsymbol{A}$ bezeichnet. Für einen Schätzer $\boldsymbol{\delta} : \mathbb{R}^{n\times p} \to \mathrm{PSD}(p)$ für $\boldsymbol{g}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$ sei durch Gleichung (6.16) eine Verlustfunktion L gegeben, wobei die Norm durch Gleichung (6.7) definiert sei. Dann ist ein Schätzer $\boldsymbol{\delta}$ mit der Gestalt $\boldsymbol{\delta}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = k\boldsymbol{A}, k > 0$, als Schätzer für $\boldsymbol{\Sigma}$ nicht P-stabil bezüglich L. Mit Wahrscheinlichkeit 1 gilt $\|k\boldsymbol{A}-\boldsymbol{\Sigma}\| \geq \lambda_p$, wobei mit λ_p hier wieder der kleinste Eigenwert von $\boldsymbol{\Sigma}$ bezeichnet sei.*

Beweis: Es seien $\boldsymbol{\mu} \in \mathbb{R}^p$ und $\boldsymbol{\Sigma} \in \mathrm{PD}(p)$, beliebig gewählt. Der Schätzer $\boldsymbol{\delta} : \mathbb{R}^{n\times p} \to \mathrm{PSD}(p)$ sei gegeben durch $\boldsymbol{\delta}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = k\boldsymbol{A}$ mit $k > 0$. Da $n \leq p$ gilt, ist $\boldsymbol{A}$ singulär und damit nicht positiv definit. Dies gilt ebenso für $k\boldsymbol{A}$, d. h., $k\boldsymbol{A} \in \mathrm{PSD}(p) \setminus \mathrm{PD}(p)$. Mit Wahrscheinlichkeit 1 gilt dann

$$L((\boldsymbol{\mu},\boldsymbol{\Sigma}), k\boldsymbol{A}) = \|k\boldsymbol{A}-\boldsymbol{\Sigma}\| \geq \lambda_p. \tag{6.47}$$

Wird hier $\varepsilon \leq \lambda_p$ gewählt, so ist die Wahrscheinlichkeit für $\{L((\boldsymbol{\mu},\boldsymbol{\Sigma}), k\boldsymbol{A}) < \varepsilon\}$ gleich Null. Es lässt sich kein $\alpha \in (0,1)$ finden, das hier die Bedingung (6.6), das Kriterium für P-Stabilität, erfüllen könnte. □

6.4 Schätzung der Kovarianzmatrix mit der Ridge-Methode

Bei einparametrigen Ridge-Schätzungen der Form (4.16) ist der Faktor λ ein Parameter, der geeignet zu wählen ist. Gleichung (4.16) lässt sich auch wie folgt schreiben:

$$\boldsymbol{S}_{\text{ridge}} = \frac{1}{1-\tilde{\lambda}}((1-\tilde{\lambda})\,\boldsymbol{S} + \tilde{\lambda}\boldsymbol{I}), \tag{6.48}$$

wobei $\tilde{\lambda} := \frac{\lambda}{1+\lambda}$ gilt. Bei der Diskriminanzanalyse kann der Faktor $\frac{1}{1-\tilde{\lambda}}$ auch weggelassen werden; dies würde ja an der erhaltenen Trennhyperebene nichts ändern. Für die Regularisierte Diskriminanzanalyse (RDA) erhält die Diskriminanzfunktion (4.17) dann die Gestalt

$$f(\boldsymbol{x}) = \Big(\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\Big)'((1-\tilde{\lambda})\,\boldsymbol{S} + \tilde{\lambda}\boldsymbol{I})^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}). \tag{6.49}$$

Man hat dann einen neuen Parameter $\tilde{\lambda}$ ($0 \leq \tilde{\lambda} < 1$). Im Grenzfall hätte man bei $\tilde{\lambda} = 0$ hier die Matrix $\boldsymbol{S}$ selbst; dies würde zur klassischen linearen Diskriminanzanalyse führen. Mit $\tilde{\lambda} = 1$ hätte man als Grenzfall auf der anderen Seite ein Vielfaches der Einheitsmatrix. Die geschätzte Kovarianzstruktur wäre dann geometrisch eine Kugel, in der Diskriminanzanalyse hätte man dann die euklidische Metrik als Abstandsmaß. Bei $\tilde{\lambda}$-Werten zwischen 0 und 1 hat man von beiden „Extremfällen“ einen gewissen Anteil. Der Parameter $\tilde{\lambda}$ gibt hierbei an, wie groß jeweils diese Anteile sind, an welcher Stelle man sich zwischen diesen beiden Extremen bewegt.

Die Anwendung der Theorie der linearen Vektorräume erlaubt jetzt eine Begründung für das Vorgehen bei der Regularisierten Diskriminanzanalyse, insbesondere bei singulärer Stichprobenkovarianzmatrix $\boldsymbol{S}$. In diesem Fall gibt es bei spezieller Wahl der Fehlerfunktion, wie wir gesehen haben, eine untere Schranke $\lambda_p > 0$, die der Fehler der Schätzung mit Wahrscheinlichkeit 1 nicht unterschreitet. Durch Addition mit einer positiv definiten Matrix versucht man hier, eine Schätzung zu bekommen, die näher am „wahren“ Parameter $\boldsymbol{\Sigma}$ liegt.

Man betrachtet jetzt nicht allein die Matrix $\boldsymbol{S}$, sondern alle Linearkombinationen der Form

$$\tilde{\boldsymbol{S}} = (1-\tilde{\lambda})\,\boldsymbol{S} + \tilde{\lambda}\boldsymbol{I} \tag{6.50}$$

entsprechend Gleichung (6.48). Ausdruck (6.50) lässt sich auch schreiben als

$$\tilde{\boldsymbol{S}} = \boldsymbol{S} + \tilde{\lambda}(\boldsymbol{I} - \boldsymbol{S}). \tag{6.51}$$

Für beliebige reelle $\tilde{\lambda}$ ist dies die Gleichung für eine Gerade im Vektorraum $\mathcal{S}(p)$ der symmetrischen $p \times p$-Matrizen. Die Gerade verbindet die beiden Punkte $\boldsymbol{S}$ und $\boldsymbol{I}$. Damit hat man hier einen endlichdimensionalen Unterraum $\mathcal{U}$, und nach dem Fundamentalsatz der linearen Approximationstheorie gibt es ein Element in

$\mathcal{U}$, das zu einem beliebigen Element $\boldsymbol{V} \in \mathcal{S}(p)$ Proximum ist, das heißt minimalen Abstand zu $\boldsymbol{V}$ hat. Unter gewissen Voraussetzungen ist dieses Proximum auch eindeutig. Gesucht wäre dann ein $\tilde{\lambda}$, das dieses Element liefert, d. h., mit dem die erhaltene Schätzung für $\boldsymbol{\Sigma}$ dann innerhalb dieses betrachteten Unterraumes einen minimalen Abstand zu $\boldsymbol{\Sigma}$ hat. Der Parameter $\tilde{\lambda}$ sollte nun so gewählt werden, dass die erhaltene Ridge-Schätzung positiv definit ist. Dies ist bei Werten von $\tilde{\lambda}$ zwischen 0 und 1 immer der Fall. Bei singulärer Matrix $\boldsymbol{S}$ muss $\tilde{\lambda}$ dann zumindest größer als Null sein. Man hofft also, hier immer ein Optimum bei einem $\tilde{\lambda}$ aus dem zulässigen Bereich zu finden.

Dieses optimale $\tilde{\lambda}$ ist allerdings zum Einen unbekannt, zum Anderen ist es immer abhängig von der jeweiligen Stichprobenkovarianzmatrix $\boldsymbol{S}$. Oftmals hat man aber keine geeignete Methode, um für jede konkrete Stichprobenkovarianzmatrix den zugehörigen optimalen Parameter zu bestimmen. Dann ist eine Methode gesucht, um $\tilde{\lambda}$ so zu bestimmen, dass die Abstände zu $\boldsymbol{\Sigma}$ *im Mittel* möglichst klein sind bzw. die zugehörige Risikofunktion minimiert wird. Diese Problemstellung soll später im Kapitel 8 noch ausführlich behandelt werden.

6.5 Schlussbemerkungen

Bei den Aussagen, die hier gemacht wurden, wurde eine ganz spezielle Verlustfunktion und spezielle Normen verwendet. Mit anderen Normen oder einer anderen Verlustfunktion lassen sich aber aufgrund der Eigenschaften solcher Funktionen (z. B. Normäquivalenz in endlichdimensionalen Räumen) ähnliche Eigenschaften vermuten.

Die letzte Aussage über das Nichterfülltsein der Bedingung für P-Stabilität der Schätzer $k\boldsymbol{A}$ für $\boldsymbol{\Sigma}$ macht die Schwierigkeit der Situation erneut deutlich. Durch die P-Instabilität wurde eine Analogie zum Schätzer $\delta_2' = \frac{1}{n}\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2 = 0$ bei $n = 1$ in Beispiel 6.1 hergestellt. Der Wert von δ_2' (bei $n = 1$) ändert sich nicht, wenn der Faktor $\frac{1}{n}$ durch eine beliebige positive reelle Zahl k ersetzt wird, P-Instabilität gilt somit allgemeiner für Schätzer δ der Form $\delta = k\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2$. Diese lassen sich somit in gewisser Weise als Spezialfall für die multivariaten Schätzer $\delta = k\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ auffassen. Im Fall $p = 1$ führt die Voraussetzung $n \leq p$ zu der Bedingung $n = 1$. Die mehrdimensionalen Schätzer $k\boldsymbol{A}$ bei $n \leq p$ stellen umgekehrt eine Verallgemeinerung des — unbrauchbaren — Schätzers $k(x_\alpha - \bar{x})^2 = 0$ bei $n = 1$ dar. Die Bedingung der P-Stabilität liefert in gewisser Weise eine Charakterisierung dieses Zusammenhangs. Man weiß nun, dass die Schätzung mit Sicherheit einen gewissen Abstand zum wahren Parameter hat — den man aber nicht kennt. Dies begründet ein weiteres Mal die Notwendigkeit für zusätzliche Regularisierungsmaßnahmen wie die der Ridge-Methode, womit man versucht, diesen Mangel auszugleichen.

Die folgenden beiden Kapitel sollen sich mit einer möglichst günstigen Wahl des „Regularisierungsparameters“ λ der Ridge-Methode beschäftigen. Ausgehend

von verschiedenen Ansätzen werden dazu verschiedene Vorgehensweisen betrachtet. Zunächst wird in Kapitel 7 versucht, eine näherungsweise Determinanten-erwartungstreue Ridge-Schätzung zu bestimmen. Da in dieser Arbeit vor allem das Klassifikationsproblem betrachtet wird, ist hier jedoch der in Kapitel 8 betrachtete Ansatz, den Klassifikationsfehler möglichst klein zu halten, von größerer Bedeutung.

Kapitel 7

Determinanten-erwartungstreue Ridge-Schätzungen

7.1 Einleitung

Bei multivariater Normalverteilung $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ liefert die Stichprobenkovarianzmatrix $\boldsymbol{S}$ eine erwartungstreue Schätzung für die Kovarianzmatrix $\boldsymbol{\Sigma}$. Soll jedoch die Inverse der Kovarianzmatrix, $\boldsymbol{\Sigma}^{-1}$, oder — wie bei der linearen Diskriminanzanalyse — eine Funktion der inversen Matrix, $\boldsymbol{f}(\boldsymbol{\Sigma}^{-1})$, geschätzt werden, kann als Schätzung $\boldsymbol{S}^{-1}$ bzw. $\boldsymbol{f}(\boldsymbol{S}^{-1})$ verwendet werden, falls $\boldsymbol{S}$ regulär ist. Dieser naheliegende Ansatz versagt, wie wir bereits festgestellt haben, falls $\boldsymbol{S}$ singulär ist (d. h., falls $n \leq p$ ist). Aber selbst bei Invertierbarkeit von $\boldsymbol{S}$ erscheint es nicht immer ratsam, als Schätzung für $\boldsymbol{f}(\boldsymbol{\Sigma}^{-1})$ einfach $\boldsymbol{f}(\boldsymbol{S}^{-1})$ zu verwenden. Speziell bei der linearen Diskriminanzanalyse haben Untersuchungen gezeigt, dass dies zu schlechten Ergebnissen führt, falls n nicht viel größer als p ist [33].

Als eine Alternative wird hier die einparametrige Ridge-Methode betrachtet. Durch Addition einer Diagonalmatrix mit positiven Diagonalelementen zu $\boldsymbol{S}$ ist die erhaltene Matrix immer invertierbar. Ein resultierendes Problem allerdings stellt der zusätzliche Parameter λ dar, der geeignet zu wählen ist.

In diesem Kapitel wird eine Möglichkeit vorgestellt, den Parameter λ so zu bestimmen, dass die Determinante (oder zumindest eine Unterdeterminante) der Schätzung von $\boldsymbol{\Sigma}$ oder ihrer Inversen — zumindest näherungsweise — erwartungstreu ist.

7.2 Die Eigenvektor-Invarianz der einparametrigen Ridge-Schätzung

Eine Schätzung $\tilde{\boldsymbol{S}}$ für die Kovarianzmatrix $\boldsymbol{\Sigma}$ sollte in jedem Fall symmetrisch sein. Damit ließe sich auf eine solche Matrix stets die Spektralzerlegung

$\tilde{\boldsymbol{S}} = \boldsymbol{GLG'}$ anwenden. Umgekehrt ließe sich eine Schätzung aus geeignet gewählten Matrizen $\boldsymbol{G}$ und $\boldsymbol{L}$ mit gewünschten Eigenschaften konstruieren.

Sei $\boldsymbol{G}$ eine Matrix, deren Spaltenvektoren die Eigenvektoren (EV) der Stichprobenkovarianzmatrix $\boldsymbol{S}$ sind. Ist $\boldsymbol{L}$ eine Diagonalmatrix mit positiven Diagonalelementen $l_1, \ldots, l_p$, so ist nach Bemerkung 2.1 die Matrix

$$\boldsymbol{S}_{\mathrm{EV}} := \boldsymbol{GLG'} \tag{7.1}$$

symmetrisch, ihre Eigenwerte sind $l_1, \ldots, l_p$ und die zugehörigen Eigenvektoren sind die Spaltenvektoren von $\boldsymbol{G}$, also die Maximum-Likelihood-Schätzer für die Eigenvektoren von $\boldsymbol{\Sigma}$. Da die Eigenwerte alle positiv sind, ist $\boldsymbol{S}_{\mathrm{EV}}$ positiv definit.

Seien $l_1 \geq \ldots \geq l_p$ die Eigenwerte von $\boldsymbol{S}$, $\boldsymbol{L}$ sei die Diagonalmatrix mit Diagonalelementen $l_1, \ldots, l_p$. Die Eigenwerte $l_1, \ldots, l_p$ des Schätzers für $\boldsymbol{\Sigma}$ können dann — unabhängig davon, ob $\boldsymbol{S}$ regulär oder singulär ist — einfach durch die Vorschrift

$$l_i^* = l_i + \lambda,\ \lambda > 0\ (i = 1, \ldots, p) \tag{7.2}$$

bestimmt werden. Man erhält hier die Diagonalmatrix

$$\boldsymbol{L}^* = \boldsymbol{L} + \lambda \boldsymbol{I} \tag{7.3}$$

(mit Einheitsmatrix $\boldsymbol{I}$). Mit $\boldsymbol{S} = \boldsymbol{GLG'}$ erhält man:

$$\boldsymbol{S}_{\mathrm{EV}} = \boldsymbol{G}(\boldsymbol{L} + \lambda \boldsymbol{I})\boldsymbol{G'} = \boldsymbol{GLG'} + \boldsymbol{G}\lambda\, \boldsymbol{IG'} = \boldsymbol{S} + \lambda \boldsymbol{I}. \tag{7.4}$$

Damit erhält man hier die einparametrige Ridge-Schätzung von $\boldsymbol{\Sigma}$:

$$\boldsymbol{S}_{\mathrm{ridge}} = \boldsymbol{S} + \lambda \boldsymbol{I}. \tag{7.5}$$

Für die Bildung von Ridge-Schätzungen gibt es auch ganz andere Ansätze. Von J. Läuter stammt z. B. der Vorschlag, statt $\boldsymbol{S}$ die Matrix

$$\boldsymbol{S}_{\mathrm{ridge}} = \boldsymbol{S} + \frac{p(n-2)}{(n-4)(n+p-3)} \mathbf{Diag}(\boldsymbol{S}[\mathbf{Diag}(\boldsymbol{S})]^{-1}\boldsymbol{S}) \tag{7.6}$$

zu verwenden, welche eine skalenangepasste Ridge-Korrektur benutzt. Näheres dazu findet man in der entsprechenden Literatur [33] [29].

7.3 Erwartungstreue Schätzer bei parametrischen Abbildungen

Gegeben sei ein statistisches Modell $(M, \mathcal{A}, (P_\vartheta)_{\boldsymbol{\vartheta} \in \Theta})$ mit $\Theta \subset \mathbb{R}^k$. Ein Schätzer δ für den Parameter ϑ wird als erwartungstreu bezeichnet, wenn $\mathrm{E}_\vartheta \delta = \vartheta$ gilt. Das Integral und der Erwartungswert im $\mathbb{R}^k$ werden hier wie üblich komponentenweise verstanden, d. h., $\mathbf{E}(\boldsymbol{T}) = \mathbf{E}(T_1, \ldots, T_k)' = (\mathrm{E}(T_1), \ldots, \mathrm{E}(T_k))'$. Sind $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$

$n \geq 2$ Beobachtungen von unabhängigen, identisch $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, dann sind die Schätzer $\bar{\boldsymbol{x}}$ und
$\boldsymbol{S} = \frac{1}{n-1} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ erwartungstreu für $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$. Das gilt unabhängig vom Rang von $\boldsymbol{S}$, also auch für den Fall, dass $\boldsymbol{S}$ singulär ist. Dies ergibt sich aus der Tatsache, dass als Erwartungswert für die Matrix $\boldsymbol{S} = [s_{ij}]_{i,j=1,\ldots,p}$ jeweils die Matrix $\mathbf{E}(\boldsymbol{S}) = [\mathrm{E}(s_{ij})]_{i,j=1,\ldots,p}$ verstanden wird.

In manchen Situationen ist es zweckmäßig, zusätzlich eine Abbildung $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^l$ zu betrachten. Ein Schätzer $\boldsymbol{T} : (M, \mathcal{A}) \to (\mathbb{R}^k, \mathcal{B}^k)$ wird dann als *$\boldsymbol{g}$-erwartungstreu für $\boldsymbol{\vartheta}$* bezeichnet, falls die Bedingung

$$\mathbf{E}_{\boldsymbol{\vartheta}} \boldsymbol{g}(\boldsymbol{T}) = \boldsymbol{g}(\boldsymbol{\vartheta}) \ \forall \boldsymbol{\vartheta} \in \Theta \tag{7.7}$$

erfüllt ist.

Sei mit $\mathcal{M}(p,p)$ der Raum aller reellen $p \times p$-Matrizen bezeichnet. Betrachtet man z. B. die Abbildung $\mathrm{tr} : \mathcal{M}(p,p) \to \mathbb{R}$, die jeder $p \times p$-Matrix $\boldsymbol{A}$ ihre Spur zuordnet (d. h., $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}\boldsymbol{A} = \sum_{i=1}^{p} a_{ii}$), so folgt aus der Eigenschaft der Additivität von Erwartungswerten für $\boldsymbol{S} = \frac{1}{n-1} \sum_{\alpha=1}^{n} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$

$$\mathrm{E}(\mathrm{tr}\boldsymbol{S}) = \mathrm{E}\left(\sum_{i=1}^{p} s_{ii}\right) = \sum_{i=1}^{p} \mathrm{E}(s_{ii}) = \sum_{i=1}^{p} \sigma_{ii} = \mathrm{tr}\boldsymbol{\Sigma}. \tag{7.8}$$

Ein Schätzer $\boldsymbol{\delta} : M \to \mathcal{M}(p,p)$, der wie $\boldsymbol{S}$ diese Eigenschaft erfüllt, heißt *Spur-erwartungstreu für $\boldsymbol{\Sigma}$*.

In manchen Fällen mag aber vielleicht eher die Determinante einer Matrix von Bedeutung sein. Diese spielt beispielsweise bei der Matrix-Invertierung eine wesentliche Rolle. In solchen Fällen kann es sinnvoll sein, eine Determinanten-erwartungstreue Schätzung einer Matrix zu bestimmen. Man betrachtet entsprechend die Abbildung $\det : \mathcal{M}(p,p) \to \mathbb{R}$. Ein Schätzer $\boldsymbol{\delta} : M \to \mathcal{M}(p,p)$, für den

$$\mathrm{E}(\det\boldsymbol{\delta}(x)) = \det\boldsymbol{\Sigma} \tag{7.9}$$

gilt, wird als Determinanten-erwartungstreu für $\boldsymbol{\Sigma}$ bezeichnet. Dazu sei hier bemerkt, dass die Stichprobenkovarianzmatrix zumeist nicht Determinanten-erwartungstreu ist.

7.4 Die Wishart-Verteilung

7.4.1 Grundlagen

Sind $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ unabhängige, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilte Zufallsvektoren, so unterliegt die Statistik $\boldsymbol{A} = \sum_{\alpha=1}^{n} (\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}})(\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}})'$ einer *Wishart*-Verteilung mit $m = n - 1$ Freiheitsgraden und Parameter $\boldsymbol{\Sigma}$ (Abk.: $W_p(\boldsymbol{\Sigma}, m)$). (Im Mehrstichprobenmodell gilt entsprechend $m = n - J$.) Für $m \geq p$ hat diese Verteilung eine

Dichte, die auf der Menge $\mathcal{S}(p)$ definiert ist. Andernfalls wird diese Verteilung als singulär bezeichnet. Die Dichte einer Wishart-Verteilung $W_p(\mathbf{\Sigma}, m)$ hat die folgende Gestalt [3]:

$$w(\boldsymbol{A}) = \begin{cases} \frac{|\boldsymbol{A}|^{\frac{1}{2}(m-p-1)} \exp\left(-\frac{1}{2}\mathrm{tr}\mathbf{\Sigma}^{-1}\boldsymbol{A}\right)}{2^{\frac{1}{2}mp}\pi^{p(p-1)/4}|\mathbf{\Sigma}|^{\frac{1}{2}m}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(m+1-i)\right]} & \text{für } \boldsymbol{A} \in \mathrm{PD}(p), \\ 0 & \text{sonst.} \end{cases} \tag{7.10}$$

Für eine Matrix $\boldsymbol{A}$, die regulär ist und einer Verteilung $W_p(\mathbf{\Sigma}, m)$ unterliegt, gilt: Die inverse Matrix $\boldsymbol{A}^{-1}$ unterliegt einer inversen Wishart-Verteilung $W_p^{-1}(\mathbf{\Sigma}^{-1}, m)$. Die inverse Wishart-Verteilung $W_p^{-1}(\mathbf{\Psi}, m)$ hat die Dichte

$$w^{-1}(\boldsymbol{B}) = \begin{cases} \frac{|\mathbf{\Psi}|^{\frac{1}{2}m}|\boldsymbol{B}|^{-\frac{1}{2}(m+p+1)} \exp\left(-\frac{1}{2}\mathrm{tr}\mathbf{\Psi}\boldsymbol{B}^{-1}\right)}{2^{\frac{1}{2}mp}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(m+1-i)\right]} & \text{für } \boldsymbol{B} \in \mathrm{PD}(p), \\ 0 & \text{sonst.} \end{cases} \tag{7.11}$$

Es kann gezeigt werden: Hat $\boldsymbol{A}$ die Verteilung $W_p(\mathbf{\Sigma}, m)$, so gilt für den Erwartungswert von $\boldsymbol{A}^{-1}$ (für $m > p+1$):

$$\mathbf{E}(\boldsymbol{A}^{-1}) = \frac{1}{m-p-1}\mathbf{\Sigma}^{-1} \tag{7.12}$$

(z. B. [3, Lemma 7.7.1]). Die Stichprobenkovarianzmatrix
$\boldsymbol{S} = \frac{1}{n-1}\sum_{\alpha=1}^{n}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}})(\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}})'$ hat die Verteilung $W_p(\frac{1}{n-1}\mathbf{\Sigma}, n-1)$, damit gilt bei $n \geq p+3$ für den Erwartungswert von $\boldsymbol{S}^{-1}$:

$$\mathbf{E}(\boldsymbol{S}^{-1}) = \frac{1}{n-p-2}\left(\frac{1}{n-1}\mathbf{\Sigma}\right)^{-1} = \frac{n-1}{n-p-2}\mathbf{\Sigma}^{-1}. \tag{7.13}$$

Das bedeutet:

$$\frac{n-p-2}{n-1}\boldsymbol{S}^{-1} = (n-p-2)\boldsymbol{A}^{-1} \tag{7.14}$$

ist ein erwartungstreuer Schätzer für die Inverse der Kovarianzmatrix, $\mathbf{\Sigma}^{-1}$.

7.4.2 Verteilung der Determinanten

Für die Verteilung der Determinanten von Wishart-verteilten Matrizen gilt der folgende Satz:

Satz 7.1 ([37], Theorem 3.4.8) *Für eine Matrix $\boldsymbol{A}$ mit $\boldsymbol{A} \sim W_p(\mathbf{\Sigma}, m)$, $m \geq p$, ist die Verteilung von $|\boldsymbol{A}|$ gleich der Determinante $|\mathbf{\Sigma}|$, multipliziert mit p stochastisch unabhängigen χ^2-verteilten Zufallsvariablen mit Freiheitsgraden $m, m-1, \ldots, m-p+1$.*

Wegen $\mathrm{E}(\chi_k^2) = k$ gilt für den Erwartungswert von $|\boldsymbol{A}|$:

$$\mathrm{E}(|\boldsymbol{A}|) = |\mathbf{\Sigma}|\prod_{i=1}^{p}(m-i+1). \tag{7.15}$$

7.4.3 Weitere Eigenschaften der Wishart-Verteilung

Satz 7.2 ([37], Theorem 3.4.1) *Gilt $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma}, m)$ und ist $\boldsymbol{B}$ eine konstante $(p \times q)$-Matrix mit Rang $q \leq p$, so gilt $\boldsymbol{B'MB} \sim W_q(\boldsymbol{B'\Sigma B}, m)$.*

Als Folgerung dieses Satzes ergibt sich, dass Teilmatrizen von Wishart-Matrizen ebenfalls eine Wishart-Verteilung haben. Dies besagt der folgende Satz:

Satz 7.3 ([3], Theorem 7.3.4) *Die $p \times p$-Matrizen $\boldsymbol{A}$ und $\boldsymbol{\Sigma}$ seien jeweils folgendermaßen partitioniert in q und $p - q$ Zeilen und Spalten:*

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Hat $\boldsymbol{A}$ die Verteilung $W_p(\boldsymbol{\Sigma}, m)$, dann hat $\boldsymbol{A}_{11}$ die Verteilung $W_q(\boldsymbol{\Sigma}_{11}, m)$.

7.5 Parameterwahl bei der Ridge-Methode

Die theoretischen Ergebnisse der bisherigen Abschnitte sollen jetzt für die Ridge-Methode angewandt werden. Es wird eine Möglichkeit angegeben, wie der Parameter λ bei der einparametrigen Ridge-Methode geeignet gewählt werden kann, so dass die Determinante der resultierenden Matrix, zumindest näherungsweise, erwartungstreu ist. Zunächst wird die Schätzung der Kovarianzmatrix $\boldsymbol{\Sigma}$ betrachtet, anschließend die Schätzung von $\boldsymbol{\Sigma}^{-1}$.

7.5.1 Schätzung für $\boldsymbol{\Sigma}$

Wir betrachten das Modell von n Beobachtungen von unabhängigen, identisch multivariat normalverteilten Zufallsvektoren $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(i = 1, \ldots, n)$. Der Ridge-Schätzer für die Kovarianzmatrix $\boldsymbol{\Sigma}$ hat allgemein die Form

$$\boldsymbol{S}_{\text{ridge}} = \boldsymbol{S} + \lambda \boldsymbol{I}, \tag{7.16}$$

wobei nun λ geeignet zu wählen ist. Wir betrachten hier wieder besonders den Fall von singulären Kovarianzmatrizen. Damit auch in diesem Fall (7.16) positiv definit ist, müsste hierbei stets $\lambda > 0$ gelten.

Wünschenswert wäre es, den Parameter $\lambda = \lambda(\boldsymbol{S})$ in Abhängigkeit von der Stichprobenkovarianzmatrix so bestimmen zu können, dass

$$\mathrm{E}(|\, \boldsymbol{S} + \lambda(\boldsymbol{S})\boldsymbol{I}|) = |\boldsymbol{\Sigma}| \tag{7.17}$$

gilt. Dies erreicht man, falls $\boldsymbol{S}$ regulär ist. Um auch den Fall von singulären Stichprobenkovarianzmatrizen $\boldsymbol{S}$ berücksichtigen zu können, werden hier allgemeiner

Teilmatrizen von $\boldsymbol{S}$ und $\boldsymbol{\Sigma}$ betrachtet. Wir betrachten dazu Partitionierungen von Matrizen $\boldsymbol{M}(p \times p)$ wie in Satz 7.3:

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{pmatrix}. \tag{7.18}$$

Ist die Ordnung q der Teilmatrix $\boldsymbol{S}_{11}$ bzw. $\boldsymbol{\Sigma}_{11}$ entsprechend klein, d. h., gilt

$$q \leq n - 1, \tag{7.19}$$

so ist deren Regularität gewährleistet. Der Parameter $\lambda > 0$ ist dann schließlich so zu bestimmen, dass

$$\mathrm{E}(|\boldsymbol{S}_{11} + \lambda \boldsymbol{I}_q|) = |\boldsymbol{\Sigma}_{11}| \tag{7.20}$$

gilt. Als wesentliches Ergebnis dieses Abschnitts wird dazu die Aussage des folgenden Satzes angegeben.

Satz 7.4 *Gegeben seien n Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ von unabhängigen, identisch $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Es gelten die gleichen Bezeichnungen wie bisher. Die Matrizen $\boldsymbol{S}$ und $\boldsymbol{\Sigma}$ seien wie in (7.18) in $2 \leq q \leq n-1$ und $p-q$ Zeilen und Spalten partitioniert, $l_1 \geq \ldots \geq l_q$ seien die Eigenwerte von $\boldsymbol{S}_{11}$. Dann hat die folgende Gleichung*

$$\prod_{i=1}^{q}(l_i + \lambda) = \prod_{i=1}^{q} l_i \frac{n-1}{n-i} \tag{7.21}$$

genau eine positive Lösung für λ. Ist λ die positive Lösung von (7.21), so ist die Bedingung (7.20) erfüllt.

Beweis: Wir betrachten die Funktion

$$f(\lambda) = \prod_{i=1}^{q}(l_i + \lambda) - \prod_{i=1}^{q} l_i \frac{n-1}{n-i}. \tag{7.22}$$

Im Bereich $\lambda \geq 0$ ist f monoton wachsend. Bei $\lambda = 0$ gilt:

$$f(\lambda) = \prod_{i=1}^{q} l_i - \prod_{i=1}^{q} l_i \frac{n-1}{n-i} = \prod_{i=1}^{q} l_i \left(1 - \prod_{i=1}^{q} \frac{n-1}{n-i}\right) < 0. \tag{7.23}$$

Für $\lambda \to \infty$ gilt:

$$\lim_{\lambda \to \infty} f(\lambda) = \infty. \tag{7.24}$$

Nach dem Zwischenwertsatz gibt es damit (mindestens) eine positive Nullstelle von f. Wegen der Monotonie hat f genau eine Nullstelle im Bereich $[0, \infty)$. Diese Nullstelle ist genau die Lösung der Gleichung (7.21).

Die linke Seite von Gleichung (7.21) ist nach Abschnitt 7.2 das Produkt der Eigenwerte von $\boldsymbol{S}_{11} + \lambda \boldsymbol{I}$, damit gilt

$$|\boldsymbol{S}_{11} + \lambda \boldsymbol{I}| = \prod_{i=1}^{q} (l_i + \lambda) = \prod_{i=1}^{q} l_i \frac{n-1}{n-i}. \tag{7.25}$$

Weiterhin gilt

$$|\boldsymbol{S}_{11}| = \prod_{i=1}^{q} l_i, \tag{7.26}$$

damit erhält man

$$|\boldsymbol{S}_{11} + \lambda \boldsymbol{I}| = |\boldsymbol{S}_{11}| \prod_{i=1}^{q} \frac{n-1}{n-i}. \tag{7.27}$$

Nach Satz 7.3 hat $\boldsymbol{S}_{11}$ die Verteilung $W_q(\frac{1}{n-1}\boldsymbol{\Sigma}_{11}, n-1)$. Nach Satz 7.1 ist die Verteilung der Determinante $|\boldsymbol{S}_{11}|$ gleich $|\frac{1}{n-1}\boldsymbol{\Sigma}_{11}|$, multipliziert mit q stochastisch unabhängigen χ^2-verteilten Zufallsvariablen mit Freiheitsgraden $n-1, n-2, \ldots, n-q$. Wegen $\mathrm{E}(\chi_k^2) = k$ gilt dann für den Erwartungswert von $|\boldsymbol{S}_{11}|$:

$$\mathrm{E}(|\boldsymbol{S}_{11}|) = \frac{1}{(n-1)^p} |\boldsymbol{\Sigma}_{11}| \prod_{i=1}^{q} (n-i) = |\boldsymbol{\Sigma}_{11}| \prod_{i=1}^{q} \frac{n-i}{n-1}. \tag{7.28}$$

Mit Gleichung (7.27) folgt dann für den Erwartungswert von $\mathrm{E}(|\boldsymbol{S}_{11} + \lambda \boldsymbol{I}_q|)$:

$$\mathrm{E}(|\boldsymbol{S}_{11} + \lambda \boldsymbol{I}_q|) = \mathrm{E}\left(|\boldsymbol{S}_{11}| \prod_{i=1}^{q} \frac{n-1}{n-i} \right) = |\boldsymbol{\Sigma}_{11}|. \tag{7.29}$$

□

Gleichung (7.21) lässt sich auch in eine Polynomialgleichung umformen. Die linke Seite lässt sich wie folgt schreiben:

$$\begin{aligned} \prod_{i=1}^{q} (l_i + \lambda) &= \sum_{i=0}^{q} \left(\sum_{C \in C^i(L_q)} \left(\prod_{j: l_j \in C} l_j \right) \lambda^{q-i} \right) && (7.30) \\ &= \sum_{i=70}^{q-1} \left(\sum_{C \in C^i(L_q)} \left(\prod_{j: l_j \in C} l_j \right) \lambda^{q-i} \right) + \prod_{i=1}^{q} l_i && (7.31) \end{aligned}$$

($C^i(L_q)$ bezeichnet die Menge aller Teilmengen von $\{l_1, \ldots, l_q\}$ mit i Elementen). Einsetzen in Gleichung (7.21) liefert:

$$\sum_{i=0}^{q-1} \left(\sum_{C \in C^i(L_q)} \left(\prod_{j: l_j \in C} l_j \right) \lambda^{q-i} \right) + \prod_{i=1}^{q} l_i = \prod_{i=1}^{q} l_i \frac{n-1}{n-i}. \tag{7.32}$$

Dies ist äquivalent zu der polynomialen Gleichung

$$\sum_{i=0}^{q-1}\left(\sum_{C\in C^i(L_q)}\left(\prod_{j:l_j\in C} l_j\right)\lambda^{q-i}\right)+\prod_{i=1}^{q} l_i\left(1-\prod_{i=1}^{q}\frac{n-1}{n-i}\right)=0. \tag{7.33}$$

In der praktischen Anwendung kann diese Polynomgleichung nun näherungsweise gelöst werden. (Die SAS/IML-Funktion POLYROOT beispielsweise verwendet dazu einen Algorithmus von Jenkins und Traub [26].) Damit ist eine Möglichkeit für die geeignete Bestimmung des Parameters λ bei der Ridge-Methode gegeben.

7.5.2 Schätzung für $\boldsymbol{\Sigma}^{-1}$

Es sei das gleiche Normalverteilungsmodell wie bisher vorausgesetzt. Mit $\boldsymbol{A}$ sei hier wieder die Statistik

$$\boldsymbol{A}=\sum_{\alpha=1}^{n}(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})'=(n-1)\boldsymbol{S} \tag{7.34}$$

bezeichnet. Ein erwartungstreuer Schätzer für $\boldsymbol{\Sigma}^{-1}$ ist nach (7.14) gegeben durch

$$(n-p-2)\boldsymbol{A}^{-1}=\Big(\frac{1}{n-p-2}\boldsymbol{A}\Big)^{-1} \tag{7.35}$$

(bei $n-2>p$). Jetzt wäre es wünschenswert, λ so bestimmen zu können, dass

$$\mathrm{E}\left(\left|\Big(\frac{1}{n-p-2}\,\boldsymbol{A}+\lambda\boldsymbol{I}\Big)^{-1}\right|\right)=|\boldsymbol{\Sigma}^{-1}| \tag{7.36}$$

bzw.

$$\mathrm{E}\left(\left|\Big(\frac{1}{n-p-2}\boldsymbol{A}_{11}+\lambda\boldsymbol{I}_q\Big)^{-1}\right|\right)=|\boldsymbol{\Sigma}_{11}^{-1}| \tag{7.37}$$

gilt. Dies gelingt nicht in gleicher Weise wie bei der Schätzung von $\boldsymbol{\Sigma}$. Wie wir noch sehen werden, kann hierbei nicht eine Teilmatrix von $\frac{1}{n-p-2}\boldsymbol{A}$ verwendet werden. Der Faktor $\frac{1}{n-p-2}$ müsste durch einen Faktor $k>0$ ersetzt werden, der bestimmten Bedingungen genügt.

Als Modell für die Ridge-Schätzung von $\boldsymbol{\Sigma}^{-1}$ betrachten wir allgemein

$$\boldsymbol{S}_{\text{ridge}}^{inv}=(k\,\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}. \tag{7.38}$$

(Durch die Bezeichnung soll hier deutlich von der Inversen der Ridge-Schätzung $\boldsymbol{S}_{\text{ridge}}^{-1}=(\boldsymbol{S}_{\text{ridge}})^{-1}$ unterschieden werden.)

Wir betrachten jetzt Teilmatrizen der Ordnung

$$q\le n-3. \tag{7.39}$$

Falls $p \leq n-3$ gilt, kann $q = p$ gesetzt werden. Der Parameter λ soll nun so bestimmt werden, dass

$$\mathrm{E}(|(k\boldsymbol{A}_{11} + \lambda \boldsymbol{I}_q)^{-1}|) = |\boldsymbol{\Sigma}_{11}^{-1}| \tag{7.40}$$

für gewisse $k > 0$ gilt. Dazu ist in dem folgenden Satz eine Aussage gegeben.

Satz 7.5 *Gegeben seien n Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ von unabhängigen, identisch $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Es gelten die gleichen Bezeichnungen wie bisher. Die Matrizen $\boldsymbol{S}$ und $\boldsymbol{\Sigma}$ seien wie in (7.18) in $2 \leq q \leq n-3$ und $p-q$ Zeilen und Spalten partitioniert, $l_1 \geq \ldots \geq l_q$ seien die Eigenwerte von $k\boldsymbol{A}_{11}$. Gilt*

$$0 < k \leq \frac{1}{n-3}, \tag{7.41}$$

so hat die folgende Gleichung

$$\prod_{i=1}^{q}(l_i + \lambda) = \prod_{i=1}^{q} l_i \frac{1}{k(n-i-2)} \tag{7.42}$$

genau eine positive Lösung für λ. Ist λ die positive Lösung von (7.42), so ist die Bedingung (7.40) erfüllt.

Beweis: Wir betrachten jetzt die Funktion

$$f(\lambda) = \prod_{i=1}^{q}(l_i + \lambda) - \prod_{i=1}^{q} l_i \frac{1}{k(n-i-2)}. \tag{7.43}$$

Im Bereich $\lambda \geq 0$ ist f ebenfalls monoton wachsend. Bei $\lambda = 0$ gilt:

$$f(0) = \prod_{i=1}^{q} l_i - \prod_{i=1}^{q} l_i \frac{1}{k(n-i-2)} = \prod_{i=1}^{q} l_i \left(1 - \prod_{i=1}^{q} \frac{1}{k(n-i-2)}\right). \tag{7.44}$$

Gilt hierbei $0 < k \leq \frac{1}{n-3}$, so folgt $f(0) < 0$. Für $\lambda \to \infty$ gilt wieder:

$$\lim_{\lambda \to \infty} f(\lambda) = \infty. \tag{7.45}$$

Nach dem Zwischenwertsatz gibt es damit (mindestens) eine positive Nullstelle von f. Wegen der Monotonie hat f genau eine Nullstelle im Bereich $[0, \infty)$. Diese Nullstelle ist genau die Lösung der Gleichung (7.42).

Die linke Seite von Gleichung (7.42) ist nach Abschnitt 7.2 das Produkt der Eigenwerte von $k\boldsymbol{A}_{11} + \lambda \boldsymbol{I}$, damit gilt

$$|k\boldsymbol{A}_{11} + \lambda \boldsymbol{I}| = \prod_{i=1}^{q}(l_i + \lambda) = \prod_{i=1}^{q} l_i \frac{1}{k(n-i-2)}. \tag{7.46}$$

Mit

$$|k\boldsymbol{A}_{11}| = \prod_{i=1}^{q} l_i \tag{7.47}$$

erhält man hier

$$|k\boldsymbol{A}_{11} + \lambda\boldsymbol{I}| = |k\boldsymbol{A}_{11}| \prod_{i=1}^{q} \frac{1}{k(n-i-2)}. \tag{7.48}$$

Dies ist äquivalent zu

$$|k\boldsymbol{A}_{11} + \lambda\boldsymbol{I}|^{-1} = |k\boldsymbol{A}_{11}|^{-1} \prod_{i=1}^{q} k(n-i-2). \tag{7.49}$$

Mit Satz 7.3 gilt $\boldsymbol{A}_{11} \sim W_q(\boldsymbol{\Sigma}_{11}, n-1)$; die Verteilung von $|\boldsymbol{A}_{11}|$ ist nach Satz 7.1 gleich $|\boldsymbol{\Sigma}_{11}|$, multipliziert mit q stochastisch unabhängigen χ^2-verteilten Zufallsvariablen mit Freiheitsgraden $n-1, n-2, \ldots, n-q$. Für $l > 2$ gilt $\mathrm{E}(\chi_l^2)^{-1} = \frac{1}{l-2}$. Damit ist, falls $n > q+2$ gilt, der Erwartungswert von $|\boldsymbol{A}_{11}^{-1}|$ gegeben durch

$$\mathrm{E}(|\boldsymbol{A}_{11}^{-1}|) = \frac{1}{|\boldsymbol{\Sigma}_{11}|} \prod_{i=1}^{q} \frac{1}{n-i-2}. \tag{7.50}$$

Damit ist eine erwartungstreue Schätzung für $|\boldsymbol{\Sigma}_{11}^{-1}|$ gegeben durch

$$|\boldsymbol{A}_{11}^{-1}| \prod_{i=1}^{q} (n-i-2) \tag{7.51}$$

bzw. durch

$$|k\boldsymbol{A}_{11}|^{-1} \prod_{i=1}^{q} k(n-i-2). \tag{7.52}$$

Mit Gleichung (7.49) folgt dann die Gültigkeit von Bedingung (7.40). □

Gleichung (7.42) lässt sich analog wieder als Polynomgleichung darstellen. Die linke Seite lässt sich wie folgt umformen:

$$\prod_{i=1}^{q}(l_i + \lambda) = \sum_{i=0}^{q} \left(\sum_{C \in C^i(L_q)} \left(\prod_{j:l_j \in C} l_j \right) \lambda^{q-i} \right) \tag{7.53}$$

$$= \sum_{i=0}^{q-1} \left(\sum_{C \in C^i(L_q)} \left(\prod_{j:l_j \in C} l_j \right) \lambda^{q-i} \right) + \prod_{i=1}^{q} l_i \tag{7.54}$$

($C^i(L_q)$ bezeichnet wieder die Menge aller Teilmengen von $\{l_1, \ldots, l_q\}$ mit i Elementen.) Einsetzen in Gleichung (7.42) liefert:

$$\sum_{i=0}^{q-1} \left(\sum_{C \in C^i(L_q)} \left(\prod_{j:l_j \in C} l_j \right) \lambda^{q-i} \right) + \prod_{i=1}^{q} l_i = \prod_{i=1}^{q} l_i \frac{1}{k(n-i-2)}. \tag{7.55}$$

Man erhält hier die polynomiale Gleichung

$$\sum_{i=0}^{q-1}\left(\sum_{C\in C^i(L_q)}\left(\prod_{j:l_j\in C} l_j\right)\lambda^{q-i}\right)+\prod_{i=1}^{q} l_i\left(1-\prod_{i=1}^{q}\frac{\frac{1}{k}}{n-i-2}\right)=0. \qquad (7.56)$$

Die Polynomgleichung (7.56) kann nun wieder numerisch gelöst werden. Als Faktor k kann dabei etwa der Faktor $\frac{1}{n}$ der Maximum-Likelihood-Schätzung verwendet werden; dieser erfüllt in jedem Fall die Bedingung (7.41).

7.5.3 Variablenauswahl

In beiden Verfahren, bei der Schätzung für $\boldsymbol{\Sigma}$ und für $\boldsymbol{\Sigma}^{-1}$, ist die Anzahl der für die Teilmatrix ausgewählten Variablen q jeweils durch Bedingung (7.19) bzw. (7.39) begrenzt. Je höher nun dieses q gewählt wird, desto näher ist die Teilmatrix $\boldsymbol{\Sigma}_{11}$ bzw. $\boldsymbol{\Sigma}_{11}^{-1}$ der gesamten Matrix $\boldsymbol{\Sigma}$ bzw. $\boldsymbol{\Sigma}^{-1}$. Sicher ist ebenso die Teilmatrix von $\boldsymbol{S}+\lambda\boldsymbol{I}$ bzw. von $k\,\boldsymbol{A}+\lambda\boldsymbol{I}$ der entsprechenden gesamten Matrix am nächsten, je größer q ist. Das heißt, man kommt in dieser Hinsicht der erwartungstreuen Schätzung der Determinante von $\boldsymbol{\Sigma}$ bzw. von $\boldsymbol{\Sigma}^{-1}$ am nächsten, wenn man q so groß wie möglich wählt. Andererseits sind für die Teilmatrix bei höherer Dimension auch mehr Parameter zu schätzen, was wiederum zu größeren Schätzfehlern führt. Daher ist es nicht unbedingt in jedem Fall angebracht, q so groß wie möglich zu wählen. Gegebenenfalls hat man da einen geeigneten Kompromiss zu suchen.

Natürlich müssen die für die Bildung der Teilmatrizen ausgewählten Variablen nicht unbedingt die nach der ursprünglichen Reihenfolge ersten sein. Die Variablen (Zeilen und Spalten) können dazu beliebig untereinander vertauscht werden. Die ausgewählten Zeilen- und Spaltenindizes müssen natürlich die gleichen sein, so dass die erhaltene Teilmatrix stets symmetrisch ist. Diese Variablenauswahl kann nach geeigneten Kriterien erfolgen. Beispielsweise können die q Variablen mit den größten empirischen Varianzen für die Bildung der Teilmatrix verwendet werden.

7.5.4 Simulationsergebnisse

Nach der beschriebenen Methode wurden durch Simulationen die entsprechenden Parameter λ bestimmt, zunächst in Abhängigkeit von der Korrelation der Variablen, dann in Abhängigkeit von der Anzahl der Variablen. Hierbei ist λ_1 jeweils die positive Lösung der Polynomgleichung (7.33), q hat entweder den maximal möglichen Wert $q_{\max}=\min(n-2,p)$ oder $\frac{1}{2}q_{\max}$ (ganzzahlig gerundet). Analog ist λ_2 die positive Lösung der Polynomgleichung (7.56), q hat den maximalen Wert $q_{\max}=\min(n-4,p)$ oder $\frac{1}{2}q_{\max}$ (ganzzahlig gerundet). Für das zweite Verfahren wurde der Faktor $k=\frac{1}{n}$ verwendet (vgl. Abschnitt 7.5.2).

Es wurden jeweils 10 Beobachtungen unabhängiger, identisch verteilter Zufallsvektoren simuliert. Für die Berechnungen in Abhängigkeit von der Korrelation der Variablen wurde eine 50-dimensionale multivariate Normalverteilung mit Mittelwertsvektor $\boldsymbol{\mu} = (1, \ldots, 1)'$ zugrundegelegt. Es wurde angenommen, dass die 50 Variablen jeweils einheitlich die Varianz 1 haben, für die paarweisen Korrelationen zwischen verschiedenen Variablen wurden einheitlich die Werte $\varrho = 0.1$, $\varrho = 0.5$ bzw. $\varrho = 0.9$ angenommen.

Die Parameter λ_1 und λ_2 wurden in jedem Simulationslauf entsprechend neu berechnet. In Tabelle 7.1 sind für λ_1 und λ_2 für die verschiedenen paarweisen Korrelationen ϱ jeweils die aus den 1000 Simulationsläufen ermittelten empirischen Mittelwerte $\bar{\lambda}_1$ und $\bar{\lambda}_2$ sowie die empirischen Standardabweichungen s_1 und s_2 dargestellt.

q	$0.5q_{\max}$				$q_{\max}$			
ϱ	$\bar{\lambda}_1$	s_1	$\bar{\lambda}_2$	s_2	$\bar{\lambda}_1$	s_1	$\bar{\lambda}_2$	s_2
0.1	2.13	0.17	0.76	0.05	6.53	0.49	1.08	0.04
0.5	1.41	0.12	0.84	0.12	3.97	0.31	1.66	0.12
0.9	0.30	0.02	1.93	0.27	0.82	0.06	7.07	0.43

Tabelle 7.1: Empirische Mittelwerte und Standardabweichungen der Parameter bei verschiedenen paarweisen Korrelationen

Für die Simulationen in Abhängigkeit von der Anzahl der Variablen wurden die gleichen Verteilungsannahmen zugrundegelegt. Hier wurden die paarweisen Korrelationen zwischen je zwei verschiedenen Variablen einheitlich auf 0.5 gesetzt, und die Anzahl der Variablen p wurde variiert. Tabelle 7.2 zeigt die empirischen Mittelwerte und Standardabweichungen der berechneten Ridge-Parameter.

q	$0.5q_{\max}$				$q_{\max}$			
p	$\bar{\lambda}_1$	s_1	$\bar{\lambda}_2$	s_2	$\bar{\lambda}_1$	s_1	$\bar{\lambda}_2$	s_2
5	0.14	0.04	0.47	0.10	0.12	0.05	0.37	0.15
10	0.33	0.06	0.72	0.08	0.41	0.11	1.51	0.26
15	0.49	0.07	0.79	0.09	0.89	0.15	2.02	0.12
20	0.63	0.08	0.82	0.10	1.34	0.18	2.01	0.12
50	1.41	0.12	0.84	0.12	3.97	0.31	1.66	0.12
100	2.61	0.16	0.84	0.12	8.25	0.43	1.45	0.11
200	4.92	0.23	0.83	0.12	16.68	0.63	1.27	0.09

Tabelle 7.2: Empirische Mittelwerte und Standardabweichungen der Ridge-Parameter in Abhängigkeit von der Variablenanzahl

Insgesamt sind die Unterschiede in den Ergebnissen nicht sehr groß. Generell

kann man sehen, dass die errechneten Parameter $\bar{\lambda}_1$ und $\bar{\lambda}_2$ — außer im Fall $p = 5$ — bei der Wahl $q = q_{\max}$ höhere Werte haben als bei $0.5q_{max}$. Bei den verschiedenen Korrelationen verhalten sich $\bar{\lambda}_1$ und $\bar{\lambda}_2$ gegenläufig; bei höheren Korrelationen wird $\bar{\lambda}_1$ kleiner und $\bar{\lambda}_2$ größer. Eine Erhöhung der Variablenanzahl führt generell zu einer Erhöhung von $\bar{\lambda}_1$, bei $\bar{\lambda}_2$ gilt dies nur bei kleineren Variablenzahlen. Bei höheren Dimensionen bleibt $\bar{\lambda}_2$ bei weiterer Erhöhung der Merkmalszahl nahezu konstant bzw. wird sogar wieder etwas kleiner.

7.6 Schlussbemerkung

In diesem Kapitel wurde für die Bestimmung des Ridge-Parameters das Kriterium der näherungsweisen Determinanten-Erwartungstreue verwendet. Speziell in der Diskriminanzanalyse ist ja die Determinanten-Erwartungstreue nicht das entscheidende Kriterium. Vielmehr ist man dabei an einer Minimierung des Klassifikationsfehlers interessiert. Daher ist es naheliegend, dafür dann auch dieses Kriterium — einen möglichst kleinen Klassifikationsfehler — zugrunde zu legen. Darum soll es vorrangig im dritten Teil dieser Arbeit gehen.

Teil III

Klassifikation bei hohen Dimensionen

In diesem letzten Teil kehren wir wieder zu der Schwerpunktthematik dieser Arbeit, dem Klassifikationsproblem, zurück. Wie bereits im ersten Teil erwähnt, steht dabei die Zwei-Klassen-Situation im Mittelpunkt unserer Betrachtungen. Es wird hier im Wesentlichen wieder das Problem der linearen Diskriminanzanalyse betrachtet. Daher wird auch, wenn nicht anders angegeben, hier stets das Zwei-Stichproben-Modell zugrundegelegt.

Das Kapitel 8 beschäftigt sich noch einmal mit der möglichst günstigen Wahl des Ridge-Parameters. Hierbei ist die Minimierung des Klassifikationsfehlers das zugrundegelegte Kriterium. Im Kapitel 9 wird schließlich an Beispielen die Anwendung verschiedener Klassifikationsverfahren demonstriert.

Kapitel 8

Klassifikationsfehleranalyse

Bei der Diskriminanzanalyse ist man gewöhnlich an einer Minimierung des Risikos (3.13), des Klassifikationsfehlers, interessiert. Da zumeist die Verteilung der verwendeten Statistiken nicht bekannt ist, ist es oft schwierig, dieses Risiko zu bestimmen. Für einige Verfahren gibt es dafür näherungsweise Berechnungsformeln, doch zumeist ist man darauf angewiesen, dieses durch Anwenden des Verfahrens auf Daten mit bekannter Klassenzugehörigkeit zu schätzen.

Auch für die Ridge-Methode lässt sich nicht ohne Weiteres ein im Sinne der Optimierung von (3.13) optimaler Parameter bestimmen. Von Š. Raudys u. M. Skurichina [43] gibt es eine asymptotische Berechnungsformel zur Bestimmung des Klassifikationsfehlers. Die Autoren haben in [43] unter anderem Untersuchungen zur Abhängigkeit des optimalen Ridge-Parameters von der Anzahl der Lernstichprobenelemente bei niedriger Dimension ($p = 8$) und unter der Bedingung $n - 2 \geq p$ angestellt.

Ausgehend von dem Ansatz von Raudys und Skurichina wird hier eine neue asymptotische Berechnungsformel für die Berechnung des Klassifikationsfehlers, die nicht an die Bedingung $n - 2 \geq p$ gebunden ist, hergeleitet. Daraus ergibt sich dann auch eine Abschätzung für die Wahl des optimalen Parameters der Ridge-Methode.

Am Ende des Kapitels werden schließlich Simulationsergebnisse vorgestellt. Es wird angegeben, mit welchen Ridge-Parametern bei höheren Merkmalsanzahlen und insbesondere unter der Bedingung $n \leq p$ die kleinsten Fehlerraten erzielt wurden. In Abhängigkeit von der Anzahl der Variablen, von der Anzahl der Lernstichprobenelemente und von der Korrelation zwischen den Variablen wurden dazu jeweils für verschiedene Werte von λ die Fehlerraten ermittelt, zum Vergleich außerdem die jeweils mit dem berechneten optimalen Parameter ermittelten Fehlerraten.

8.1 Klassifikationsrisiko bei bekannten Parametern

Das Risiko (3.3) des Schätzers für den Klassenparameter j stellt ein wesentliches Kriterium für die Beurteilung der Güte einer Klassifikation dar. Oft verwendet man hierbei die einfache Kostenfunktion

$$C(j,i) = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases} \tag{8.1}$$

Für Klasse j entspricht dann das Risiko $R(j,\delta)$ der Fehlklassifikationswahrscheinlichkeit, d. h. der Wahrscheinlichkeit der fehlerhaften Klassifikation bei wahrem Parameter j. Diese Wahrscheinlichkeit lässt sich bei linearer Diskriminanzanalyse mit bekannten Verteilungen wie folgt bestimmen. Gegeben sei das Modell

$$(\mathbb{R}^p, \mathcal{B}^p, (N(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}))_{j=1,2}). \tag{8.2}$$

Das Risiko (3.3) lässt sich dann schreiben als

$$R(j,\delta) = \sum_{i=1}^{2} C(j,i) p_{ji}(\delta) = p_{ji:i\neq j} \ (j = 1,2) \tag{8.3}$$

mit

$$p_{ji} = P_j(\{\boldsymbol{x} \in \mathbb{R}^p : \delta(\boldsymbol{x}) = i\})$$

und

$$P_j = N(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}) \ (j = 1,2).$$

Für die einzelnen Wahrscheinlichkeiten p_{12} und p_{21} gilt dann

$$p_{12} = P_1(\{\boldsymbol{x} \in \mathbb{R}^p : f(\boldsymbol{x}) \leq 0\})$$

und

$$p_{21} = P_2(\{\boldsymbol{x} \in \mathbb{R}^p : f(\boldsymbol{x}) > 0\}),$$

wobei f durch (3.11) gegeben ist. Es lässt sich nun feststellen, dass in diesem Fall beide Wahrscheinlichkeiten gleich groß sind und dass

$$p_{12} = p_{21} = \Phi\left(-\frac{1}{2}\Delta\right) \tag{8.4}$$

gilt (vgl. dazu z. B. [37]). Φ bezeichnet hier die Verteilungsfunktion der Standardnormalverteilung $N(0,1)$. Die als *Mahalanobis-Abstand* bezeichnete Größe Δ^2 ist durch

$$\Delta^2 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \tag{8.5}$$

gegeben, damit erhält man dann $\Delta = \sqrt{\Delta^2}$.

8.2 Asymptotische Betrachtungen in der Literatur

Es konnte gezeigt werden, dass (8.4) das minimale bayessche und damit optimale Risiko der linearen Diskriminanzanalyse ist. Es ist zugleich das asymptotische Risiko (für $n \to \infty$) der Fisher'schen linearen Diskriminanzregel [42].

Für einige andere Verfahren gibt es auch asymptotische Näherungsformeln für $n \to \infty$ für die Bestimmung des Risikos. Diese sind aber für kleineres n oftmals recht ungenau. Genauere Abschätzungen ergeben sich aus einer doppelt asymptotischen Betrachtung, wobei die Anzahl der Lernstichprobenelemente und die Anzahl der Variablen gleichmäßig gegen ∞ streben. A. D. Deev hat dies wie folgt formalisiert: Sind n_1 und n_2 jeweils die Anzahlen der Lernstichprobenelemente für die beiden Klassen ($n = n_1 + n_2$) und ist p die Anzahl der Variablen, so werden die Grenzübergänge $n_1 \to \infty$, $n_2 \to \infty$, $p \to \infty$ bei $p/n = const.$ und konstantem Mahalanobis-Abstand Δ^2 betrachtet [42]. Es konnte festgestellt werden, dass auf diese Weise erhaltene Abschätzungen oft auch für kleinere n und p sehr genau sind [43].

Lässt man bei der Ridge-Methode $\lambda \to \infty$ streben, so erhält man als Grenzfall die folgende Diskriminanzfunktion:

$$f(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\right)' (\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}). \tag{8.6}$$

Hierbei wird der euklidische Abstand zu den Mittelwertsvektoren als Entscheidungskriterium zugrundegelegt. Das Klassifikationsverfahren, das diese Diskriminanzfunktion verwendet, wird daher auch als *Euklid-Distanz*-Klassifikationsverfahren (engl.: Euclidian distance classifier, EDC) bezeichnet [42].

Für das EDC-Verfahren hat Raudys [42] für den Fall $n_1 = n_2$ und $q_1 = q_2 = \frac{1}{2}$ die folgende doppelt asymptotische Näherung angegeben:

$$r(Q, \mathrm{EDC}) \approx \Phi\left(-\frac{\Delta^*}{2}\frac{1}{\sqrt{T_{\boldsymbol{\mu}}}}\right) \tag{8.7}$$

mit

$$\Delta^* = \frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{\mu}}} \qquad (\boldsymbol{\mu} = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}),$$

$$T_{\boldsymbol{\mu}} = 1 + \frac{4p^*}{\Delta^{*2} n}$$

und der *effektiven Dimension*

$$p^* = \frac{(\boldsymbol{\mu}'\boldsymbol{\mu})^2 (\mathrm{tr}\boldsymbol{\Sigma}^2)}{(\boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{\mu})^2}.$$

Der Begriff der effektiven Dimension gründet sich auf die folgende Betrachtungsweise: Es sei zunächst der einfache Fall $\mathbf{\Sigma} = \boldsymbol{I}\sigma^2$ angenommen. Dann gilt

$$\Delta^{*2} = \frac{(\boldsymbol{\mu'\mu})^2}{\sigma^2 \boldsymbol{\mu'\mu}} = \boldsymbol{\mu'}\mathbf{\Sigma}^{-1}\boldsymbol{\mu} = \Delta^2 \tag{8.8}$$

(vgl. (8.5)), d. h., $\Delta^* = \Delta$, und

$$p^* = \frac{\sigma^4 p}{\sigma^4} = p. \tag{8.9}$$

Das heißt weiter, es gilt hier

$$T_{\boldsymbol{\mu}} = 1 + \frac{4p}{\Delta^2 n}. \tag{8.10}$$

Der Fehler (8.7) hängt dann von p und n nur über das Verhältnis $\frac{p}{n}$ ab. Im allgemeinen Fall, wenn $\mathbf{\Sigma}$ nicht notwendig von der Form $\boldsymbol{I}\sigma^2$ ist, wird die Dimension p durch p^* ersetzt. Auf gleiche Weise hängt der Fehler (8.7) dann vom Verhältnis $\frac{p^*}{n}$ ab. Daher wird p^* auch als *effektive Dimension* bezeichnet (vgl. [42]).

In ähnlicher Weise erhält Raudys [42] auch für den Fall $n_1 = n_2$ für die Fisher'sche Lineare Diskriminanzanalyse (LDA, vgl. Abschnitt 3.3) die doppelt asymptotische Näherung

$$r(Q, \mathrm{LDA}) \approx \Phi\left(-\frac{\Delta}{2}\frac{1}{\sqrt{T_{\boldsymbol{\mu}} T_{\mathbf{\Sigma}}}}\right) \tag{8.11}$$

mit

$$T_{\boldsymbol{\mu}} = 1 + \frac{4p}{\Delta^2 n}$$

und

$$T_{\mathbf{\Sigma}} = 1 + \frac{p}{n-p}.$$

Die Korrekturterme $T_{\boldsymbol{\mu}}$ und $T_{\mathbf{\Sigma}}$ beschreiben dabei den Fehler, der durch die Schätzung der Parameter $\boldsymbol{\mu}$ und $\mathbf{\Sigma}$ entsteht.

Für die Regularisierte Diskriminanzanalyse (RDA), die die Ridge-Methode verwendet und so ja gewissermaßen „zwischen“ diesen beiden Verfahren liegt, haben schließlich Raudys und Skurichina [43] folgende — ebenfalls doppelt asymptotische — Näherung angegeben:

$$r(Q, \mathrm{RDA}) \approx \Phi\left\{-\frac{\Delta}{2}\left(\frac{n}{n-p}\left(1 + \frac{4p}{n\Delta^2}\right)\right)^{-\frac{1}{2}}\frac{\sqrt{1+\lambda B}}{1+\lambda C}\right\}. \tag{8.12}$$

B und C sind dabei wie folgt gegeben:

$$B = \frac{2}{1-y}\beta_2 + \frac{\mathrm{tr}\mathbf{\Delta}^{-1}}{n}, \qquad C = \frac{1}{1-y}\beta_1, \qquad y = \frac{p}{n}$$

$$\beta_i = \frac{\boldsymbol{m'}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}}{\Delta^2}\alpha_i + \frac{\mathrm{tr}\boldsymbol{\Lambda}^{-1}}{n-p} \ (i = 1, 2)$$

$$\alpha_1 = 1, \qquad \alpha_2 = \left(1 + \frac{4\mathrm{tr}\boldsymbol{\Lambda}^{-1}}{n\boldsymbol{m'}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}}\right)\left(1 + \frac{4p}{n\Delta^2}\right)^{-1}$$

$\boldsymbol{\Lambda}$ ist hier die Diagonalmatrix mit den Eigenwerten von $\boldsymbol{\Sigma}$, so dass $\boldsymbol{\Gamma\Sigma\Gamma'} = \boldsymbol{\Lambda}$ gilt ($\boldsymbol{\Gamma}$ ist die Orthogonalmatrix der zugehörigen Eigenvektoren), und es gilt $\boldsymbol{m} := \boldsymbol{\Gamma'}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$.

Durch Differenzieren von (8.12) nach λ und Nullsetzen der Ableitung erhält man als optimalen Parameter

$$\lambda_{opt} = \frac{B - 2C}{BC}. \tag{8.13}$$

Damit ist eine asymptotische Berechnungsformel für den optimalen Wert für λ gegeben. Der Parameter λ ist hierbei unabhängig von den Beobachtungen bzw. von der Stichprobenkovarianzmatrix $\boldsymbol{S}$, d. h. hängt nur vom „wahren“ Verteilungsparameter ab. Sind die Parameter $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ und $\boldsymbol{\Sigma}$ nicht bekannt, so ist es sicher auch hier möglich, diese durch die entsprechenden Schätzungen zu ersetzen. Allerdings ist hier zu bemerken, dass Ausdruck (8.12) nur für den Fall $n > p$ definiert ist, so dass (8.13) auch nur für diesen Fall gilt.

8.3 Ein neuer asymptotischer Ansatz

Ausgehend von dem Ansatz von Raudys und Skurichina wird hier auf ähnliche Weise eine Berechnungsformel für die Berechnung des Klassifikationsfehlers hergeleitet. Im Unterschied zu dem Ansatz von Raudys und Skurichina wird hier keine doppelte Asymptotik, sondern nur der einfache Grenzübergang $n \to \infty$ betrachtet. Die hergeleitete Berechnungsformel ist andererseits aber nicht durch die Bedingung $n > p$ eingeschränkt. Zugleich ergibt sich auch wieder eine Abschätzung für die Wahl eines optimalen Parameters λ_0.

Raudys und Skurichina [43] verwenden für die Ridge-Regel die mittels Taylor-Entwicklung an der Stelle $\lambda = 0$ erhaltene Näherung

$$(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1} \approx \boldsymbol{S}^{-1} - \lambda\boldsymbol{S}^{-2} \tag{8.14}$$

unter Vernachlässigung der Terme höherer Ordnung. Um wieder eine geeignete Approximation mittels Reihenentwicklung zu bekommen, erweist sich zunächst eine Umparametrisierung als zweckmäßig. Der Term $(\boldsymbol{S}+\lambda)^{-1}$ wird zunächst mit $1 + \lambda$ multipliziert — dies ändert nichts an der Diskrimination. Man erhält den Term

$$(1 + \lambda)\,(\boldsymbol{S} + \lambda\boldsymbol{I})^{-1} = \left(\frac{1}{1+\lambda}\,\boldsymbol{S} + \frac{\lambda}{1+\lambda}\boldsymbol{I}\right)^{-1}. \tag{8.15}$$

Mittels $\tilde{\lambda} := \frac{1}{1+\lambda}$ erfolgt dann eine Umparametrisierung. Man erhält:

$$(\tilde{\lambda}\, \boldsymbol{S} + (1-\tilde{\lambda})\boldsymbol{I})^{-1} = (\, \tilde{\lambda}(\boldsymbol{S}-\boldsymbol{I}) + \boldsymbol{I})^{-1}. \tag{8.16}$$

Für hinreichend kleine Werte von $\tilde{\lambda}$ lässt sich die erhaltene Matrix als eine unendliche Reihe darstellen. Dies lässt sich direkt mit folgendem Lemma zeigen. (Mit $\mathcal{S}(p)$ wird hier wieder die Menge der symmetrischen $p \times p$-Matrizen bezeichnet.)

Lemma 8.1 *Sei $\boldsymbol{M} \in \mathcal{S}(p)$ und $\|.\|$ eine Norm auf $\mathcal{S}(p)$, für die die Bedingung*

$$\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\| \tag{8.17}$$

gilt, $\tilde{\lambda} \in \mathbb{R}$. Falls die Bedingung

$$|\tilde{\lambda}| < \frac{1}{\|\boldsymbol{M}\|} \tag{8.18}$$

erfüllt ist, gilt die folgende Identität:

$$(\tilde{\lambda}\boldsymbol{M} + \boldsymbol{I})^{-1} = \sum_{k=0}^{\infty} (-\tilde{\lambda})^k \boldsymbol{M}^k. \tag{8.19}$$

Beweis: Zunächst wird die Identität der so genannten Neumannschen Reihe gezeigt (nach [47, Chapter 5], [10, Bemerkung 2.9]).

Sei $\boldsymbol{C}$ eine Matrix $(p \times p)$ mit $\|\boldsymbol{C}\| < 1$, es gelte (8.17). Für beliebiges $n \in \mathbb{N}$ gilt:

$$\begin{aligned}
&(\boldsymbol{I} + \boldsymbol{C} + \boldsymbol{C}^2 + \ldots + \boldsymbol{C}^n)(\boldsymbol{I} - \boldsymbol{C}) \\
&= \boldsymbol{I} + \boldsymbol{C} + \boldsymbol{C}^2 + \ldots + \boldsymbol{C}^n - (\boldsymbol{C} + \boldsymbol{C}^2 + \boldsymbol{C}^3 + \ldots \boldsymbol{C}^{n+1}) && (8.20)\\
&= \boldsymbol{I} - \boldsymbol{C}^{n+1}. && (8.21)
\end{aligned}$$

Gleiches gilt für die Multiplikation von $(\boldsymbol{I} - \boldsymbol{C})$ von links. Wegen $\|\boldsymbol{A}\| < 1$ gilt mit Bedingung (8.17)

$$\lim_{n\to\infty} \|\boldsymbol{A}^n\| = 0. \tag{8.22}$$

Aus der Normkonvergenz folgt die Konvergenz

$$\lim_{n\to\infty} \boldsymbol{A}^n = \boldsymbol{0}\,. \tag{8.23}$$

Damit gilt für $n \to \infty$:

$$\Big(\sum_{n=0}^{\infty} \boldsymbol{C}^n\Big)(\boldsymbol{I} - \boldsymbol{C}) = \boldsymbol{I}, \tag{8.24}$$

damit folgt die Identität der Neumannschen Reihe:

$$\sum_{n=0}^{\infty} \boldsymbol{C}^n = (\boldsymbol{I} - \boldsymbol{C})^{-1}\,. \tag{8.25}$$

Ist Bedingung (8.18) erfüllt, so gilt

$$\|\tilde{\lambda}\boldsymbol{M}\| < 1. \tag{8.26}$$

Setzt man nun

$$\boldsymbol{C} := -\tilde{\lambda}\boldsymbol{M}, \tag{8.27}$$

so gilt $\|\boldsymbol{C}\| < 1$; die Neumannsche Reihe hat die Gestalt

$$(\tilde{\lambda}\boldsymbol{M} + \boldsymbol{I})^{-1} = (\boldsymbol{I} - \boldsymbol{C})^{-1} = \sum_{k=0}^{\infty} \boldsymbol{C}^k = \sum_{k=0}^{\infty} (-\tilde{\lambda})^k \boldsymbol{M}^k. \tag{8.28}$$

□

Mit $\boldsymbol{M} := \boldsymbol{S} - \boldsymbol{I}$ erhält man dann die Reihen-Entwicklung

$$(\tilde{\lambda}(\boldsymbol{S} - \boldsymbol{I}) + \boldsymbol{I})^{-1} = \sum_{k=0}^{\infty} \tilde{\lambda}^k (\boldsymbol{I} - \boldsymbol{S})^k. \tag{8.29}$$

Verwendet man nur die ersten beiden Glieder, erhält man die Approximation

$$(\tilde{\lambda}(\boldsymbol{S} - \boldsymbol{I}) + \boldsymbol{I})^{-1} \approx \boldsymbol{I} + \tilde{\lambda}(\boldsymbol{I} - \boldsymbol{S}). \tag{8.30}$$

Mit Rücktransformation mittels

$$\tilde{\lambda} = \frac{1}{1+\lambda} \tag{8.31}$$

erhält man

$$\left(\frac{1}{1+\lambda}\,\boldsymbol{S} + \frac{\lambda}{1+\lambda}\boldsymbol{I}\right)^{-1} \approx \boldsymbol{I} + \frac{1}{1+\lambda}(\boldsymbol{I} - \boldsymbol{S}), \tag{8.32}$$

weitere Multiplikation mit $1+\lambda$ ergibt die näherungsweise Diskriminanzfunktion:

$$\tilde{h}^{\mathrm{RDA}}(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\right)' ((\lambda+2)\boldsymbol{I} - \boldsymbol{S})(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}). \tag{8.33}$$

Das Klassifikationsverfahren, das man erhält, wenn man (8.33) als Diskriminanzfunktion verwendet (bei Entscheidung für Klasse 1, falls $\tilde{h}(\boldsymbol{x}) \geq 0$ ist, für Klasse 2 sonst), sei mit RDA* bezeichnet. Gesucht ist jetzt eine Berechnungsformel für das asymptotische Risiko $r_\infty(Q, \mathrm{RDA}^*)$ (zu gegebener Verteilung Q auf $\{1,2\}$). Dazu wird im folgenden Satz ein Resultat angegeben. Zunächst wird dafür der Begriff der stochastischen Konvergenz erklärt.

Definition 8.1 (Stochastische Konvergenz) *Gegeben seien Zufallsvektoren* $\boldsymbol{X}_n = (X_{1,n}, \ldots, X_{k,n})' : (M, \mathcal{A}, P) \to (\mathbb{R}^k, \mathcal{B}^k)$, $n \geq 1$. *Die Folge* $(\boldsymbol{X}_n)_{n\geq 1}$ *konvergiert stochastisch oder in Wahrscheinlichkeit gegen den Vektor* $\boldsymbol{c} = (c_1, \ldots, c_k)'$, *wenn für alle* $\varepsilon > 0$

$$\lim_{n\to\infty} P(\{|\boldsymbol{X}_n - \boldsymbol{c}| > \varepsilon\}) = 0 \tag{8.34}$$

gilt. (Bezeichnung: $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{c}$*)*

Satz 8.1 *Gegeben sei das statistische Modell (3.14), Q sei die Gleichverteilung auf $\{1,2\}$. Es gelte zusätzlich, dass $\boldsymbol{\mu} := \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$ kein Eigenvektor von $\boldsymbol{\Sigma}$ zum Eigenwert $\lambda+2$ ist. Das asymptotische Risiko unter $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ (für $n_1 \to \infty, n_2 \to \infty$), $r_\infty(Q, \mathrm{RDA}^*)$, ist dann gegeben durch*

$$r_\infty(Q, \mathrm{RDA}^*) = \Phi\left(-\frac{1}{2} \frac{\boldsymbol{\mu}'((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}'((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\mu}}} \right) \tag{8.35}$$

(Φ ist die Verteilungsfunktion der Standardnormalverteilung $N(0,1)$).

Beweis: $\bar{\boldsymbol{x}}^{(1)}$, $\bar{\boldsymbol{x}}^{(2)}$ und $\boldsymbol{S}$ sind konsistente Schätzer für $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ und $\boldsymbol{\Sigma}$. Das heißt, es gilt

$$\bar{\boldsymbol{x}}^{(1)} \xrightarrow{P} \boldsymbol{\mu}^{(1)} \quad (n_1 \to \infty), \tag{8.36}$$

$$\bar{\boldsymbol{x}}^{(2)} \xrightarrow{P} \boldsymbol{\mu}^{(2)} \quad (n_2 \to \infty) \tag{8.37}$$

und

$$\boldsymbol{S} \xrightarrow{P} \boldsymbol{\Sigma} \quad (n_1, n_2 \to \infty). \tag{8.38}$$

Nach Cramér [7] sind die Grenzwerte in Wahrscheinlichkeit von Summen, Differenzen und Produkten von Zufallsvariablen jeweils die Summen, Differenzen bzw. Produkte der Grenzwerte in Wahrscheinlichkeit der Zufallsvariablen. Damit gilt

$$(\lambda+2)\boldsymbol{I} - \boldsymbol{S} \xrightarrow{P} (\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma}, \tag{8.39}$$

$$((\lambda+2)\boldsymbol{I} - \boldsymbol{S})(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) \xrightarrow{P} ((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \tag{8.40}$$

sowie

$$(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})'((\lambda+2)\boldsymbol{I} - \boldsymbol{S})(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) \xrightarrow{P} (\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})'((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \tag{8.41}$$

(jeweils für $n_1, n_2 \to \infty$). Die asymptotische Verteilung P_j $(j = 1,2)$ von $\tilde{h}^{\mathrm{RDA}}$ für $n_1, n_2 \to \infty$ ist also $N(\beta_{\tilde{h}}^{(j)}, \sigma_{\tilde{h}}^2)$, wobei der Erwartungswert $\beta_{\tilde{h}}^{(j)}$ $(j = 1,2)$ durch

$$\mu_{\tilde{h}}^{(j)} = -(-1)^j \frac{1}{2} (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \tag{8.42}$$

und die Varianz jeweils durch

$$\sigma_{\tilde{h}}^2 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \tag{8.43}$$

gegeben ist. Die asymptotische Verteilung $N(\beta_{\tilde{h}}^{(j)}, \sigma_{\tilde{h}}^2) = N_{\beta_{\tilde{h}}^{(j)}, \sigma_{\tilde{h}}^2}$ $(j = 1,2)$ ist hier stets auf dem Messraum $(\mathbb{R}, \mathcal{B})$ aufgefasst; für $B \in \mathcal{B}$ sei mit $N_{\beta_{\tilde{h}}^{(j)}, \sigma_{\tilde{h}}^2}(B)$ das Wahrscheinlichkeitsmaß von B bezüglich $N_{\beta_{\tilde{h}}^{(j)}, \sigma_{\tilde{h}}^2}$ bezeichnet $(j = 1,2)$.

Ist nun Q die Gleichverteilung, d. h. betragen die Wahrscheinlichkeiten $q_j = Q(\{j\}) = \frac{1}{2}$ $(j = 1, 2)$, so erhält man für die asymptotische Fehlerwahrscheinlichkeit (das bayessche Risiko) $r_\infty(Q, \mathrm{RDA}^*)$:

$$\begin{aligned}
r_\infty(Q, \mathrm{RDA}^*) &= \frac{1}{2}P_1(\{\tilde{h} < 0\}) + \frac{1}{2}P_2(\{\tilde{h} \geq 0\}) \\
&= \frac{1}{2}N_{\beta_{\tilde{h}}^{(1)}, \sigma_{\tilde{h}}^2}(\{x \in \mathbb{R} : x < 0\}) \\
&\quad + \frac{1}{2}N_{\beta_{\tilde{h}}^{(2)}, \sigma_{\tilde{h}}^2}(\{x \in \mathbb{R} : x \geq 0\}) \\
&= \frac{1}{2}N_{0,1}\left(\left\{x \in \mathbb{R} : x < \frac{-\mu_{\tilde{h}}^{(1)}}{\sqrt{\sigma_{\tilde{h}}^2}}\right\}\right) \\
&\quad + \frac{1}{2}N_{0,1}\left(\left\{x \in \mathbb{R} : x \geq \frac{-\mu_{\tilde{h}}^{(2)}}{\sqrt{\sigma_{\tilde{h}}^2}}\right\}\right) \\
&= \frac{1}{2}\Phi\left(-\frac{\mu_{\tilde{h}}^{(1)}}{\sqrt{\sigma_{\tilde{h}}^2}}\right) + \frac{1}{2}\left(1 - \Phi\left(-\frac{\mu_{\tilde{h}}^{(2)}}{\sqrt{\sigma_{\tilde{h}}^2}}\right)\right) \\
&= \frac{1}{2}\Phi\left(-\frac{\mu_{\tilde{h}}^{(1)}}{\sqrt{\sigma_{\tilde{h}}^2}}\right) + \frac{1}{2}\Phi\left(-\frac{\mu_{\tilde{h}}^{(1)}}{\sqrt{\sigma_{\tilde{h}}^2}}\right) \\
&= \Phi\left(-\frac{1}{2}\frac{\boldsymbol{\mu'}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu'}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\mu}}}\right) \quad (8.44)
\end{aligned}$$

(mit $\boldsymbol{\mu} := \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$). □

Damit ist eine Darstellung für das asymptotische Risiko der näherungsweisen Ridge-Regel gegeben. Gesucht ist jetzt der optimale Parameter λ_{opt}, der (8.35) minimiert. Dazu ist im folgenden Satz ein Resultat angegeben.

Satz 8.2 *Es sei* $\boldsymbol{\Sigma} \in \mathrm{PD}(p)$, $\boldsymbol{0} \neq \boldsymbol{\mu} \in \mathbb{R}^p$ *sei kein Eigenvektor von* Σ. *Das asymptotische Risiko (8.35) wird minimiert mit der Wahl* $\lambda = \lambda_0$, *gegeben durch*

$$\lambda_0 = \frac{\boldsymbol{\mu'\mu\mu'\Sigma^3\mu} - \boldsymbol{\mu'\Sigma\mu\mu'\Sigma^2\mu}}{\boldsymbol{\mu'\mu\mu'\Sigma^2\mu} - (\boldsymbol{\mu'\Sigma\mu})^2} - 2. \quad (8.45)$$

Beweis: Wir betrachten die Funktion $f : \mathbb{R} \to [0, 1]$, gegeben durch

$$f(\lambda) = \Phi\left(-\frac{1}{2}\frac{\boldsymbol{\mu'}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu'}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}((\lambda+2)\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{\mu}}}\right). \quad (8.46)$$

Definiere zusätzlich die Funktion z mit

$$z(\lambda) = -\frac{1}{2}\frac{\boldsymbol{\mu'}((\lambda+2)\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu'}((\lambda+2)\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\Sigma}((\lambda+2)\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}}}, \tag{8.47}$$

so dass dann $f(\lambda) = \Phi(z(\lambda))$ gilt.

Die Verteilungsfunktion Φ ist streng monoton wachsend. Eine Minimierung von $\Phi(z(\lambda))$ bezüglich λ ist damit äquivalent zur Minimierung von $z(\lambda)$ bezüglich λ. Mit der Substitution $\lambda^* := \lambda + 2$ erhält man die folgende Funktion $z(\lambda^*)$:

$$z(\lambda^*) = -\frac{1}{2}\frac{\boldsymbol{\mu'}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu'}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\Sigma}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}}}. \tag{8.48}$$

Mit Hilfe der Differentialrechnung kann nun das Minimum bestimmt werden.

Die Ableitung von $z(\lambda^*)$ ist wie folgt:

$$\frac{\mathrm{d}z(\lambda^*)}{\mathrm{d}\lambda^*} = \frac{-\frac{1}{2}\boldsymbol{\mu'}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu'}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\Sigma}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}}} + \frac{\frac{1}{2}\boldsymbol{\mu'}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}}{(\boldsymbol{\mu'}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\Sigma}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu})^{\frac{3}{2}}} \tag{8.49}$$

Damit gilt:

$$\begin{aligned}
&\frac{\mathrm{d}z(\lambda^*)}{\mathrm{d}\lambda^*} \leq 0 \\
&\Leftrightarrow \quad \frac{1}{2}\boldsymbol{\mu'}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu} \leq \frac{1}{2}\boldsymbol{\mu'}\boldsymbol{\mu}(\boldsymbol{\mu'}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\Sigma}(\lambda^*\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{\mu}) \\
&\Leftrightarrow \quad \lambda^{*2}\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu} - \lambda^*\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu} - \lambda^*\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu} + \boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu} \\
&\qquad \leq \lambda^{*2}\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu} - 2\lambda^*\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu} + \boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^3\boldsymbol{\mu} \qquad (8.50) \\
&\Leftrightarrow \quad \lambda^*(\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu} - (\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu})^2) \leq \boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^3\boldsymbol{\mu} - \boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu}. \qquad (8.51)
\end{aligned}$$

Auf den Term $\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu} - (\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu})^2$, der auf der linken Seite innerhalb der Klammern steht, lässt sich nun die Cauchy-Schwarz'sche Ungleichung anwenden. Es gilt:

$$(\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu})^2 \leq \boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{\mu}. \tag{8.52}$$

Die Gleichheit gilt genau dann, wenn $\boldsymbol{\Sigma}\boldsymbol{\mu} = \gamma\boldsymbol{\mu}$ für ein $\gamma \in \mathbb{R}$ gilt (vgl. z. B. [48, Abschnitt 29]). Dies wäre genau dann der Fall, wenn $\boldsymbol{\mu}$ Eigenvektor von $\boldsymbol{\Sigma}$ wäre. Da dies nach Voraussetzung nicht so ist, gilt in (8.52) die strikte Ungleichung. Damit folgt

$$\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu} - (\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu})^2 > 0. \tag{8.53}$$

Für die erste Ableitung von λ^* folgt damit

$$\frac{\mathrm{d}z(\lambda^*)}{\mathrm{d}\lambda^*} \leq 0 \Leftrightarrow \lambda^* \leq \frac{\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^3\boldsymbol{\mu} - \boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu}}{\boldsymbol{\mu'}\boldsymbol{\mu}\boldsymbol{\mu'}\boldsymbol{\Sigma}^2\boldsymbol{\mu} - (\boldsymbol{\mu'}\boldsymbol{\Sigma}\boldsymbol{\mu})^2}. \tag{8.54}$$

Analog lässt sich die entsprechende Bedingung für die Gleichheit herleiten; es gilt:

$$\frac{\mathrm{d}z(\lambda^*)}{\mathrm{d}\lambda^*} = 0 \Leftrightarrow \lambda^* = \frac{\boldsymbol{\mu'\mu\mu'\Sigma^3\mu} - \boldsymbol{\mu'\Sigma\mu\mu'\Sigma^2\mu}}{\boldsymbol{\mu'\mu\mu'\Sigma^2\mu} - (\boldsymbol{\mu'\Sigma\mu})^2}. \tag{8.55}$$

Damit folgt nun: Falls $\boldsymbol{\mu}$ kein Eigenvektor von $\boldsymbol{\Sigma}$ ist, ist λ_0^* Minimalstelle von $z(\lambda^*)$. Mit Rücktransformation $\lambda = \lambda^* - 2$ erhält man als Minimalstelle:

$$\lambda_0 = \frac{\boldsymbol{\mu'\mu\mu'\Sigma^3\mu} - \boldsymbol{\mu'\Sigma\mu\mu'\Sigma^2\mu}}{\boldsymbol{\mu'\mu\mu'\Sigma^2\mu} - (\boldsymbol{\mu'\Sigma\mu})^2} - 2. \tag{8.56}$$

□

Es bleibt nun zu untersuchen, ob bzw. unter welchen Bedingungen λ_0 im Konvergenzbereich der Neumannschen Reihe (8.19) liegt und unter welchen Bedingungen λ_0 positiv ist. Diese Fragestellungen sind unter allgemeinen Bedingungen schwierig zu beantworten; für spezielle Verteilungsannahmen lassen sich dazu jedoch Untersuchungen anstellen. Es seien folgende zusätzliche Bedingungen erfüllt:

1. Die Differenz der Erwartungswerte $\boldsymbol{\mu}$ ist von der Form $\boldsymbol{\mu} = (m, \ldots, m)'$, $m \in \mathbb{R} \setminus \{0\}$.

2. Die Kovarianzmatrix $\boldsymbol{\Sigma} \in \mathrm{PD}(p)$ ist von der Form $\boldsymbol{\Sigma} = [\sigma_{ij}]_{i,j=1,\ldots,p}$ mit

$$\sigma_{ii} = \begin{cases} v_1 > 0 & \text{für } i = 1, \ldots, q \\ v_2 > 0 & \text{für } i = q+1, \ldots, p \end{cases} \tag{8.57}$$

   (mit $q \leq p$), und $\sigma_{ij} = c \geq 0$, falls $i \neq j$ $(i, j = 1, \ldots, p)$.

Der Fall $v_1 = v_2$ ist hier von besonderem Interesse; aus technischen Gründen wird aber zunächst die Unterscheidung vorgenommen. Im folgenden Satz sind nun die dazu erzielten Resultate zusammengefasst.

Satz 8.3 *Es gelten die Bedingungen 1. und 2. Gilt $v_1 = v_2$, so sind in (8.45) sowohl Zähler als auch Nenner 0. Ist v_1 fest und λ_0 eine Funktion der Variablen v_2, so gilt*

$$\lim_{v_2 \to v_1} \lambda_0(v_2) = 2v_1 + (p-2)c - 2. \tag{8.58}$$

Beweis: Es gelten die Voraussetzungen 1. und 2. Der Term $\boldsymbol{\mu'\Sigma\mu}$ hat dann die Gestalt

$$\boldsymbol{\mu'\Sigma\mu} = m^2(qv_1 + rv_2 + p(p-1)c) \tag{8.59}$$

(mit $r := p - q$). Weiterhin gilt

$$\boldsymbol{\mu'\Sigma^2\mu} = m^2(qv_1^2 + rv_2^2 + p(p-1)^2c^2 + 2q(q-1)v_1c + 2r(r-1)v_2c). \tag{8.60}$$

Für $\boldsymbol{\mu'\Sigma}^3\boldsymbol{\mu}$ erhält man schließlich nach Zusammenfassen den Ausdruck

$$\begin{aligned}\boldsymbol{\mu'\Sigma}^3\boldsymbol{\mu} &= m^2[qv_1^3 + rv_2^3 + (q^4 + r^4 - 3q^3 - 3r^3 + 4q^3r + 4qr^3 + 6q^2r^2 \\ &\quad -9q^2r - 9qr^2 + 3q^2 + 3r^2 + 6qr - q - r)c^3 \\ &\quad +(3q^3 - 6q^2 + 6q^2r + 3qr^2 - 6qr + 3q)v_1c^2 \\ &\quad +(3r^3 - 6r^2 + 6qr^2 + 3q^2r - 6qr + 3r)v_2c^2 \\ &\quad +(3q^2 - 3q + 2qr)v_1^2c + (3r^2 - 3r + 2qr)v_2^2c + 2qrv_1v_2c].\end{aligned} \tag{8.61}$$

Es können jetzt alle in (8.45) vorkommenden Produkte berechnet werden (mit $\boldsymbol{\mu'\mu} = pm^2$). Auf eine ausführliche Darstellung der recht aufwändigen elementaren Berechnungen soll hier verzichtet werden. Für (8.45) erhält man schließlich als Zähler

$$\begin{aligned}\text{Zähler} &= m^4[qrv_1^3 + qrv_2^3 - qrv_1^2v_2 - qrv_1v_2^2 + (q^2r + qr^2 - 2qr)v_1^2c \\ &\quad +(q^2r + qr^2 - 2qr)v_2^2c + (-2q^2r - 2qr^2 + 4qr)v_1v_2c]\end{aligned} \tag{8.62}$$

und als Nenner

$$\text{Nenner} = m^4[qrv_1^2 + qrv_2^2 - 2qrv_1v_2]. \tag{8.63}$$

Man sieht jetzt leicht, dass sowohl Zähler als auch Nenner Null werden, wenn $v_1 = v_2$ gilt.

Sei jetzt angenommen, dass die Kovarianz c und die Varianz v_1 fest seien. Wir betrachten jetzt λ_0 als eine Funktion von v_2. Für $v_2 \to v_1$ gilt dann (mit $f(v_2) := \text{Zähler}(v_2)$ und $g(v_2) := \text{Nenner}(v_2)$):

$$\lim_{v_2 \to v_1} \lambda_0^*(v_2) = \lim_{v_2 \to v_1} \frac{f(v_2)}{g(v_2)}. \tag{8.64}$$

Mit zweimaliger Anwendung des Satzes von l'Hospital erhält man:

$$\begin{aligned}\lim_{v_2 \to v_1} \lambda_0^*(v_2) &= \lim_{v_2 \to v_1} \frac{f''(v_2)}{g''(v_2)} \\ &= \lim_{v_2 \to v_1} \frac{6qrv_2 - 2qrv_1 + (2q^2r + 2qr^2 - 4qr)c}{2qr} \\ &= \lim_{v_2 \to v_1} 3v_2 - v_1 + (q + r - 2)c \\ &= 2v_1 + (p - 2)c.\end{aligned} \tag{8.65}$$

Die Funktion $\lambda_0^*(v_2)$ hat also hier eine Unstetigkeitsstelle, ist aber in einer Umgebung von $v_2 = v_1 = v$ stetig, der Grenzwert existiert. Somit kann in dieser Situation der Wert für λ_0^* gemäß

$$\lambda_0^* = 2v + (p - 2)c \tag{8.66}$$

(bzw. für λ_0 gemäß

$$\lambda_0 = 2v + (p - 2)c - 2) \tag{8.67}$$

berechnet werden. □

Man sieht jetzt leicht, dass unter den genannten Voraussetzungen λ_0 genau dann positiv ist, wenn

$$2v + (p-2)c > 2 \tag{8.68}$$

gilt. Dies ist gesichert, wenn die Varianzen, die Kovarianzen und die Dimension p nicht zu klein sind.

Es verbleibt die Frage, unter welchen Bedingungen λ_0 innerhalb des Konvergenzradius der Reihe (8.19) liegt. Mit der Identität (8.31) lässt sich für $\boldsymbol{M} = \boldsymbol{S}$ die Konvergenzbedingung (8.18) schreiben als

$$\left|\frac{1}{1+\lambda}\right| < \frac{1}{\|\boldsymbol{S} - \boldsymbol{I}\|}. \tag{8.69}$$

Dies ist äquivalent zu

$$|1+\lambda| > \|\boldsymbol{S} - \boldsymbol{I}\|. \tag{8.70}$$

Als Resultat lässt sich hierzu die Aussage des folgenden Satzes formulieren. Um Verwechslungen mit dem Ridge-Parameter λ_0 zu vermeiden, werden hier abweichend die Eigenwerte von $\boldsymbol{\Sigma}$ mit $\delta_1, \ldots, \delta_p$ und die Eigenwerte von $\boldsymbol{S}$ mit $d_1, \ldots, d_p$ bezeichnet.

Satz 8.4 *Es gelten die Bedingungen 1. und 2., es sei $v_1 = v_2 = v$. Sei $\boldsymbol{S} \in$ PSD(p). Mit d_1 sei der größte Eigenwert von $\boldsymbol{S}$ und mit δ_1 der größte Eigenwert von $\boldsymbol{\Sigma}$ bezeichnet. Falls die Bedingung*

$$d_1 < \delta_1 + v - c \tag{8.71}$$

erfüllt ist, liegt λ_0 im Konvergenzbereich der Potenzreihe (8.19).

Beweis: Sind d_i $(i = 1, \ldots, p)$ die Eigenwerte von $\boldsymbol{S}$, lassen sich die Eigenwerte von $\boldsymbol{S} - \boldsymbol{I}$ gemäß $d_i^* = d_i - 1$ $(i = 1, \ldots, p)$ berechnen. Für $\boldsymbol{S} \in$ PSD(p) gilt dann

$$-1 \leq d_i^* \leq d_1 - 1 \ (i = 1, \ldots, p), \tag{8.72}$$

wobei d_1 der größte Eigenwert von $\boldsymbol{S}$ ist. Der größte Betrag der Eigenwerte und damit die Spektralnorm von $\boldsymbol{S} - \boldsymbol{I}$ ist dann durch

$$\|\boldsymbol{S} - \boldsymbol{I}\| = \max\{d_1 - 1; 1\} \tag{8.73}$$

gegeben. Für $\lambda > 0$ ist die Bedingung $|1+\lambda| > 1$ immer erfüllt. Bedingung (8.70) ist also für $\lambda > 0$ erfüllt, falls

$$\lambda > d_1 - 2 \tag{8.74}$$

gilt.

Für λ im Bereich $-2 \leq \lambda \leq 0$ gilt $|1+\lambda| \leq 1$ und damit

$$|1+\lambda| \leq \|\boldsymbol{S} - \boldsymbol{I}\|. \tag{8.75}$$

Für $\lambda < -2$ gilt

$$|1 + \lambda| = -\lambda - 1 > 1. \tag{8.76}$$

Dann liegt λ innerhalb des Konvergenzbereichs, wenn

$$-\lambda > d_1 \tag{8.77}$$

gilt.

Wird λ_0 nach (8.67) berechnet, so gilt stets $\lambda_0 \geq -2$. Der Wert λ_0 kann dann also nur im Konvergenzbereich liegen, wenn λ_0 positiv ist.

Unter den Voraussetzungen 1. und 2. hat die Matrix $\mathbf{\Sigma}$ mit Varianzen v und Kovarianzen c die Eigenwerte

$$\delta_1 = v + (p-1)c \tag{8.78}$$

und

$$\delta_i = v - c \ (i = 2, \ldots, p). \tag{8.79}$$

Für den nach (8.67) berechneten Wert $\lambda_0 = 2v + (p-2)c - 2$ gilt dann

$$\lambda_0 = \delta_1 + v - c - 2. \tag{8.80}$$

λ_0 erfüllt demnach die Bedingung (8.70), falls $d_1 < \lambda_0 + 2$, d. h.

$$d_1 < \delta_1 + v - c \tag{8.81}$$

gilt. □

Unter diesen speziellen Voraussetzungen ist also für λ_0 die Gültigkeit der Approximationsformel (8.32) gesichert, wenn die Kovarianzen nicht allzu groß sind.

8.4 Simulationsexperimente

Wie wir gesehen haben, gibt es zwar für einige Klassifikationsverfahren Berechnungsformeln, die unter bestimmten Bedingungen eine näherungsweise Berechnung des Klassifikationsfehlers erlauben. Jedoch setzen diese zumeist die genaue Kenntnis der Parameter voraus und sind daher nur bedingt anwendbar. Außerdem gibt es oft auch andere Einschränkungen, wie bei der Näherungsformel (8.12) etwa die Bedingung $n - 2 \geq p$. Aus diesem Grund wird das Risiko der Klassifikation meistens durch Anwendung der Diskriminationsregel(n) auf Teststichproben mit bekannter Klassenzugehörigkeit geschätzt. Die dafür verwendeten Daten sind entweder Beobachtungen aus irgendwelchen Experimenten, oder sie werden durch Simulationsrechnungen erzeugt. Im letzteren Fall werden mittels durch Zufallsgenerator erzeugter Folgen von Pseudozufallszahlen künstlich Daten generiert, die annähernd einer vorgegebenen Verteilung unterliegen. Darauf werden dann die verschiedenen Verfahren zur Klassifikation angewandt, es wird jeweils eine Fehlerrate ermittelt. Dies wird mehrfach wiederholt, und schließlich erhält man die geschätzten Fehlerraten als Mittelwerte aus den Fehlerraten der einzelnen Simulationsläufe.

8.4.1 Verwendetes Modell

Durch Simulationen wurden die Fehlerraten der beiden Verfahren RDA und RDA* bei höheren Merkmalszahlen ermittelt. In jedem Durchlauf wurden dazu wieder Beobachtungen von stochastisch unabhängigen, multivariat normalverteilten Zufallsvektoren $\boldsymbol{X}_\alpha^{(1)} \sim N_p(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ $(\alpha = 1, \ldots, n_1)$ und $\boldsymbol{X}_\alpha^{(2)} \sim N_p(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ $(\alpha = 1, \ldots, n_2)$ als Lernstichproben simuliert, außerdem eine zu klassifizierende Beobachtung eines Zufallsvektors $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma})$ $(j \in \{1, 2\})$. Die Parameter wurden wie folgt bestimmt:

$$\boldsymbol{\mu}^{(1)} = (\,1, \ldots, 1)', \; \boldsymbol{\mu}^{(2)} = (\,0, \ldots, 0)',$$

$$\boldsymbol{\Sigma} = [\sigma_{ij}]_{i,j=1,\ldots,p} \text{ mit } \begin{array}{rcl} \sigma_{ii} & = & v_1 = 1 \text{ für } i = 1, \ldots, p/2, \\ \sigma_{ii} & = & v_2 \text{ für } i = p/2 + 1, \ldots, p, \\ \sigma_{ij} & = & \varrho\sqrt{\sigma_{ii}\sigma_{jj}}, \text{ falls } i \neq j \; (i, j = 1, \ldots, p). \end{array}$$

Die Anzahl der Lernstichprobenelemente war jeweils mit $n_1 = n_2 = n/2$ bestimmt. Für p, n, v_2 und ϱ wurden nacheinander die folgenden Werte verwendet:

- $p = 10$, $p = 100$, $p = 500$
- $n = 20, n = 40, n = 60$
- $v_2 = 1, v_2 = 2, v_2 = 3$
- $\varrho = 0.1$, $\varrho = 0.9$

Für jede dieser Parameterkombinationen wurden jeweils 10 000 Wiederholungen durchgeführt.

Für die Klassifikation wurde die Ridge-Regel RDA mit der Kovarianzschätzung

$$\boldsymbol{S}_{\text{RDA}} = \boldsymbol{S} + \lambda \boldsymbol{I}, \tag{8.82}$$

außerdem die Regel RDA*, die die Diskriminanzfunktion (8.33) verwendet, angewandt. Für λ wurden Zehner-Potenzen mit den Exponenten -2, -1.5, -1, ..., 3.5, 4 (d. h. $\lambda = 10^{-2}, 10^{-1.5}, \ldots, 10^4$) verwendet, zum Vergleich zusätzlich der nach (8.45) bzw. (8.67) ermittelte näherungsweise optimale Wert λ_0 sowie als Grenzfall $\lambda = \infty$. Außerdem wurden jeweils die optimalen Fehlerwahrscheinlichkeiten nach (8.4) berechnet. Exemplarisch sind in Tabelle 8.1 die Ergebnisse für $p = 500$, $n = 20$ und $v_2 = 2$ dargestellt. (Die kleinsten Fehlerraten sind jeweils hervorgehoben.) Die gesamten Ergebnisse dieser Simulationen sind in den Tabellen im Anhang A zu sehen.

$p = 500, n = 20, v_2 = 2$

ϱ	0.1		0.9	
λ_0	75.7		673.3	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.0981	0.8901	**0.0375**	0.6488
$10^{-1.5}$	0.0981	0.8901	0.0379	0.6488
10^{-1}	0.0981	0.8901	0.0389	0.6488
$10^{-0.5}$	0.098	0.89	0.0422	0.6488
10^{0}	0.0978	0.8898	0.0527	0.6488
$10^{0.5}$	0.0981	0.8893	0.0842	0.6488
10^{1}	**0.0971**	0.8874	0.1521	0.6487
$10^{1.5}$	0.0983	0.8563	0.2383	0.6481
10^{2}	0.0988	0.2046	0.3022	0.6468
$10^{2.5}$	0.0999	**0.099**	0.3273	0.6205
10^{3}	0.1006	0.1006	0.3385	**0.3407**
$10^{3.5}$	0.1006	0.1006	0.3422	0.3414
10^{4}	0.1012	0.1012	0.3432	0.3432
∞	0.1015	0.1015	0.3441	0.3441
λ_0	0.0986	0.406	0.3359	0.4269

Tabelle 8.1: Simulationsergebnisse für $p = 500$, $n = 20$, $v_2 = 2$

8.4.2 Wertung der Ergebnisse

Mit Hilfe der Pseudozufallszahlen erreicht man, dass die Daten annähernd die gewünschte Verteilung haben. Da man aber nur endlich viele Wiederholungen durchführen kann, sind die Schätzungen der Fehlerwahrscheinlichkeiten mit einer gewissen Varianz behaftet. Da das Verfahren aber mit den verschiedenen Parameterwerten jeweils an den gleichen simulierten Daten angewandt wurde, kann man mit wesentlich größerer Sicherheit feststellen, wo das Optimum bezüglich der Parameterwahl liegt. Wenn dies auch durch die nötige Diskretisierung des Parameterraums eingeschränkt ist, so lassen sich zumindest Tendenzen und Muster erkennen.

Bei den Simulationen waren die Parameter der Verteilung ja stets bekannt, und so konnte der Wert λ_0 auch direkt aus diesen berechnet werden. Es ist aber zu beachten, dass λ_0 nach den Resultaten in diesem Kapitel die Fehlerrate für das Verfahren RDA* minimiert, nicht aber für das Ridge-Verfahren RDA. Ferner ist zu bedenken, dass λ_0 das asymptotische Risiko minimiert, während bei den Simulationen die Anzahl der Beobachtungen immer endlich ist. Daher ist es nicht unbedingt zu erwarten, dass die hier erreichten Fehlerraten tatsächlich auch bei λ_0 am kleinsten sind.

Sind die Varianzen in allen Variablen gleich ($v_2 = 1$), so lässt sich bei beiden Verfahren kaum ein lokales Minimum der Fehlerrate feststellen. Die Fehlerrate nimmt zunächst mit größer werdendem λ ab, ab einer bestimmten Stelle liegt der Fehler nahezu konstant bei dem Grenzwert für $\lambda \to \infty$ und ändert sich mit weiterer Erhöhung kaum noch. Dies lässt sich vielleicht damit erklären, dass in dieser Situation die Kovarianzmatrix der Einheitsmatrix sehr nahe ist, so dass mit dem Grenzfall $\lambda \to \infty$ stets eine gute Klassifikation erreicht wird. Je größer nun die Unterschiede in den Varianzen sind, desto deutlicher ist auch ein Minimum in der Fehlerrate zu erkennen.

Die beiden Verfahren RDA und RDA* unterscheiden sich in ihren Fehlerraten besonders stark bei kleinen λ-Werten. Dort sind die Fehlerraten von RDA* stets höher. Mit größeren λ-Werten nähern sich beide Verfahren immer mehr an. Die Fehlerraten von RDA* liegen bei kleinen λ-Werten oftmals über 0.5, so dass man da mit Umkehrung des Vorzeichens eine bessere Klassifikation erreichen würde. Auffallend ist hier auch, dass bei einem gewissen λ-Wert das Fehlerniveau sprunghaft stark abfällt, während es im übrigen λ-Bereich nahezu konstant ist. Hier deutet sich wohl die Umkehrung des Vorzeichens an: In der Diskriminanzfunktion (9.2) von RDA* sind die Eigenwerte von $\big((\lambda+2)\boldsymbol{I} - \boldsymbol{S}\big)$ bei kleinen λ-Werten negativ, bei großen λ-Werten positiv. Dazwischen gibt es einen Übergangsbereich, in dem es zu einem Vorzeichenwechsel kommt. Das negative Vorzeichen führt bei kleineren λ-Werten überwiegend zu einer Fehlklassifikation; dadurch kommen solche Fehlerraten von über 0.5 zustande.

Die Unterschiede in den Varianzen der einzelnen Variablen machen sich hauptsächlich bei dem Verfahren RDA bemerkbar. Bei hohen Korrelationen wird bei $v_2 = 2$ und $v_2 = 3$ — im Gegensatz zu dem Fall $v_2 = v_1 = 1$ — das Minimum stets bei sehr kleinen λ-Werten erreicht, besonders bei hohen Dimensionen ($p = 500$).

Ein Einfluss der Lernstichprobenumfänge auf den Wert des optimalen λ, wie es in anderen Arbeiten untersucht wurde [42] [43], ist hier kaum zu beobachten. Offenbar ist dieser so gering, dass er gegenüber den anderen Faktoren nicht ins Gewicht fällt.

Die jeweils mit dem Wert λ_0 erreichten Fehlerraten sind bei RDA in fast allen Fällen niedriger als bei RDA*. Bei gleichen Varianzen in den Variablen liegen hier bei RDA die Fehlerraten nur wenig über den minimalen erreichten Fehlerraten, bei RDA* sind diese Unterschiede zumeist größer. Bei unterschiedlichen Varianzen und hohen Korrelationen dagegen liegen die von RDA* erreichten Fehlerraten meistens näher an den minimalen Fehlerraten. Mit dem Verfahren RDA werden hier sehr kleine Fehlerraten erreicht, teilweise sogar Fehler von 0.0, so dass die mit λ_0 ermittelten Fehler deutlich darüber liegen.

Insgesamt kann man feststellen, dass der Wert λ_0 bei beiden verwendeten Verfahren zwar meistens nicht das exakte Optimum liefert, unter geeigneten Bedingungen aber eine mehr oder weniger gute Näherung.

Bei diesen Simulationen waren die zugrundegelegten Verteilungen bekannt; daher konnte λ_0 direkt daraus berechnet werden. Im nächsten Kapitel, wo die

Verfahren RDA und RDA* mit anderen Klassifikationsverfahren verglichen werden sollen, werden jedoch zur Berechnung von λ_0 statt der wahren Parameter ihre entsprechenden Schätzungen verwendet.

Kapitel 9

Vergleich mit anderen Verfahren

Um die Güte eines Verfahrens bewerten zu können, ist es sinnvoll, es mit anderen Verfahren zu vergleichen. Für ein bestimmtes Klassifikationsproblem, d. h. unter bestimmten Verteilungsvoraussetzungen, können dazu jeweils die Klassifikationsrisiken, die erreichten Fehlerraten, verglichen werden. Wie schon erwähnt, werden diese Fehlerraten zumeist durch Anwendung der Verfahren auf Teststichproben mit bekannter Klassenzugehörigkeit geschätzt.

Es werden im Folgenden die Ergebnisse verschiedener Untersuchungen vorgestellt. Zunächst wurden die verwendeten Daten wieder durch Simulationen mit Pseudozufallszahlen erzeugt, außerdem wurden die verschiedenen Verfahren auch auf vorhandene Beispieldatensätze angewandt.

9.1 Verwendete Verfahren

Sowohl mit den simulierten Daten als auch mit den vorhandenen Datensätzen wurden für die Klassifikation jeweils die folgenden Verfahren verwendet (Erläuterungen zu den einzelnen Verfahren folgen später):

1. Im Verfahren *PCA* wird mittels Hauptkomponentenanalyse eine Dimensionreduzierung durchgeführt; mit den neuen Variablen wird schließlich die lineare Diskriminanzregel nach Fisher (3.19) angewandt.

2. *PLS* (Partial Least Squares, vgl. dazu Abschnitt 4.2.1). Hierbei wird die SAS/STAT-Prozedur PLS verwendet [45]. Die Klassenvariable wird dazu einfach als binäre abhängige Variable modelliert. Die Anzahl der verwendeten Faktoren wird dabei in jedem Durchlauf der Kreuzvalidierung jeweils durch eine zusätzliche Kreuzvalidierungsprozedur ermittelt (Option CV=ONE).

3. Die *Ridge*-Regel nach J. Läuter [33] (vgl. Abschnitt 7.2).

4. Die *Mehrfaktor*-Regel, ebenfalls nach Läuter [33].

5. Die Methode der *Variablen-Korrelations-Analyse VCA*, bei der nach paarweisen Korrelationen Variablenmengen und daraus neue Datenvektoren gebildet werden (vgl. Abschnitt 4.2.2).

   Die folgenden Verfahren sind verschiedene Varianten der *Regularisierten Diskriminanzanalyse (RDA)* bzw. ihrer Modifikation *RDA**. Die RDA-Regeln verwenden jeweils die Diskriminanzfunktion

$$f(\boldsymbol{x}) = \Big(\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\Big)' (\boldsymbol{S} + \lambda \boldsymbol{I})^{-1} (\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}), \tag{9.1}$$

   die RDA*-Regeln die Diskriminanzfunktion

$$f(\boldsymbol{x}) = \Big(\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\Big)' ((\lambda + 2)\boldsymbol{I} - \boldsymbol{S})(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}). \tag{9.2}$$

   Der Parameter λ wird dazu auf unterschiedliche Weise ermittelt:

6. Bei *RDA-det1* ist λ die positive Lösung der Polynomgleichung (7.33),

7. bei *RDA-det2* entsprechend die positive Lösung von (7.56). (Die beiden Verfahren RDA-det1 und RDA-det2 werden nur bei den Simulationen mit $p \leq 20$ Merkmalen verwendet.)

   Bei den Regeln

8. *RDA-*λ_0 und

9. *RDA*-*λ_0 wird als Parameter λ schließlich der nach (8.45) berechnete Wert λ_0 verwendet. Statt $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ und $\boldsymbol{\Sigma}$ werden die entsprechenden Schätzungen aus den Lernstichproben verwendet.

   Bei den Beispieldatensätzen werden schließlich noch die beiden Variationen

10. *RDA-CV* und

11. *RDA*-CV* verwendet (Erklärung siehe unten).

Zu einigen Verfahren seien nun noch im Einzelnen die folgenden Erläuterungen gegeben:

Zu *PCA*: Seien $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}, \boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$ die $n = n_1 + n_2$ Lernstichprobenvektoren. Es wird die Kovarianzmatrix, bezogen auf den Gesamtmittelwertsvektor $\bar{\boldsymbol{x}}$,

$$\boldsymbol{S}_0 = \frac{1}{n-1} \sum_{j=1}^{2} \sum_{\alpha=1}^{n_j} (\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha^{(j)} - \bar{\boldsymbol{x}})', \tag{9.3}$$

verwendet. Betrachtet wird nun das verallgemeinerte Eigenwertproblem

$$\boldsymbol{S}_0 \boldsymbol{G} = \mathbf{Diag}(\boldsymbol{S}_0) \boldsymbol{G} \boldsymbol{\Lambda}. \tag{9.4}$$

Die Hauptkomponenten werden durch Multiplikation mit den zugehörigen Eigenvektoren, d. h., mit den Spaltenvektoren von $\boldsymbol{G}$, gebildet (nach [29]).

In der Regel *Ridge* wird nach dem Modell der linearen Diskriminanzanalyse nach Abschnitt 3.3 dort in Gleichung (3.19) die Stichprobenkovarianzmatrix (3.18) durch die folgendermaßen berechnete Matrix ersetzt:

$$\boldsymbol{S}_{\text{Ridge}} := \boldsymbol{S} + \frac{p(n-2)}{(n-4)(n+p-3)} \mathbf{Diag}(\boldsymbol{S}[\mathbf{Diag}(\boldsymbol{S})]^{-1}\boldsymbol{S}). \tag{9.5}$$

Mit s_{ij} $(i, j = 1, \ldots, p)$ seien die Elemente der Matrix $\boldsymbol{S}$ bezeichnet. Die Diagonalelemente s_{ii} $(i = 1, \ldots, p)$ von $\boldsymbol{S}$ sind gegeben durch

$$s_{ii} = \frac{1}{n-2} \sum_{j=1}^{2} \sum_{\alpha=1}^{n_j} (x_{i\alpha}^{(j)} - \bar{x}_i^{(j)})^2 \ (i = 1, \ldots, p); \tag{9.6}$$

diese sind mit Wahrscheinlichkeit 1 alle positiv. Damit ist mit Wahrscheinlichkeit 1 $\mathbf{Diag}(\boldsymbol{S})$ invertierbar. Die Diagonalelemente von $\boldsymbol{S}[\mathbf{Diag}(\boldsymbol{S})]^{-1}\boldsymbol{S}$ berechnen sich wie folgt:

$$[\boldsymbol{S}[\mathbf{Diag}(\boldsymbol{S})]^{-1}\boldsymbol{S}]_{ii} = \sum_{k=1}^{p} s_{kk}^{-1} s_{ik}^2 \ (i = 1, \ldots, p). \tag{9.7}$$

Mit Wahrscheinlichkeit 1 sind diese ebenfalls alle positiv; damit ist mit Wahrscheinlichkeit 1 $\mathbf{Diag}(\boldsymbol{S}[\mathbf{Diag}(\boldsymbol{S})]^{-1}\boldsymbol{S})$ positiv definit — auch bei singulärer Stichprobenkovarianzmatrix $\boldsymbol{S}$.

Die so genannte *Mehrfaktorregel* verwendet statt $\boldsymbol{S}$ in Gleichung (3.19) die folgende, in der Regel *Ridge* als Korrekturterm vorkommende Diagonalmatrix:

$$\boldsymbol{S}_{\text{Mehrfaktor}} = \mathbf{Diag}(\boldsymbol{S}[\mathbf{Diag}(\boldsymbol{S})]^{-1}\ \boldsymbol{S}). \tag{9.8}$$

Beim Klassifikationsverfahren *VCA* wird nach Abschnitt 4.2.2 zunächst eine Dimensionsreduzierung durchgeführt. Nach der Bedingung (4.14) werden die Variablenmengen M_i gebildet; hier wird stets $c = 0.5$ gesetzt. Bei Version 2 (VCA-II) werden hierbei im Unterschied zu Version 1 (VCA-I) die einelementigen Variablenmengen nicht mit berücksichtigt. Die neuen Datenvektoren werden dann nach (4.15) berechnet. Identische Variablenmengen M_i, die genau dieselben Variablen enthalten, werden nur einmal zur Bildung eines neuen Datenvektors verwendet. Um die Anzahl der Datenvektoren zu beschränken, wird, falls es mehr als $n/6$ (n ist Gesamtanzahl der Lernstichprobenelemente) verschiedene Variablenmengen gibt, die Anzahl der Datenvektoren auf $n/6$ (ganzzahlig gerundet) reduziert. Dazu werden die Datenvektoren $\boldsymbol{z}_i$ mit den betragsmäßig größten Werten der t-Statistik

$$|t_i| = \frac{|\bar{z}_i^{(1)} - \bar{z}_i^{(2)}|}{s_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{9.9}$$

ausgewählt, wobei hier $\bar{z}_i^{(1)}$, $\bar{z}_i^{(2)}$ und

$$s_i = \sqrt{\frac{1}{n-2}\sum_{j=1}^{2}\sum_{\alpha=1}^{n_j}(z_{\alpha,i}^{(j)} - \bar{z}^{(j)})^2} \tag{9.10}$$

die empirischen Mittelwerte bzw. die empirische Standardabweichung der Komponenten der neuen Datenvektoren sind. Gibt es gar keine Variablenmenge mit mindestens zwei Variablen, so wird bei Version 2 von den ursprünglichen Variablen diejenige mit der betragsmäßig größten t-Statistik als neuer Datenvektor gewählt. Sind die neu gebildeten Datenvektoren linear abhängig, so wird der Datenvektor mit der betragsmäßig kleinsten t-Statistik weggelassen. Das wird gegebenenfalls mehrfach wiederholt, bis die verbleibenden Datenvektoren linear unabhängig sind.

Mit den letztlich verbleibenden neuen Datenvektoren wird dann schließlich die klassische lineare Diskriminanzregel (3.19) angewandt.

Bei den Verfahren RDA-CV und RDA*-CV wird zunächst per Kreuzvalidierung innerhalb der Lernstichprobenelemente ein optimaler Parameter λ_{CV} bestimmt. Dazu wird aus den Lernstichprobenelementen nacheinander jeweils einer als „zu klassifizierend“ deklariert, die restlichen verbleiben als Lernstichprobenelemente. Die Verfahren RDA und RDA* werden dazu nacheinander mit verschiedenen λ-Werten angewandt. Ausgehend vom Startwert λ_0, der wie bei RDA und RDA* berechnet wird, wird λ dann schrittweise verkleinert bzw. vergrößert. Aus allen diesen Durchläufen werden jeweils für alle der verwendeten Parameter λ für beide Verfahren die Fehlerraten bestimmt. Schließlich werden zur eigentlichen Klassifikation die Diskriminanzfunktionen (9.1) (bei RDA-CV) bzw. (9.2) (bei RDA*-CV) mit den Parametern λ_{CV}, die hier jeweils die niedrigsten Fehlerraten hatten, verwendet.

9.2 Simulationen

Bei jedem Durchlauf wurden jeweils dieselben Daten für die verschiedenen Verfahren verwendet. Gemäß dem Modell der linearen Diskriminanzanalyse wurden jeweils Zufallsvektoren $\boldsymbol{X}_1^{(1)}, \ldots, \boldsymbol{X}_{n_1}^{(1)} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\boldsymbol{X}_1^{(2)}, \ldots, \boldsymbol{X}_{n_2}^{(2)} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ sowie ein weiterer Zufallsvektor $\boldsymbol{X}$ simuliert, wobei abwechselnd $\boldsymbol{X} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ bzw. $\boldsymbol{X} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ angenommen wurde.

Die Anzahl der Lernstichprobenelemente war jeweils mit $n_1 = n_2 = 10$ bestimmt. Jede Simulation wurde mit 10 000 Wiederholungen durchgeführt. Zum Vergleich wurden zusätzlich jedes Mal auch die theoretischen minimalen Fehlklassifikationswahrscheinlichkeiten, wie in Abschnitt 8.1 beschrieben, berechnet.

9.2.1 Klassifikation in Abhängigkeit von der Anzahl der Variablen

Die Frage der Abhängigkeit der Fehlklassifikationswahrscheinlichkeit von der Dimension des Beobachtungsraumes, d. h. von der Anzahl der Variablen p, ist sicher von besonderem Interesse. Nach approximativen Berechnungen und aus Simulationsuntersuchungen ist bekannt, dass bei der klassischen linearen Diskriminanzanalyse mit geschätzten Parametern bei Erhöhung der Dimension p die Fehlerrate größer wird, wenn der Mahalanobis-Abstand (8.5) konstant gehalten wird [33]. Wenn zu bestehenden Variablen einfach neue dazugenommen werden oder umgekehrt einige der vorhandenen Variablen weggelassen werden, bleibt der Mahalanobis-Abstand jedoch nicht notwendigerweise konstant. Wie sich in diesem Fall die Fehlerrate verändert, wurde hier untersucht.

Die Parameter wurden hierbei wie folgt bestimmt:

$$\boldsymbol{\mu}^{(1)} = (\,1,\ldots,1)', \;\; \boldsymbol{\mu}^{(2)} = (\,0,\ldots,0)',$$

$$\boldsymbol{\Sigma} = [\sigma_{ij}]_{i,j=1,\ldots,p} \text{ mit } \begin{array}{rcl} \sigma_{ii} & = & 1 \text{ für } i = 1,\ldots,p, \\ \sigma_{ij} & = & 0.5, \text{ falls } i \neq j \;\; (i,j = 1,\ldots,p). \end{array}$$

Die paarweisen Korrelationen zwischen je zwei verschiedenen Variablen wurden einheitlich auf $\varrho = 0.5$ gesetzt, und die Anzahl der Variablen p wurde variiert. Tabelle 9.1 zeigt die erhaltenen Fehlerraten.

	Anzahl der Variablen						
	5	10	15	20	50	100	200
theor. optimale Rate	0.259	0.250	0.247	0.245	0.242	0.241	0.240
PCA	0.264	0.261	0.264	0.267	0.278	0.272	0.264
PLS	0.303	0.295	0.297	0.303	0.282	0.276	0.263
Ridge	0.285	0.279	0.272	0.261	0.250	0.253	0.245
Mehrfaktor	0.267	0.263	0.258	0.254	0.245	0.250	0.245
VCA-I	0.296	0.286	0.283	0.277	0.268	0.260	0.256
VCA-II	0.305	0.292	0.288	0.278	0.268	0.260	0.256
RDA-det1	0.328	0.366	0.411	0.377			
RDA-det2	0.322	0.356	0.346	0.307			
RDA-λ_0	0.269	0.262	0.258	0.252	0.245	0.251	0.244
RDA*-λ_0	0.275	0.266	0.263	0.254	0.245	0.251	0.244

Tabelle 9.1: Durch Simulation ermittelte Fehlerraten in Abhängigkeit von der Variablenanzahl

9.2.2 Klassifikation bei zwei unabhängigen Variablenblöcken

Um auch eine etwas andere Verteilungsstruktur zu verwenden, wurde für weitere Simulationen ein „Zwei-Block-Modell“ mit zwei voneinander unabhängigen Variablenblöcken zugrundegelegt. Die Parameter wurden dazu wie folgt bestimmt:

$$\boldsymbol{\mu}^{(1)} = (\,1,\ldots,1)',\ \boldsymbol{\mu}^{(2)} = (\,0,\ldots,0)',$$

die Kovarianzmatrix hatte die Form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

wobei die Teilmatrizen $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$, jeweils $p/2 \times p/2$, wie folgt bestimmt wurden:

$$\boldsymbol{\Sigma}_{kk} = [\sigma_{ij}]_{i,j=1,\ldots,p/2} \text{ mit } \begin{array}{rcl} \sigma_{ii} & = & 1 \text{ für } i = 1,\ldots,p/2, \\ \sigma_{ij} & = & \varrho > 0, \text{ falls } i \neq j \ (i,j = 1,\ldots,p/2) \end{array} \ (k = 1,2).$$

Hier war die Anzahl der Variablen mit $p = 50$ fest, und die paarweisen Korrelationen innerhalb der zwei voneinander unabhängigen Variablenblöcke wurden in den verschiedenen Simulationsläufen mit $\varrho = 0.1$, $\varrho = 0.5$ sowie $\varrho = 0.9$ variiert.

Tabelle 9.2 zeigt die Klassifikationsergebnisse dieser Simulationen.

	paarweise Korrelationen		
	0.1	0.5	0.9
theoretische optimale Rate	0.028	0.163	0.229
PCA	0.047	0.207	0.257
PLS	0.050	0.210	0.292
Ridge	0.047	0.191	0.254
Mehrfaktor	0.040	0.182	0.254
VCA-I	0.205	0.255	0.258
VCA-II	0.196	0.256	0.258
RDA-λ_0	0.035	0.176	0.250
RDA*-λ_0	0.035	0.175	0.252

Tabelle 9.2: Durch Simulation ermittelte Fehlerraten im Zwei-Block-Modell

9.3 Anwendungsbeispiele

Die Verfahren wurden schließlich auch an realen Datensätzen angewandt. Die Daten stellen jeweils eine Reihe von mehrdimensionalen Beobachtungen dar, die

jeweils einer von zwei Klassen angehören. In jedem Fall war die Gesamtanzahl der Lernstichprobenelemente beider Klassen nicht größer als die Anzahl der Merkmale, d. h., es galt hier immer

$$n = n_1 + n_2 \leq p. \tag{9.11}$$

Da hier keine separaten Teststichproben zur Verfügung standen, wurden nach dem Prinzip der Kreuzvalidierung die Klassifikationen in mehreren Durchläufen ausgeführt. Dabei wurde jedes Mal eine andere Beobachtung als „zu klassifizierend“ und die verbleibenden als Lernstichprobenelemente deklariert („leave one out“). Aus allen Durchläufen wurden dann die Fehlerraten berechnet. Bei den Verfahren RDA-CV und RDA*-CV wurde etwa die Hälfte der Datenvektoren zur Bestimmung des optimalen Parameters λ_{CV} verwendet. Dies geschah auch jeweils per leave-one-out-Methode. Mit den optimalen Parametern wurden dann an den restlichen Daten mit der leave-one-out-Methode wieder die Fehlerraten ermittelt.

9.3.1 Verwendete Datensätze

Für die Beispielberechnungen wurden drei verschiedene Datensätze verwendet. Zwei davon enthalten aus Microarray-Experimenten gewonnene Genexpressionsdaten, der dritte enthält die Spektraldaten verschiedener Stoffproben.

Der Datensatz „Colon“ wurde erstmalig von U. Alon et al. [1] verwendet und ist jetzt über das Internet frei verfügbar (http://www.sph.uth.tmc.edu/hgc, vgl. [52]). Der Datensatz enthält die Expressionsdaten von 2000 Genabschnitten ($p = 2000$) aus dem Darmgewebe von insgesamt 62 Personen (22 gesunde, 40 tumorerkrankte). Nach Alon et al. haben bereits zahlreiche andere Autoren diese Daten für verschiedene Untersuchungen, u. a. über Klassifikationsverfahren, verwendet. Um solche Genexpressionsdaten besser analysieren zu können, werden sie üblicherweise zunächst vorbehandelt. Hier wurden sie, entsprechend dem Vorgehen in [9], durch Logarithmieren mit dekadischem Logarithmus transformiert und anschließend über beide Klassen hinweg standardisiert, so dass schließlich alle Variablen einheitlich den empirischen Gesamtmittelwert 0 und die empirische Varianz 1 haben.

Der Datensatz „Knoten“ stammt aus Microarray-Experimenten, die an der Medizinischen Fakultät der Universität Leipzig durchgeführt wurden [14] [15]. Der Datensatz beinhaltet die Genexpressionsdaten aus dem Schilddrüsengewebe von 30 Personen. Davon haben je 15 Personen kalte bzw. heiße Knoten. Für die Analyse wurden jeweils die Differenzen zwischen den logarithmierten Werten von Knotengewebe und Umgebungsgewebe verwendet. Um den Rechenaufwand zu beschränken, wurden hier für die Berechnungen von den ursprünglich 12625 Genen nach dem Kriterium der Beträge der t-Statistiken,

$$|t_i| = \frac{|\bar{x_i}^{(1)} - \bar{x_i}^{(2)}|}{s_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \; (i = 1, \ldots, p), \tag{9.12}$$

3000 ausgewählt.

Der Datensatz „Gasoil“ enthält die Spektraldaten von insgesamt 115 Stoffproben, die zu drei Untergruppen gehören. Er ist auch über das Internet erhältlich (http://myweb.dal.ca/pdwentze/). Zu jeder Stoffprobe sind in den Daten jeweils die UV-Spektren über 572 Kanäle aufgezeichnet (d. h. $p = 572$). Für die Berechnungen zur Klassifizierung wurden zwei dieser Untergruppen verwendet, so dass wieder zwei Klassen vorhanden waren. Es verblieben für die Analyse 56 Proben — 30 zu der einen Klasse und 26 zur zweiten.

9.3.2 Klassifikationsergebnisse

In Tabelle 9.3 sind die durch die verschiedenen Verfahren erreichten Fehlerraten angegeben.

Datensatz	Colon	Knoten	Gasoil
PCA	0.15	0.07	0.04
PLS	0.15	0.07	0.04
Ridge	0.13	0.10	0.14
Mehrfaktor	0.27	0.13	0.16
VCA-I	0.11	0.17	0.12
VCA-II	0.11	0.17	0.12
RDA-λ_0	0.27	0.10	0.21
RDA-CV	0.26	0.12	0.00
RDA*-λ_0	0.24	0.10	0.16
RDA*-CV	0.32	0.12	0.00

Tabelle 9.3: An Beispieldaten ermittelte Fehlerraten

Die Ergebnisse lassen darauf schließen, dass der Schwierigkeitsgrad der Klassifikation in den einzelnen Datensätzen unterschiedlich ist. Für jeden Datensatz separat lassen sich aber die einzelnen Klassifikationsverfahren bezüglich der erreichten Fehlerraten vergleichen.

9.4 Auswertung der Ergebnisse

Welches Klassifikationsverfahren die kleinsten Fehlerraten erzielt, hängt sicher immer stark von den jeweiligen Bedingungen ab. Dies zeigt sich besonders bei der Verwendung von realen Datensätzen, die in ihrer Struktur doch recht unterschiedlich sein können.

Die beiden Verfahren RDA-det1 und RDA-det2 wurden wegen des hohen Rechenaufwandes nur bei den Simulationen bis zu einer Merkmalsanzahl von $p = 20$

angewandt. Der verwendete Parameter λ wurde hierbei ja nach dem Kriterium einer (näherungsweisen) Determinanten-Erwartungstreue bestimmt. Im Kapitel 7 wurde bereits angemerkt, dass dieses nicht das für die Klassifikation entscheidende Kriterium ist. Daher sind auf diese Weise auch nicht unbedingt gute Klassifikationsergebnisse zu erwarten. Dies hat sich in den Simulationen auch bestätigt. Die erreichten Fehlerraten liegen jeweils deutlich über denen der anderen Verfahren. Es wurden insgesamt keine befriedigenden Ergebnisse erreicht, so dass dieser Ansatz wohl für die Klassifikation nicht als geeignet erscheint.

Bis auf RDA-det1 und RDA-det2 verhalten sich die anderen verwendeten Verfahren bei den Simulationen sehr ähnlich zueinander. Bei Veränderung der Korrelationen zwischen den Variablen oder der Variablenanzahl verändern sich jeweils die Fehlerraten in die gleiche Richtung. Sie liegen nur auf unterschiedlichem Niveau, und die Änderung ist auch unterschiedlich stark. Eine Erhöhung der Korrelationen zwischen den Variablen führt bei allen verwendeten Verfahren zu einer Erhöhung der Fehlerraten. Dieses Verhalten entspricht auch etwa dem der berechneten theoretischen optimalen Fehlerraten.

Bei den Simulationen in Abhängigkeit von der Merkmalsanzahl p lassen sich zwei verschiedene Muster erkennen: Bei einigen Verfahren nehmen die Fehlerraten mit Erhöhung der Merkmalszahlen nahezu kontinuierlich ab. Bei anderen werden die Fehlerraten mit Erhöhung der Merkmalszahlen zunächst größer, erreichen ein Maximum und nehmen dann mit weiterer Erhöhung der Merkmalszahlen wieder ab.

Vergleicht man hier die Verfahren untereinander, so zeigt sich, dass bei den Simulationen die Mehrfaktorregel und RDA-λ_0 in allen Fällen die besten Ergebnisse erzielt haben. Sie unterscheiden sich auch nur unwesentlich voneinander. Bei hohen Korrelationen und bei hohen Merkmalszahlen erreichen die Verfahren — ohne RDA-det1 und RDA-det2 — nahezu die gleichen Fehlerraten. Bei kleineren Korrelationen hat das Verfahren VCA die höchsten Fehlerraten, ansonsten hat zumeist das Verfahren PLS die höchsten Fehlerraten.

Mit den verwendeten Datensätzen wurden jeweils unterschiedliche Ergebnisse erreicht. Auffällig ist, dass hier die Verfahren PCA und PLS jeweils genau die gleichen Fehlerraten haben.

Das Verfahren Ridge liegt mit den erreichten Ergebnissen jeweils im Mittelfeld. Im Gegensatz dazu sind die Ergebnisse des Verfahrens VCA in den einzelnen Datensätzen extrem unterschiedlich. Beim Datensatz Colon hat es die niedrigsten Fehlerraten, beim Datensatz Knoten dagegen die höchsten.

Die Verfahren Mehrfaktor, RDA-λ_0 und RDA*-λ_0 haben hier überall ähnliche Ergebnisse erreicht. Beim Datensatz Knoten wurden mittlere Fehlerraten erreicht, bei den anderen Datensätzen waren diese Verfahren am schlechtesten.

Zumeist wurden mit den beiden Versionen von VCA jeweils ähnliche oder sogar die gleichen Ergebnisse erreicht. Bei den Simulationen war bei niedrigen Korrelationen die Version II, bei der zur Bildung der Faktoren die einelementigen Variablenmengen nicht mit verwendet wurden, etwas besser, bei niedrigen

Dimensionen hingegen war sie etwas schlechter.

Der Unterschied der Verfahren RDA-CV und RDA*-CV, bei denen im Vergleich zu RDA-λ_0 und RDA*-λ_0 der optimale Parameters λ_{CV} durch Kreuzvalidierung ermittelt wurde, war beim Datensatz Gasoil besonders deutlich. Bei den anderen Datensätzen wurde dadurch keine wesentliche Verbesserung erreicht, manchmal sogar eine leichte Verschlechterung. Die Verfahren RDA-λ_0 und RDA*-λ_0 unterscheiden sich bei der Klassifizierung mit den Datensätzen insgesamt nur unwesentlich.

Als Fazit lässt sich für alle der verwendeten Verfahren — mit Ausnahme von RDA-det1 und RDA-det2 — feststellen, dass sie die Situation der hohen Dimension recht gut bewältigen. Im Gegensatz zur klassischen Methode LDA, bei der die Fehlerraten groß werden, wenn der Gesamtumfang n der Lernstichproben nicht viel größer als die Merkmalsanzahl p ist (vgl. [33]), liefern die Verfahren hier auch bei hohen Merkmalszahlen brauchbare Ergebnisse. Das zeigt sich besonders bei den Simulationen in Abhängigkeit von der Merkmalsanzahl p. Die Erhöhung von p führte insgesamt nicht zu einer wesentlichen Verschlechterung der Ergebnisse, d. h. zu wesentlich höheren Fehlerraten.

Kapitel 10

Zusammenfassung

Die Situation hoher Dimensionen und daraus entstehende Probleme der statistischen Analyse bildeten den inhaltlichen Schwerpunkt dieser Arbeit. Insbesondere wurde die Situation einer singulären Stichprobenkovarianzmatrix betrachtet.

Die Spektralzerlegung $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$ zeigte sich als eine gute Möglichkeit, den Parameter $\boldsymbol{\Sigma}$ als Linearkombination aus Eigenwerten und Eigenvektoren darzustellen. So konnte die Frage nach der Existenz von Maximum-Likelihood-Schätzern für die Teilparameter $\boldsymbol{\Gamma}$ und $\boldsymbol{\Lambda}$ erörtert werden. Es konnte gezeigt werden, dass in jedem Fall — nicht notwendig eindeutige — Maximum-Likelihood-Schätzer für die Eigenvektoren von $\boldsymbol{\Sigma}$ existieren, bei singulärer Stichprobenkovarianzmatrix $\boldsymbol{S}$ aber nicht für die Eigenwerte. Dies sind zunächst Aussagen von theoretischer Bedeutung. Da in jedem Fall die Eigenvektoren von $\boldsymbol{S}$ die Likelihoodfunktion maximieren, erscheint es darüber hinaus als naheliegend, sie als Schätzer für die Eigenvektoren von $\boldsymbol{\Sigma}$ zu verwenden. Da auch solche Ridge-Schätzungen der Form

$$\boldsymbol{S}_{\text{ridge}} = \boldsymbol{S} + \lambda \boldsymbol{I} \tag{10.1}$$

dieselben Eigenvektoren haben wie $\boldsymbol{S}$, ist hiermit zugleich eine Begründung für die Verwendung dieses heuristischen Ansatzes gegeben.

Weiterhin konnte ein Stabilitätskonzept entwickelt werden, das eine Charakterisierung der besonderen Situation bei singulärer Stichprobenkovarianzmatrix ermöglicht. Hier konnte für den Schätzfehler eine untere Schranke in Abhängigkeit vom wahren Parameter $\boldsymbol{\Sigma}$ angegeben werden. Ist $\boldsymbol{S}$ singulär, so gilt mit Wahrscheinlichkeit 1

$$\|\boldsymbol{S} - \boldsymbol{\Sigma}\| \geq \lambda_p, \tag{10.2}$$

wobei λ_p der kleinste Eigenwert von $\boldsymbol{\Sigma}$ ist. Bei positiv definiter Matrix $\boldsymbol{\Sigma}$ ist $\lambda_p > 0$, somit hat man damit eine untere Schranke für den Schätzfehler, wenn $\boldsymbol{S}$ (oder eine andere singuläre Matrix) als Schätzung für $\boldsymbol{\Sigma}$ verwendet wird. Der Kern der Aussage ist hierbei, dass bei singulärer Schätzungsmatrix mit Wahrscheinlichkeit 1 ein positiver Abstand zum wahren Parameter besteht. Es können auch nicht

ohne Weiteres auf sonst übliche Weise Konfidenzbereiche für die Kovarianzmatrix angegeben werden.

Sollen als Alternative Ridge-Schätzungen der Form

$$\boldsymbol{S}_{\text{ridge}} = \boldsymbol{S} + \lambda \boldsymbol{I} \tag{10.3}$$

verwendet werden, hat man den zusätzlichen Parameter λ geeignet zu wählen. Dies gilt ebenso bei Verwendung der Ridge-Klassifikationsmethode RDA, bei der in der Diskriminanzfunktion der klassischen linearen Diskriminationsmethode statt der Stichprobenkovarianzmatrix $\boldsymbol{S}$ eine Ridge-Matrix der Form (10.3) verwendet wird. Die Untersuchungen zur Verwendung der näherungsweisen Determinanten-Erwartungstreue als Kriterium zur Bestimmung von λ können hier wohl als interessante zusätzliche Betrachtungen angesehen werden; für die Klassifikation scheint dieses Kriterium eher nicht geeignet. Die in Simulationen ermittelten λ-Werte sind hier sehr viel kleiner als bei Verwendung des Klassifikationsfehlers als Kriterium, und entsprechend hoch waren dann auch die ermittelten Klassifikationsfehler.

Die Klassifikationsfehleranalyse lieferte zunächst eine asymptotische Berechnungsformel für den Fehler der Klassifikationsregel RDA*, die wiederum eine Approximation der Regel RDA darstellt. Eine Minimierung dieses Fehlers liefert schließlich den gesuchten optimalen Wert des Ridge-Parameters λ. Diese Möglichkeit der Parameterbestimmung für λ ließ sich dann in zwei Klassifikationsregeln, RDA-λ_0 und RDA*-λ_0, anwenden. Die wahren Verteilungsparameter wurden dabei jeweils durch ihre Schätzungen ersetzt. Dieses heuristische „Plug-in"-Vorgehen kann so zumindest auf Grundlage theoretischer Resultate erfolgen.

In der Theorie werden zur Ableitung von Verfahren ja oftmals ideale Voraussetzungen angenommen, die in Wirklichkeit nicht volständig erfüllt sind. Für einige der verwendeten Verfahren trifft das für die zugrundegelegten Verteilungsannahmen zu (multivariate Normalverteilung mit gemeinsamer Kovarianzmatrix). Speziell für die Verfahren RDA-λ_0 und RDA*-λ_0 ist weiter zu bedenken, dass zur Ermitlung von λ_0 die Reihenentwicklung nur für hinreichend großes λ gültig ist, dass dabei nur die ersten beiden Glieder der Reihe verwendet wurden und dass hier in der Berechnungsformel für λ_0 außerdem die wahren Parameter einfach durch ihre Schätzungen ersetzt wurden. Daher sind nicht unbedingt in jedem Fall optimale Ergebnisse bei der Klassifikation zu erwarten. Da auch die anderen verwendeten Klassifikationsmethoden überwiegend heuristischer Natur sind, bestehen hier jedoch ähnliche Voraussetzungen, die einen fairen Vergleich ermöglichen.

Raudys und Skurichina [43] sehen die Bedeutung ihrer Berechnungsformel für einen optimalen Parameter λ_{opt} vor allem darin, dass sie eine Orientierung für einen guten Startwert liefert und so hilft, den Rechenaufwand für die Ermittlung eines optimalen Parameters per Kreuzvalidierung zu reduzieren. Die hier angegebene Berechnungsformel liefert dazu eine Erweiterung auf den Fall $n - 2 < p$. Sie

kann nun ebenso für die Bestimmung eines guten Startwertes für die anschließende Kreuzvalidierung verwendet werden. Damit können, wie bei dem Datensatz Gasoil, im Vergleich zur direkten Anwendung des Verfahrens deutlich bessere Ergebnisse erzielt werden. Auf der anderen Seite bedeutet dies aber — im Vergleich zu anderen Verfahren — immer noch einen relativ hohen Rechenaufwand, besonders bei hohen Merkmalszahlen. Wirklich brauchbar wäre das Verfahren, wenn der nach der Formel berechnete Parameter λ_0 direkt für die Klassifikation angewendet werden könnte. Die Simulationsrechnungen und besonders das Ergebnis des Datensatzes Knoten zeigen, dass dieses Vorgehen unter geeigneten Voraussetzungen im Vergleich zu anderen Verfahren zu relativ guten Ergebnissen führen kann. Es bleibt nun das Problem, festzustellen, ob solche „geeignete Voraussetzungen“ vorliegen oder nicht.

Bei den Simulationsrechnungen wurden unter den speziellen verwendeten Modellbedingungen mit den Verfahren RDA-λ_0 und RDA*-λ_0, verglichen mit den anderen Verfahren, relativ gute Ergebnisse erzielt. Zumindest für den Fall, dass von annähernd solchen wie den verwendeten Bedingungen, d. h. von annähernd gleichen Varianzen in den einzelnen Variablen und nicht sehr großen paarweisen Korrelationen, ausgegangen werden kann, spricht nichts gegen eine direkte Anwendung dieser Verfahren. Die Anwendbarkeit unter anderen Verteilungsannahmen zu untersuchen, bleibt wohl vorerst noch eine Aufgabe der zukünftigen Forschung.

Literaturverzeichnis

[1] Alon, U., Barkai, N., Notterman, D., Gish, K., Mack, S. and Levine, J. (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *PNAS*, **96**, 6745 - 6750.

[2] Anderson, T. W. (1963). Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics*, **34**, 122 – 148.

[3] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley and Sons, New York.

[4] Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*, John Wiley and Sons, New York.

[5] Bronstein, I. N., Semendjajew, K. A., Musiol, G. und Mühlig, H. (1993). *Taschenbuch der Mathematik*, Verlag Harri Deutsch, Thun, Frankfurt am Main.

[6] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

[7] Cramér, H. (1957). *Mathematical methods of statistics*, Princeton University, Princeton.

[8] Das Gupta, S. (1982). Optimum Rules for Classification into Two Multivariate Normal Populations with the Same Covariance Matrix. In P. R. Krishnaiah and L. N. Kanal, *Handbook of Statistics 2*, North-Holland, Amsterdam, 47 – 60.

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

[10] Deuflhard, P. und Hohmann, A. (2002). *Numerische Mathematik I*, 3. Auflage, Walter de Gruyter, Berlin.

[11] Deuflhard, P. und Bornemann, F. (2002). *Numerische Mathematik II*, 2. Auflage, Walter de Gruyter, Berlin.

[12] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In J. Kittler and F. Roli (Ed.), *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, pp. 1–15, Springer Verlag, New York.

[13] Dieudonné, J. (1971). *Grundzüge der modernen Analysis*, Deutscher Verlag der Wissenschaften, Berlin.

[14] Eszlinger, M., Krohn, K., Frenzel, R., Kropf, S., Tonjes, A., Paschke, R. (2004). Gene expression analysis reveals evidence for inactivation of the TGF-beta signaling cacade in autonomously functioning thyroid nodules, *Oncogene*, **23**(3), 795 – 804.

[15] Eszlinger, M., Krohn, K., Berger, K., Läuter, J., Kropf, S., Beck, M., Führer, D., Paschke, R. (2005). Gene expression analysis reveals evidence for increased expression of cell cycle associated genes and Gq-Protein-Protein Kinase C signaling in cold thyroid nodules, *J Clin Endocrinol Metab.*, **90**(2), 1163 – 1170.

[16] Fang, K-T. and Zang, Y-T. (1990). *Generalized Multivariate Analysis*, Science Press, Beijing, and Springer-Verlag, Berlin.

[17] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179 – 188.

[18] Frank, I. E. and Friedman, J. H. (1993). A statistical View of Some Chemometrics Regression Tools, *Technometrics*, **35**, 109 – 149.

[19] Fuentes Rodriguez, A. (1988). Admissibility and Unbiasedness of the Ridge Classification Rules for Two Normal Populations with Equal Covariance Matrices, *Statistics*, **19**(3), 383 – 388.

[20] Hämmerlin, G. und Hoffmann, K.-H. (1994). *Numerische Mathematik*, 4. Auflage, Springer-Verlag, Berlin Heidelberg.

[21] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

[22] Healy, M. J. R. (1986). *Matrices for Statistics*, Clarendon Press, Oxford.

[23] Helland, I. S. (1990). Partial Least Squares Regression and Statistical Models, *Scand. J. Statist.*, **17**, 97 – 114.

[24] Henderson, H. V. and Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics, *Can. J. Stat.*, **7**, 65 – 81.

[25] Hothorn, T. (2003). *Bundling Classifiers with an Application to Glaucoma Diagnosis*, Dissertation, Universität Dortmund, Fachbereich Statistik.

[26] Jenkins, M. A. and Traub, J. F. (1970). A Three-stage Algorithm for Real Polynomials using Quadratic Iteration, *SIAM Journal of Numerical Analysis*, **7**, 545 – 566.

[27] Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley and Sons, New York.

[28] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, 2 Vols., 2nd ed., John Wiley and Sons, New York.

[29] Kropf, S. (2000). *Hochdimensionale multivariate Verfahren in der medizinischen Statistik*, Shaker Verlag, Aachen.

[30] Krzanowski, W. J., Jonathan, P., McCarty, W. V. and Thomas, M. R. (1995). Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied statistics*, **44**, 101 – 115.

[31] Krzanowski, W. J. and Marriott, F. H. C. (1994). *Multivariate Analysis*, Part 1 Distributions, Ordination and Inference, Arnold, London.

[32] Krzanowski, W. J. and Marriott, F. H. C. (1995). *Multivariate Analysis*, Part 2 Classification, Covariance Structures and Repeated Measurements, Arnold, London.

[33] Läuter, J. (1992). *Stabile multivariate Verfahren. Diskriminanzanalyse, Regressionsanalyse, Faktoranalyse*, Akademie Verlag, Berlin.

[34] Läuter, J., Glimm, E. and Eszlinger, M. (2005). Search for Relevant Sets of Variables in a High-Dimensional Setup Keeping the Familywise Error Rate, submitted to *Statistica Neerlandica*.

[35] Lehmann, E. L. (1983). *Theory of Point Estimation*, John Wiley and Sons, New York.

[36] Lehmann, E. L. (1986). *Testing statistical hypotheses*, 2nd ed., John Wiley and Sons, New York.

[37] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.

[38] Markowetz, F., Edler, L. and Vingron, M. (2003). Support Vector Machines for Protein Fold Class Prediction, *Biometrical Journal*, **45**(3), 377 – 389.

[39] Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (eds.) (1994). *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York.

[40] Pruscha, H. (2000). *Vorlesungen über Mathematische Statistik*, Teubner, Stuttgart.

[41] Pugh, C. C. (2002). *Real Mathematical Analysis*, Springer Verlag, New York.

[42] Raudys, Š. (2001). *Statistical and Neural Classifiers: An Integrated Approach to Design*, Springer Verlag, London.

[43] Raudys, Š., Skurichina, M. (1994). Small sample properties of ridge estimate of the covariance matrix in statistical and neural net classification. In E. M. Tiit, T. Kollo, H. Niemi (editors), *New Trends in Probability and Statistics: Multivariate statistics and matrices in statistics*, **3**, 237 – 245, TEV, Vilnius and VSP, Utrecht.

[44] SAS Institute Inc. (2003). *JMP Statistics and Graphics Guide, Version 5.1*, SAS Institute Inc., Cary, NC.

[45] SAS Institute Inc. *Examples Using the PLS Procedure*, SAS Institute Inc., Cary, NC. (http://support.sas.com/rnd/app/papers/plsex.pdf)

[46] Schwarz, H. R. (1969). *Numerik symmetrischer Matrizen*, Teubner, Leipzig.

[47] Searle, S. R. (1982). *Matrix algebra useful for statistics*, John Wiley and Sons, New York.

[48] Smirnow, W. I. (1971). *Lehrgang der höheren Mathematik*, Teil III/1, 6. Auflage, VEB Deutscher Verlag der Wissenschaften, Berlin.

[49] Speed, T. (ed.) (2003). *Statistical analysis of gene expression microarray data*, Chapman and Hall, London.

[50] Stone, M. and Brooks, R. J. (1990). Continuum Regression: Cross-validated Sequantially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression, *J. Stast. Soc. B*, **52**(2), 237 – 269.

[51] Wold, H. O. (1985). Partial least squares. In *Encyclopedia of Statistical Sciences, Volume 6* (eds S. Kotz and N. L. Johnson), pp. 581 – 591, Wiley, New York.

[52] Xiong, M., Fang, X. and Zhao, J. (2001). Biomarker Identification by Feature Wrappers, *Genome Research*, **11**(11), 1878 – 1887.

Anhang A

Ermittelte Fehlerraten für die Verfahren RDA und RDA*

Es folgen jetzt die Simulationsergebnisse für die beiden Klassifikationsverfahren RDA und RDA* mit verschiedenen Parameterwerten λ. Die Einzelheiten zur Simulation sind in Abschnitt 8.4 beschrieben. In den folgenden Tabellen sind jeweils die ermittelten Fehlerraten dargestellt. Die unter den jeweiligen Ausgangsbedingungen kleinsten Fehlerraten sind jeweils hervorgehoben.

$p = 10, n = 20, v_2 = 1$

ϱ	0.1		0.9	
λ_0	0.8		7.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.2359	0.4371	0.4002	0.6959
$10^{-1.5}$	0.2271	0.4297	0.3767	0.6958
10^{-1}	0.2096	0.4074	0.3444	0.6955
$10^{-0.5}$	0.1798	0.3508	0.3222	0.6952
10^{0}	0.1588	0.231	0.3079	0.695
$10^{0.5}$	0.1433	0.149	0.3049	0.6733
10^{1}	0.1372	0.1382	0.3042	0.3721
$10^{1.5}$	**0.1357**	**0.1357**	**0.3039**	**0.3038**
10^{2}	0.1361	0.1362	0.3044	0.3044
$10^{2.5}$	0.1359	0.1359	0.3045	0.3045
10^{3}	0.1359	0.1359	0.3045	0.3045
$10^{3.5}$	0.1358	0.1359	0.3045	0.3045
10^{4}	0.1358	0.1358	0.3045	0.3045
∞	0.1358	0.1358	0.3045	0.3045
λ_0	0.1628	0.2535	0.3041	0.3041

$p = 10, n = 20, v_2 = 2$

ϱ	0.1		0.9	
λ_0	2.2		11.8	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.2721	0.6078	0.3069	0.6537
$10^{-1.5}$	0.2673	0.6038	0.2895	0.6536
10^{-1}	0.2506	0.5922	0.2751	0.6536
$10^{-0.5}$	0.229	0.5535	**0.2702**	0.6535
10^{0}	0.2111	0.4268	0.2901	0.6523
$10^{0.5}$	0.1955	0.2364	0.3133	0.6425
10^{1}	0.1889	0.19	0.3261	0.4953
$10^{1.5}$	0.1888	0.1885	0.3295	**0.3275**
10^{2}	0.1881	0.1882	0.3327	0.3323
$10^{2.5}$	**0.1878**	**0.1878**	0.3323	0.3323
10^{3}	0.188	0.188	0.3324	0.3324
$10^{3.5}$	0.188	0.188	0.3324	0.3324
10^{4}	0.188	0.188	0.3325	0.3325
∞	0.1881	0.1881	0.3325	0.3325
λ_0	0.2006	0.2889	0.3267	0.442

$p = 10, n = 20, v_2 = 3$

ϱ	0.1		0.9	
λ_0	3.6		16.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.2938	0.6398	0.2335	0.6221
$10^{-1.5}$	0.2884	0.6371	0.2195	0.622
10^{-1}	0.2767	0.6305	**0.2083**	0.6216
$10^{-0.5}$	0.259	0.6096	0.2132	0.6206
10^{0}	0.2416	0.5421	0.2486	0.6191
$10^{0.5}$	0.2338	0.3368	0.2955	0.6109
10^{1}	**0.2329**	**0.2351**	0.3319	0.5315
$10^{1.5}$	0.2356	0.2354	0.3462	**0.3347**
10^{2}	0.2373	0.2372	0.3522	0.3518
$10^{2.5}$	0.2373	0.2372	0.3547	0.3547
10^{3}	0.2365	0.2365	0.3551	0.3551
$10^{3.5}$	0.2364	0.2364	0.3553	0.3553
10^{4}	0.2365	0.2365	0.3555	0.3555
∞	0.2366	0.2366	0.3556	0.3556
λ_0	0.2337	0.3124	0.3392	0.4035

$p = 10, n = 40, v_2 = 1$

ϱ	0.1		0.9	
λ_0	0.8		7.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1811	0.4277	0.36	0.6933
$10^{-1.5}$	0.1781	0.4176	0.3466	0.6933
10^{-1}	0.1731	0.3875	0.3271	0.6932
$10^{-0.5}$	0.1593	0.304	0.3112	0.6932
10^{0}	0.1428	0.183	0.3063	0.6931
$10^{0.5}$	0.1341	0.1359	0.3067	0.685
10^{1}	**0.131**	**0.131**	0.3069	0.3435
$10^{1.5}$	0.1311	0.1311	0.3063	0.3065
10^{2}	0.1313	0.1313	0.3062	0.3062
$10^{2.5}$	0.1314	0.1314	0.3063	0.3063
10^{3}	0.1314	0.1314	0.3062	0.3062
$10^{3.5}$	0.1313	0.1313	**0.3061**	**0.3061**
10^{4}	0.1313	0.1313	**0.3061**	**0.3061**
∞	0.1313	0.1313	**0.3061**	**0.3061**
λ_0	0.1459	0.2016	0.3067	0.4777

$p = 10, n = 40, v_2 = 2$

ϱ	0.1		0.9	
λ_0	2.2		11.8	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.2166	0.6702	0.2515	0.6641
$10^{-1.5}$	0.2139	0.6669	0.2437	0.6641
10^{-1}	0.2107	0.6532	**0.2393**	0.6639
$10^{-0.5}$	0.202	0.5985	0.2534	0.6638
10^{0}	0.1888	0.4233	0.2822	0.6634
$10^{0.5}$	0.1814	0.1996	0.3064	0.6601
10^{1}	**0.1801**	0.1805	0.3207	0.5079
$10^{1.5}$	0.1804	**0.1801**	0.3254	**0.3233**
10^{2}	0.1821	0.182	0.3278	0.3279
$10^{2.5}$	0.1826	0.1826	0.3283	0.3283
10^{3}	0.1829	0.1829	0.3284	0.3284
$10^{3.5}$	0.1831	0.1831	0.3285	0.3285
10^{4}	0.1831	0.1831	0.3286	0.3286
∞	0.1831	0.1831	0.3286	0.3286
λ_0	0.1815	0.2408	0.3218	0.427

$p = 10, n = 40, v_2 = 3$

ϱ	0.1		0.9	
λ_0	3.6		16.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.2275	0.6772	0.181	0.6338
$10^{-1.5}$	0.227	0.6757	0.1782	0.6338
10^{-1}	0.2239	0.6724	**0.1736**	0.6337
$10^{-0.5}$	0.2162	0.6526	0.1885	0.6336
10^{0}	**0.2073**	0.5719	0.2379	0.6331
$10^{0.5}$	0.2082	0.2911	0.2929	0.6286
10^{1}	0.2139	**0.2132**	0.3273	0.5734
$10^{1.5}$	0.219	0.2184	0.3419	**0.3321**
10^{2}	0.2203	0.2204	0.3468	0.3467
$10^{2.5}$	0.2219	0.2219	0.3486	0.3485
10^{3}	0.222	0.222	0.3493	0.3493
$10^{3.5}$	0.2221	0.2221	0.3493	0.3493
10^{4}	0.2221	0.2221	0.3493	0.3493
∞	0.2221	0.2221	0.3493	0.3493
λ_0	0.2086	0.2631	0.3362	0.3863

$p = 10, n = 60, v_2 = 1$

ϱ	0.1		0.9	
λ_0	0.8		7.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1595	0.4158	0.3467	0.7007
$10^{-1.5}$	0.1583	0.4021	0.3365	0.7007
10^{-1}	0.1553	0.3645	0.3196	0.7005
$10^{-0.5}$	0.1468	0.271	0.3069	0.7005
10^{0}	0.1379	0.1579	0.3027	0.7004
$10^{0.5}$	0.1314	0.1348	0.3013	0.6988
10^{1}	0.1303	0.1305	0.3003	0.3221
$10^{1.5}$	0.1291	0.1291	0.2998	0.2999
10^{2}	**0.1289**	**0.1289**	0.2996	0.2996
$10^{2.5}$	0.129	0.129	0.2996	0.2996
10^{3}	0.129	0.129	**0.2994**	**0.2994**
$10^{3.5}$	0.1291	0.1291	**0.2994**	**0.2994**
10^{4}	0.1291	0.1291	**0.2994**	**0.2994**
∞	0.1291	0.1291	**0.2994**	**0.2994**
λ_0	0.1393	0.1747	0.3003	0.4867

$p = 10, n = 60, v_2 = 2$

ϱ	0.1		0.9	
λ_0	2.2		11.8	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1943	0.7015	0.2344	0.6617
$10^{-1.5}$	0.1938	0.6971	0.2323	0.6617
10^{-1}	0.1916	0.6817	**0.2283**	0.6617
$10^{-0.5}$	0.1845	0.626	0.2487	0.6616
10^{0}	0.1787	0.4136	0.2886	0.6611
$10^{0.5}$	**0.1747**	0.184	0.314	0.6585
10^{1}	0.1767	**0.1766**	0.3246	0.5176
$10^{1.5}$	0.1781	0.1776	0.3299	**0.3282**
10^{2}	0.1787	0.1787	0.3309	0.331
$10^{2.5}$	0.179	0.179	0.3314	0.3314
10^{3}	0.1789	0.1789	0.3316	0.3316
$10^{3.5}$	0.1787	0.1787	0.3318	0.3318
10^{4}	0.1787	0.1787	0.3318	0.3318
∞	0.1786	0.1786	0.3318	0.3318
λ_0	0.1754	0.2157	0.3258	0.4223

$p = 10, n = 60, v_2 = 3$

ϱ	0.1		0.9	
λ_0	3.6		16.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1986	0.6974	0.1582	0.638
$10^{-1.5}$	0.1981	0.6954	**0.1551**	0.6379
10^{-1}	0.196	0.692	**0.1551**	0.6379
$10^{-0.5}$	0.1954	0.6744	0.1745	0.6374
10^{0}	**0.1925**	0.5869	0.2364	0.6363
$10^{0.5}$	0.1996	0.2589	0.2961	0.6328
10^{1}	0.2042	**0.202**	0.3264	0.5921
$10^{1.5}$	0.207	0.2065	0.3403	**0.3312**
10^{2}	0.2086	0.2083	0.3456	0.3448
$10^{2.5}$	0.2091	0.2092	0.3473	0.3472
10^{3}	0.2092	0.2092	0.3476	0.3476
$10^{3.5}$	0.2091	0.2091	0.3477	0.3477
10^{4}	0.2091	0.2091	0.3477	0.3477
∞	0.2092	0.2092	0.3478	0.3478
λ_0	0.1986	0.232	0.3347	0.3683

$p = 100, n = 20, v_2 = 1$

ϱ	0.1		0.9	
λ_0	9.8		88.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1089	0.9229	0.3691	0.6923
$10^{-1.5}$	0.1084	0.9226	0.3667	0.6923
10^{-1}	0.1072	0.9223	0.3561	0.6923
$10^{-0.5}$	0.1033	0.9217	0.3431	0.6923
10^{0}	0.0946	0.9169	0.3232	0.6922
$10^{0.5}$	0.083	0.8757	0.3128	0.6923
10^{1}	0.0742	0.3835	**0.3078**	0.6923
$10^{1.5}$	0.0716	0.0737	0.3084	0.6901
10^{2}	0.0706	0.0709	0.3085	0.4252
$10^{2.5}$	**0.0697**	**0.0697**	0.3083	0.3083
10^{3}	0.0698	0.0698	0.3083	0.3083
$10^{3.5}$	0.0698	0.0698	0.3082	**0.3082**
10^{4}	0.0698	0.0698	0.3083	0.3083
∞	0.0699	0.0699	0.3083	0.3083
λ_0	0.0744	0.3993	0.3086	0.4769

$p = 100, n = 20, v_2 = 2$

ϱ	0.1		0.9	
λ_0	15.7		133.3	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1484	0.8717	**0.1038**	0.6568
$10^{-1.5}$	0.1482	0.8717	0.1045	0.6568
10^{-1}	0.1475	0.8716	0.1049	0.6568
$10^{-0.5}$	0.145	0.871	0.1102	0.6567
10^{0}	0.1396	0.8693	0.1386	0.6567
$10^{0.5}$	0.1268	0.8578	0.2029	0.6563
10^{1}	0.1175	0.6563	0.2675	0.656
$10^{1.5}$	0.112	0.1278	0.3067	0.6527
10^{2}	0.1109	0.1107	0.3242	0.5412
$10^{2.5}$	0.1107	0.1107	0.3303	**0.3279**
10^{3}	0.111	0.111	0.3322	0.332
$10^{3.5}$	**0.1106**	**0.1106**	0.3331	0.333
10^{4}	**0.1106**	**0.1106**	0.3332	0.3332
∞	0.1107	0.1107	0.3333	0.3333
λ_0	0.1145	0.3858	0.3268	0.4286

$p = 100, n = 20, v_2 = 3$

ϱ	0.1		0.9	
λ_0	21.6		178.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1722	0.8351	**0.0271**	0.623
$10^{-1.5}$	0.1717	0.8351	0.0275	0.623
10^{-1}	0.1705	0.8349	0.0283	0.623
$10^{-0.5}$	0.1695	0.8349	0.0322	0.6229
10^{0}	0.1646	0.8325	0.0496	0.6226
$10^{0.5}$	0.1575	0.8229	0.1052	0.6219
10^{1}	0.1492	0.7302	0.2004	0.6204
$10^{1.5}$	**0.1436**	0.2015	0.2852	0.6154
10^{2}	**0.1436**	0.144	0.3294	0.5506
$10^{2.5}$	**0.1436**	0.1437	0.3484	**0.3256**
10^{3}	**0.1436**	**0.1436**	0.3542	0.3537
$10^{3.5}$	0.1437	0.1437	0.3562	0.3562
10^{4}	0.1437	0.1437	0.3569	0.3569
∞	0.1439	0.1439	0.357	0.357
λ_0	0.1447	0.3783	0.3402	0.3707

$p = 100, n = 40, v_2 = 1$

ϱ	0.1		0.9	
λ_0	9.8		88.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1421	0.9253	0.3878	0.7037
$10^{-1.5}$	0.1397	0.9253	0.3762	0.7037
10^{-1}	0.1321	0.9256	0.3535	0.7037
$10^{-0.5}$	0.1159	0.925	0.3257	0.7037
10^{0}	0.0947	0.9238	0.305	0.7037
$10^{0.5}$	0.0778	0.9147	0.2991	0.7037
10^{1}	0.0741	0.3442	0.2972	0.7037
$10^{1.5}$	0.0738	0.0736	**0.2964**	0.7039
10^{2}	0.0727	0.0726	0.2966	0.4033
$10^{2.5}$	0.0723	**0.0721**	0.2967	**0.2968**
10^{3}	**0.0721**	**0.0721**	0.2971	0.2971
$10^{3.5}$	0.0722	0.0722	0.2971	0.2971
10^{4}	**0.0721**	**0.0721**	0.2971	0.2971
∞	**0.0721**	**0.0721**	0.2971	0.2971
λ_0	0.0741	0.3661	0.2966	0.4877

$p = 100, n = 40, v_2 = 2$

ϱ	0.1		0.9	
λ_0	15.7		133.3	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1717	0.8744	0.0786	0.6558
$10^{-1.5}$	0.1689	0.8744	0.0722	0.6557
10^{-1}	0.1646	0.874	**0.0642**	0.6557
$10^{-0.5}$	0.1544	0.8738	0.0678	0.6557
10^{0}	0.1335	0.8726	0.1071	0.6557
$10^{0.5}$	0.1147	0.8662	0.1918	0.6556
10^{1}	0.1093	0.7195	0.2755	0.6551
$10^{1.5}$	0.1107	**0.11**	0.3135	0.6534
10^{2}	0.1118	0.1118	0.3281	0.571
$10^{2.5}$	0.1136	0.1136	0.3339	**0.3319**
10^{3}	0.1141	0.1141	0.3369	0.3366
$10^{3.5}$	0.114	0.114	0.3374	0.3374
10^{4}	0.114	0.114	0.3374	0.3374
∞	0.114	0.114	0.3375	0.3375
λ_0	**0.1084**	0.3386	0.3308	0.4166

$p = 100, n = 40, v_2 = 3$

ϱ	0.1		0.9	
λ_0	21.6		178.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1766	0.8319	0.017	0.6302
$10^{-1.5}$	0.1747	0.8319	0.0151	0.6302
10^{-1}	0.1722	0.8317	0.0114	0.6302
$10^{-0.5}$	0.1632	0.8311	**0.0108**	0.6302
10^{0}	0.1484	0.8295	0.0246	0.6301
$10^{0.5}$	0.1302	0.8245	0.0857	0.6301
10^{1}	**0.1274**	0.7717	0.2002	0.6294
$10^{1.5}$	0.1332	0.1438	0.2863	0.626
10^{2}	0.1377	**0.1366**	0.3301	0.596
$10^{2.5}$	0.1398	0.1398	0.3476	**0.3323**
10^{3}	0.1405	0.1405	0.3521	0.3516
$10^{3.5}$	0.1409	0.1409	0.3536	0.3535
10^{4}	0.1412	0.1412	0.354	0.354
∞	0.1411	0.1411	0.3543	0.3543
λ_0	0.1303	0.3149	0.3406	0.3634

$p = 100, n = 60, v_2 = 1$

ϱ	0.1		0.9	
λ_0	9.8		88.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.1954	0.9293	0.4104	0.6966
$10^{-1.5}$	0.1851	0.9293	0.3925	0.6966
10^{-1}	0.1636	0.929	0.3602	0.6966
$10^{-0.5}$	0.1292	0.9292	0.3248	0.6966
10^{0}	0.0967	0.9283	0.3084	0.6966
$10^{0.5}$	0.0762	0.9238	0.3047	0.6966
10^{1}	0.0701	0.3277	**0.3026**	0.6965
$10^{1.5}$	0.0681	0.0687	0.3031	0.6965
10^{2}	0.0673	0.0675	0.3033	0.3925
$10^{2.5}$	**0.0667**	**0.0667**	0.3031	**0.3031**
10^{3}	**0.0667**	**0.0667**	0.3034	0.3034
$10^{3.5}$	0.0668	0.0668	0.3034	0.3034
10^{4}	0.0668	0.0668	0.3034	0.3034
∞	0.0668	0.0668	0.3034	0.3034
λ_0	0.0701	0.3513	0.3033	0.4902

$p = 100, n = 60, v_2 = 2$

ϱ	0.1		0.9	
λ_0	15.7		133.3	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.215	0.8754	0.0797	0.6542
$10^{-1.5}$	0.2067	0.8754	0.0593	0.6542
10^{-1}	0.1901	0.8754	**0.0434**	0.6542
$10^{-0.5}$	0.1639	0.8752	0.0439	0.6542
10^{0}	0.1315	0.8734	0.0852	0.6542
$10^{0.5}$	0.1123	0.8717	0.1894	0.654
10^{1}	**0.1071**	0.7658	0.2801	0.6538
$10^{1.5}$	0.1093	**0.1073**	0.3182	0.6525
10^{2}	0.1113	0.1112	0.3333	0.5926
$10^{2.5}$	0.1118	0.1116	0.3387	**0.3368**
10^{3}	0.112	0.112	0.3399	0.3398
$10^{3.5}$	0.112	0.112	0.34	0.34
10^{4}	0.112	0.112	0.3401	0.3401
∞	0.112	0.112	0.3401	0.3401
λ_0	0.1087	0.3218	0.3355	0.421

$p = 100, n = 60, v_2 = 3$

ϱ	0.1		0.9	
λ_0	21.6		178.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.2066	0.8342	0.0185	0.6319
$10^{-1.5}$	0.2001	0.8342	0.011	0.6319
10^{-1}	0.1869	0.8342	0.005	0.6319
$10^{-0.5}$	0.1663	0.8339	**0.0039**	0.6319
10^{0}	0.1408	0.8326	0.0121	0.6319
$10^{0.5}$	**0.124**	0.8292	0.0719	0.6317
10^{1}	0.1266	0.7929	0.2027	0.631
$10^{1.5}$	0.1342	**0.1316**	0.2978	0.6294
10^{2}	0.1397	0.1388	0.3365	0.6097
$10^{2.5}$	0.1411	0.141	0.3486	**0.3398**
10^{3}	0.1416	0.1415	0.3544	0.3538
$10^{3.5}$	0.1421	0.1421	0.3557	0.3557
10^{4}	0.1421	0.1421	0.3562	0.3562
∞	0.1421	0.1421	0.3565	0.3565
λ_0	0.1307	0.2883	0.344	0.3596

$p = 500, n = 20, v_2 = 1$

ϱ	0.1		0.9	
λ_0	49.8		448.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.0687	0.9372	0.3369	0.6897
$10^{-1.5}$	0.0687	0.9372	0.3367	0.6897
10^{-1}	0.0687	0.9372	0.3352	0.6897
$10^{-0.5}$	0.0684	0.9373	0.3322	0.6897
10^{0}	0.0681	0.9371	0.326	0.6897
$10^{0.5}$	0.0672	0.9365	0.3155	0.6897
10^{1}	0.0655	0.9346	0.3098	0.6896
$10^{1.5}$	0.0634	0.7929	**0.3097**	0.6896
10^{2}	**0.062**	0.0723	0.3112	0.6893
$10^{2.5}$	**0.062**	0.0623	0.3102	0.617
10^{3}	0.0622	0.0621	0.3099	0.311
$10^{3.5}$	**0.062**	**0.062**	0.3101	**0.3101**
10^{4}	0.0621	0.0621	0.3101	**0.3101**
∞	0.0621	0.0621	0.3101	**0.3101**
λ_0	0.0623	0.4395	**0.3097**	0.4847

$p = 500, n = 20, v_2 = 2$

ϱ	0.1		0.9	
λ_0	75.7		673.3	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.0981	0.8901	**0.0375**	0.6488
$10^{-1.5}$	0.0981	0.8901	0.0379	0.6488
10^{-1}	0.0981	0.8901	0.0389	0.6488
$10^{-0.5}$	0.098	0.89	0.0422	0.6488
10^{0}	0.0978	0.8898	0.0527	0.6488
$10^{0.5}$	0.0981	0.8893	0.0842	0.6488
10^{1}	**0.0971**	0.8874	0.1521	0.6487
$10^{1.5}$	0.0983	0.8563	0.2383	0.6481
10^{2}	0.0988	0.2046	0.3022	0.6468
$10^{2.5}$	0.0999	**0.099**	0.3273	0.6205
10^{3}	0.1006	0.1006	0.3385	**0.3407**
$10^{3.5}$	0.1006	0.1006	0.3422	0.3414
10^{4}	0.1012	0.1012	0.3432	0.3432
∞	0.1015	0.1015	0.3441	0.3441
λ_0	0.0986	0.406	0.3359	0.4269

$p = 500, n = 20, v_2 = 3$

ϱ	0.1		0.9	
λ_0	101.6		898.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.113	0.8535	**0.0062**	0.6267
$10^{-1.5}$	0.1129	0.8534	0.0063	0.6267
10^{-1}	0.1128	0.8534	0.0065	0.6267
$10^{-0.5}$	0.113	0.8534	0.0073	0.6267
10^{0}	0.1128	0.8533	0.01	0.6266
$10^{0.5}$	0.1128	0.8528	0.0209	0.6266
10^{1}	**0.1127**	0.8507	0.0569	0.6264
$10^{1.5}$	0.1157	0.8318	0.1517	0.6257
10^{2}	0.1201	0.3745	0.2516	0.6225
$10^{2.5}$	0.1264	**0.122**	0.3128	0.608
10^{3}	0.1285	0.1285	0.3361	0.3457
$10^{3.5}$	0.1289	0.1288	0.3465	**0.3431**
10^{4}	0.1291	0.1291	0.3505	0.3503
∞	0.1289	0.1289	0.3519	0.3519
λ_0	0.1203	0.3637	0.3351	0.3767

$p = 500, n = 40, v_2 = 1$

ϱ	0.1		0.9	
λ_0	49.8		448.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.0731	0.9381	0.3306	0.6936
$10^{-1.5}$	0.0731	0.9381	0.3307	0.6936
10^{-1}	0.0729	0.9381	0.3292	0.6936
$10^{-0.5}$	0.0726	0.9381	0.3208	0.6936
10^{0}	0.0721	0.9379	0.312	0.6936
$10^{0.5}$	0.0683	0.938	0.3024	0.6936
10^{1}	0.0665	0.9385	0.3026	0.6936
$10^{1.5}$	0.0631	0.8854	**0.3001**	0.6936
10^{2}	0.0638	0.0632	0.3016	0.6935
$10^{2.5}$	**0.0624**	0.0625	0.3016	0.6561
10^{3}	**0.0624**	**0.0624**	0.3017	**0.3066**
$10^{3.5}$	**0.0624**	**0.0624**	0.3018	**0.3066**
10^{4}	**0.0624**	**0.0624**	0.3018	**0.3066**
∞	**0.0624**	**0.0624**	0.3019	**0.3066**
λ_0	0.0627	0.4609	0.3018	0.4793

$p = 500, n = 40, v_2 = 2$

ϱ	0.1		0.9	
λ_0	75.7		673.3	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.0932	0.8934	**0.0033**	0.6572
$10^{-1.5}$	0.0932	0.8934	0.0034	0.6572
10^{-1}	0.0931	0.8934	0.0037	0.6572
$10^{-0.5}$	0.0932	0.8934	0.0054	0.6572
10^{0}	0.0925	0.8934	0.0118	0.6571
$10^{0.5}$	**0.0921**	0.8932	0.042	0.6571
10^{1}	**0.0921**	0.8923	0.1275	0.6571
$10^{1.5}$	0.0932	0.8828	0.2372	0.657
10^{2}	0.098	0.1423	0.3002	0.6566
$10^{2.5}$	0.1	**0.0996**	0.3252	0.6481
10^{3}	0.1	0.1	0.3327	**0.3245**
$10^{3.5}$	0.1003	0.1003	0.3352	0.3346
10^{4}	0.1003	0.1003	0.3358	0.3358
∞	0.1003	0.1003	0.3365	0.3365
λ_0	0.0969	0.4032	0.3314	0.4114

$p = 500, n = 40, v_2 = 3$

ϱ	0.1		0.9	
λ_0	101.6		898.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.0948	0.8531	**0.0**	0.6323
$10^{-1.5}$	0.0948	0.8531	**0.0**	0.6323
10^{-1}	0.0949	0.8531	**0.0**	0.6323
$10^{-0.5}$	0.0948	0.8531	**0.0**	0.6322
10^{0}	**0.0945**	0.853	0.0001	0.6321
$10^{0.5}$	0.0954	0.8526	0.0025	0.632
10^{1}	0.0991	0.851	0.0339	0.632
$10^{1.5}$	0.1084	0.8435	0.139	0.6318
10^{2}	0.1191	0.3563	0.2565	0.6305
$10^{2.5}$	0.1245	**0.1227**	0.3182	0.6226
10^{3}	0.1268	0.1267	0.3414	**0.3222**
$10^{3.5}$	0.1281	0.128	0.35	0.3479
10^{4}	0.1285	0.1284	0.3519	0.3519
∞	0.1286	0.1286	0.3531	0.3531
λ_0	0.1194	0.3385	0.3403	0.356

$p = 500, n = 60, v_2 = 1$

ϱ	0.1		0.9	
λ_0	49.8		448.2	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.0778	0.9386	0.3443	0.6982
$10^{-1.5}$	0.0779	0.9386	0.3439	0.6982
10^{-1}	0.0776	0.9386	0.3397	0.6982
$10^{-0.5}$	0.077	0.9387	0.3319	0.6982
10^{0}	0.0741	0.9388	0.3194	0.6982
$10^{0.5}$	0.0689	0.9388	0.3174	0.6982
10^{1}	0.0653	0.9385	**0.3165**	0.6982
$10^{1.5}$	0.0625	0.9137	**0.3165**	0.6982
10^{2}	0.0614	0.0619	0.3167	0.6982
$10^{2.5}$	0.0611	0.0611	0.3174	0.6796
10^{3}	**0.0606**	**0.0606**	0.3173	0.3026
$10^{3.5}$	0.0607	0.0607	0.3175	**0.3022**
10^{4}	**0.0606**	**0.0606**	0.3175	**0.3022**
∞	**0.0606**	**0.0606**	0.3175	**0.3022**
λ_0	0.0615	0.4517	0.3173	0.4913

$p = 500, n = 60, v_2 = 2$

ϱ	0.1		0.9	
λ_0	75.7		673.3	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	0.087	0.8905	**0.0**	0.6463
$10^{-1.5}$	0.0871	0.8905	**0.0**	0.6463
10^{-1}	0.0871	0.8905	0.0002	0.6463
$10^{-0.5}$	0.0865	0.8905	0.0009	0.6463
10^{0}	**0.0852**	0.8905	0.0029	0.6463
$10^{0.5}$	0.0857	0.8906	0.0258	0.6463
10^{1}	0.0869	0.8898	0.124	0.6463
$10^{1.5}$	0.0929	0.8834	0.2517	0.6463
10^{2}	0.0978	0.1167	0.3163	0.6458
$10^{2.5}$	0.1019	**0.1011**	0.3364	0.6413
10^{3}	0.1024	0.1024	0.3444	**0.3336**
$10^{3.5}$	0.1028	0.1028	0.3466	0.3464
10^{4}	0.1027	0.1027	0.3477	0.3477
∞	0.1027	0.1027	0.3477	0.3477
λ_0	0.0972	0.3774	0.3421	0.4101

$p = 500, n = 60, v_2 = 3$

ϱ	0.1		0.9	
λ_0	101.6		898.4	
λ	RDA	RDA*	RDA	RDA*
10^{-2}	**0.0751**	0.8504	**0.0**	0.6397
$10^{-1.5}$	0.0752	0.8504	**0.0**	0.6397
10^{-1}	0.0752	0.8504	**0.0**	0.6397
$10^{-0.5}$	0.0755	0.8503	**0.0**	0.6397
10^{0}	0.0773	0.8501	**0.0**	0.6397
$10^{0.5}$	0.0794	0.8498	0.0007	0.6396
10^{1}	0.0867	0.8477	0.0213	0.6394
$10^{1.5}$	0.1036	0.8411	0.1366	0.6391
10^{2}	0.1195	0.326	0.261	0.6377
$10^{2.5}$	0.1267	**0.1241**	0.3178	0.6313
10^{3}	0.1294	0.1292	0.3396	**0.3058**
$10^{3.5}$	0.1307	0.1307	0.3463	0.3453
10^{4}	0.1309	0.1309	0.3478	0.3476
∞	0.1311	0.1311	0.3487	0.3487
λ_0	0.1196	0.3035	0.3384	0.3597

Anhang B

Die SAS/IML-Module rda_lambda0, rdamod_lambda0 und vca

B.1 Beschreibung

Innerhalb der SAS/IML-Umgebung können zur Anwendung der Verfahren RDA-λ_0, RDA*-λ_0 und VCA für die Klassifikation die Module rda_lambda0, rdamod_lambda0 bzw. vca verwendet werden. Eingelesen wird jeweils eine Datenmatrix, die die Lernstichprobenelemente als Zeilenvektoren enthält, ein Vektor, in dem die Klassenzugehörigkeit der Lernstichprobenelemente kodiert ist, und der zu klassifizierende Datenvektor. Die ermittelte Klassenzugehörigkeit wird jeweils als Funktionswert zurückgegeben und kann innerhalb der SAS/IML-Umgebung weiter verwendet werden. Vom Modul vca wird als Ergebnis ein zweielementiger Vektor zurückgegeben, der als Komponenten die Klassifizierungen nach den beiden verschiedenen Versionen des Verfahrens VCA enthält, bei den anderen beiden Modulen sind die Ergebnisse die jeweiligen skalaren Klassifizierungsresultate.

B.2 Programmtext der Module

B.2.1 rda_lambda0

```
start rda_lambda0(x,y,cl); /*Beginn Modul*/
      ncl=2;               /*Anzahl der Lernstichproben*/
      p=ncol(x);           /*Anzahl der Variablen*/
      n=nrow(x);           /*Gesamt-Lernstichprobenumfang*/
      df=n-ncl;
      mean=j(ncl,p,0);
      num=j(ncl,1,0);
```

```
      s=0;
      f=0;
      do k=1 to n;
         mean[cl[k],]=mean[cl[k],]+x[k,];
         num[cl[k]]=num[cl[k]]+1;
         end;
      do j=1 to ncl;
         if num[j]^=0 then mean[j,]=mean[j,]/num[j];
         end;
      do k=1 to n;
         s=s+(x[k,]-mean[cl[k],])‘*(x[k,]-mean[cl[k],]);
         end;
      if df>0 then s=1/df*s;
      dmean=mean[1]-mean[2];
      term1=dmean*dmean‘;
      term2=dmean*s*dmean‘;
      term3=dmean*s**2*dmean‘;
      term4=dmean*s**3*dmean‘;
      zähler=term2*term3-term1*term4;
      nenner=(term2)**2-term1*term3;
      if nenner^=0 then tlambda0=zähler/nenner;
      else tlambda0=trace(s);
      lambda0=tlambda0-2;
      s=s+lambda0*i(p);
      d=inv(s)*(mean[1,]-mean[2,])‘;
      f=(y-1/2*(mean[1,]+mean[2,]))*d;
      if f>0 then discr=1;
      else discr=2;
      return(discr);
finish rda_lambda0; /*Ende Modul*/
```

B.2.2 rdamod_lambda0

```
start rdamod_lambda0(x,y,cl); /*Beginn Modul*/
      ncl=2;                  /*Anzahl der Lernstichproben*/
      p=ncol(x);              /*Anzahl der Variablen*/
      n=nrow(x);              /*Gesamt-Lernstichprobenumfang*/
      df=n-ncl;
      mean=j(ncl,p,0);
      num=j(ncl,1,0);
      s=0;
      f=0;
      do k=1 to n;
```

```
         mean[cl[k],]=mean[cl[k],]+x[k,];
         num[cl[k]]=num[cl[k]]+1;
         end;
      do j=1 to ncl;
         if num[j]^=0 then mean[j,]=mean[j,]/num[j];
         end;
      do k=1 to n;
         s=s+(x[k,]-mean[cl[k],])‘*(x[k,]-mean[cl[k],]);
         end;
      if df>0 then s=1/df*s;
      dmean=mean[1]-mean[2];
      term1=dmean*dmean‘;
      term2=dmean*s*dmean‘;
      term3=dmean*s**2*dmean‘;
      term4=dmean*s**3*dmean‘;
      zähler=term2*term3-term1*term4;
      nenner=(term2)**2-term1*term3;
      if nenner^=0 then tlambda0=zähler/nenner;
      else tlambda0=trace(s);
      d=(tlambda0*i(p)-s)*(mean[1,]-mean[2,])‘;
      f=(y-1/2*(mean[1,]+mean[2,]))*d;
      if f>0 then discr=1;
      else discr=2;
      return(discr);
finish rdamod_lambda0; /*Ende Modul*/
```

B.2.3 vca

```
start vca(x,y,cl); /*Beginn Modul*/
      ncl=2;       /*Anzahl der Lernstichproben*/
      c=0.5;       /*Schranke für Variablenbildung*/
      p=ncol(x);   /*Anzahl der Variablen*/
      n=nrow(x);   /*Gesamt-Lernstichprobenumfang*/
      pmax=n/6;    /*maximale Anzahl der verwendeten Variablen*/
      df=n-ncl;
      discr=j(1,2,0);
      m=j(p,p,0);
      xm1=j(n+1,p);
      xm2=j(n+1,p);
      mean=0;
      do jn=1 to n;
         mean=mean+x[jn,];
         end;
```

```
meanges=j(n,1,1)*(mean/n);
xd=x-meanges;
w=inv(root(diag(xd‘*xd)));
r=w*xd‘*xd*w;
rq=r##2;
nm1=0;
nm2=0;
x=x//y;

/*i=1*/
a=j(n+1,p);
ad=j(n,p);
l=0;
do k=1 to p;
   if rq[1,k]>=c then do;
      m[1,k]=1;
      l=l+1;
      a[,l]=x[,k];
      ad[,l]=xd[,k];
      end;
   end;
a=a[,1:l];
ad=ad[,1:l];
w=ad‘*ad;
call eigen(ew,e,w);
nm1=nm1+1;
xm1[,nm1]=a*e[,1];
if l>1 then do;
   nm2=nm2+1;
   xm2[,nm2]=a*e[,1];
   end;

do i=2 to p;
   a=j(n+1,p);
   ad=j(n,p);
   l=0;
   do k=1 to p;
      if rq[i,k]>=c then m[i,k]=1;
      end;
   mneu=1;
   do mi=1 to i-1;
      if m[i,]=m[mi,] then mneu=0;
      end;
```

```
   if mneu=1 then do;
      do k=1 to p;
         if m[i,k]=1 then do;
            l=l+1;
            a[,l]=x[,k];
            ad[,l]=xd[,k];
            end;
         end;
      a=a[,1:l];
      ad=ad[,1:l];
      w=ad‘*ad;
      call eigen(ew,e,w);
      nm1=nm1+1;
      xm1[,nm1]=a*e[,1];
      if l>1 then do;
         nm2=nm2+1;
         xm2[,nm2]=a*e[,1];
         end;
      end;
   end;

/*Version 1*/
pm=nm1;
xm=xm1[,1:pm];
mean=j(ncl,pm,0);
num=j(ncl,1,0);
t=0;
s=0;
f=0;
do jn=1 to n;
   mean[cl[jn],]=mean[cl[jn],]+xm[jn,];
   num[cl[jn]]=num[cl[jn]]+1;
   end;
do j=1 to ncl;
   if num[j]^=0 then mean[j,]=mean[j,]/num[j];
   end;
if pm>pmax then do;
   va=0;
   do jn=1 to n;
      va=va+(xm[jn,]-mean[cl[jn],])##2;
      end;
   if df>0 then va=1/df*va;
   t=abs(mean[1,]-mean[2,])/sqrt((1/num[1]+1/num[2])*va);
```

```
   do i=1 to pmax;
      imax=t[<:>];
      xm[,i]=xm1[,imax];
      t[imax]=0;
      end;
   xm=xm[,1:pmax];
   mean=j(ncl,pmax,0);
   do jn=1 to n;
      mean[cl[jn],]=mean[cl[jn],]+xm[jn,];
      end;
   do j=1 to ncl;
      if num[j]^=0 then mean[j,]=mean[j,]/num[j];
      end;
   end;
do jn=1 to n;
   s=s+(xm[jn,]-mean[cl[jn],])`*(xm[jn,]-mean[cl[jn],]);
   end;
do while(det(s)=0);
   pm=ncol(xm)-1;
   va=0;
   do jn=1 to n;
      va=va+(xm[jn,]-mean[cl[jn],])##2;
      end;
   if df>0 then va=1/df*va;
   t=abs(mean[1,]-mean[2,])/sqrt((1/num[1]+1/num[2])*va);
   do i=1 to pm;
      imax=t[<:>];
      xm[,i]=xm1[,imax];
      t[imax]=0;
      end;
   xm=xm[,1:pm];
   mean=j(ncl,pm,0);
   do jn=1 to n;
      mean[cl[jn],]=mean[cl[jn],]+xm[jn,];
      end;
   do j=1 to ncl;
      if num[j]^=0 then mean[j,]=mean[j,]/num[j];
      end;
   s=0;
   do jn=1 to n;
      s=s+(xm[jn,]-mean[cl[jn],])`*(xm[jn,]-mean[cl[jn,]]);
      end;
   end;
```

```
if df>0 then s=1/df*s;
d=inv(s)*(mean[1,]-mean[2,])‘;
f=(xm[in,]-1/2*(mean[1,]+mean[2,]))*d;
if f>0 then discr[1]=1;
else discr[1]=2;

/*Version 2 (keine Verwendung der einelementigen
Variablenmengen)*/
if nm2=0 then do;
   pm=p;
   pmax=1;
   xm2=x;
   end;
else pm=nm2;
xm=xm2[,1:pm];
mean=j(ncl,pm,0);
num=j(ncl,1,0);
t=0;
s=0;
f=0;
do jn=1 to n;
   mean[cl[jn],]=mean[cl[jn],]+xm[jn,];
   num[cl[jn]]=num[cl[jn]]+1;
   end;
do j=1 to ncl;
   if num[j]^=0 then mean[j,]=mean[j,]/num[j];
   end;
if pm>pmax then do;
   va=0;
   do jn=1 to n;
      va=va+(xm[jn,]-mean[cl[jn],])##2;
      end;
   if df>0 then va=1/df*va;
   t=abs(mean[1,]-mean[2,])/sqrt((1/num[1]+1/num[2])*va);
   do i=1 to pmax;
      imax=t[<:>];
      xm[,i]=xm2[,imax];
      t[imax]=0;
      end;
   xm=xm[,1:pmax];
   mean=j(ncl,pmax,0);
   do jn=1 to n;
      mean[cl[jn],]=mean[cl[jn],]+xm[jn,];
```

```
            end;
         do j=1 to ncl;
            if num[j]^=0 then mean[j,]=mean[j,]/num[j];
            end;
         end;
      do jn=1 to n;
         s=s+(xm[jn,]-mean[cl[jn],])‘*(xm[jn,]-mean[cl[jn],]);
         end;
      do while(det(s)=0);
         pm=ncol(xm)-1;
         va=0;
         do jn=1 to n;
            va=va+(xm[jn,]-mean[cl[jn],])##2;
            end;
         if df>0 then va=1/df*va;
         t=abs(mean[1,]-mean[2,])/sqrt((1/num[1]+1/num[2])*va);
         do i=1 to pm;
            imax=t[<:>];
            xm[,i]=xm1[,imax];
            t[imax]=0;
            end;
         xm=xm[,1:pm];
         mean=j(ncl,pm,0);
         do jn=1 to n;
            mean[cl[jn],]=mean[cl[jn],]+xm[jn,];
            end;
         do j=1 to ncl;
            if num[j]^=0 then mean[j,]=mean[j,]/num[j];
            end;
         s=0;
         do jn=1 to n;
            s=s+(xm[jn,]-mean[cl[jn],])‘*(xm[jn,]-mean[cl[jn,]]);
            end;
         end;
      if df>0 then s=1/df*s;
      d=inv(s)*(mean[1,]-mean[2,])‘;
      f=(xm[n+1,]-1/2*(mean[1,]+mean[2,]))*d;
      if f>0 then discr[2]=1;
      else discr[2]=2;
      return(discr);
finish vca; /*Ende Modul*/
```

B.3 Argumente

`x` Datenmatrix, deren Spalten die Variablen und deren Zeilen die einzelnen Lernstichprobenelemente darstellen

`cl` Spaltenvektor, der zu jedem Lernstichprobenvektor (d. h. zu jeder Zeile der Matrix `x`) die Klassenzugehörigkeit angibt. Die Klassen müssen hierbei jeweils durch die Zahlenwerte 1 bzw. 2 kodiert sein.

`y` zu klassifizierender Datenvektor als Zeilenvektor

B.4 Beispiel

Im folgendenden Beispiel soll die Anwendung des Moduls rda_lambda0 demonstriert werden. Durch Simulationen soll die Fehlerrate des Verfahrens RDA-λ_0 ermittelt werden. Es wird hier ein Normalverteilungsmodell verwendet, wie es auch bei den Simulationen in Abschnitt 9.2.1 dieser Arbeit beschrieben ist. In jedem der 10 000 Simulationsläufe werden sowohl die Lerndatenvektoren als auch der zu klassifizierende Datenvektor jeweils mit Hilfe von Pseudozufallszahlen generiert. Anschließend wird das Modul rda_lambda0 aufgerufen, das mit Anwendung des Verfahrens RDA-λ_0 die Klassifizierung ausführt und das Ergebnis zurückgibt. Von allen Simulationsläufen werden die falschen Zuordnungen gezählt. Die Division durch die Anzahl der Simulationsläufe liefert schließlich die Fehlerrate des Verfahrens RDA-λ_0.

B.4.1 Anwendung

```
/*Simulationen zur Fehlerratenschätzung*/
/*Verfahren: RDA-lambda0*/

proc iml;
start rda_lambda0(x,y,cl); /*Beginn Modul*/
      ncl=2;               /*Anzahl der Lernstichproben*/
      p=ncol(x);           /*Anzahl der Variablen*/
      n=nrow(x);           /*Gesamt-Lernstichprobenumfang*/
      df=n-ncl;
      mean=j(ncl,p,0);
      num=j(ncl,1,0);
      s=0;
      f=0;
      do k=1 to n;
         mean[cl[k],]=mean[cl[k],]+x[k,];
         num[cl[k]]=num[cl[k]]+1;
```

```
         end;
      do j=1 to ncl;
         if num[j]^=0 then mean[j,]=mean[j,]/num[j];
         end;
      do k=1 to n;
         s=s+(x[k,]-mean[cl[k],])‘*(x[k,]-mean[cl[k],]);
         end;
      if df>0 then s=1/df*s;
      dmean=mean[1]-mean[2];
      term1=dmean*dmean‘;
      term2=dmean*s*dmean‘;
      term3=dmean*s**2*dmean‘;
      term4=dmean*s**3*dmean‘;
      zähler=term2*term3-term1*term4;
      nenner=(term2)**2-term1*term3;
      if nenner^=0 then tlambda0=zähler/nenner;
      else tlambda0=trace(s);
      lambda0=tlambda0-2;
      s=s+lambda0*i(p);
      d=inv(s)*(mean[1,]-mean[2,])‘;
      f=(y-1/2*(mean[1,]+mean[2,]))*d;
      if f>0 then discr=1;
      else discr=2;
      return(discr);
finish rda_lambda0; /*Ende Modul*/

nsim=10000; /*Anzahl der Simulationen*/
n1=10;      /*Umfang der ersten Lernstichprobe*/
n2=10;      /*Umfang der zweiten Lernstichprobe*/
p=50;       /*Anzahl der Variablen*/
corr=0.5;   /*paarweise Korrelationen zwischen den Variablen*/
mu=1;       /*Element des Differenzvektors der Erwartungswerte
            (my1 - my2 = (mu,mu,...,mu))*/
n=n1+n2;    /*Gesamt-Lernstichprobenumfang*/
cl1=j(n1,1,1);
cl2=j(n2,1,2);
cl=cl1//cl2;

/*Kovarianzmatrix Sigma*/
sigma=j(p,p,corr);
do i=1 to p;
   sigma[i,i]=1;
   end;
```

```
b=root(sigma);
mu1=j(1,p,mu);
mu2=j(1,p,0);
d=(mu1-mu2)*inv(sigma)*(mu1-mu2)‘;

/*Simulationen*/
count=0;
do sim=1 to nsim;
   j=j(n+1,p);
   z=normal(j);
   x=j(n,p);
   do in=1 to n;
      if cl[in]=1 then x[in,]=z[in,]*b+mu1;
      else x[in,]=z[in,]*b+mu2;
      end;
   cly=mod(sim,2)+1;
   if cly=1 then y=z[n+1,]*b+mu1;
   else y=z[n+1,]*b+mu2;
   if rda_lambda0(x,y,cl)^=cly then count=count+1;
   end;
fehler_rda_lambda0=count/nsim; /*Fehlerrate von RDA-lambda0*/
print corr p fehler_rda_lambda0;
quit;

run;
```

B.4.2 Ergebnisausdruck

Als Ergebnis erhält man den folgenden Ausdruck auf dem Bildschirm:

CORR	P	FEHLER_RDA_LAMBDA0
0.5	50	0.2451

Die Bedeutung ist hierbei wie folgt:

CORR:	paarweise Korrelation zwischen verschiedenen Variablen
P:	Anzahl der Variablen
FEHLER_RDA_LAMBDA0:	ermittelte Fehlerrate des Verfahrens RDA-λ_0 (das eigentliche Ergebnis)