Flexible Modellierung kategorialer Responsevariablen

Dissertation

zur Erlangung des Grades Doctor rerum naturalium
(Dr. rer. nat.)
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

vorgelegt von
Torsten Scholz

*

21. November 2003

Flexible Modellierung kategorialer Responsevariablen

Dissertation

zur Erlangung des Grades Doctor rerum naturalium
(Dr. rer. nat.)
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

vorgelegt von

Torsten Scholz

*

21. November 2003

Gutachter:	Prof. Dr. Gerhard Tutz
Koreferent:	PD Dr. Helmut Küchenhoff
Externer Gutachter:	Prof. Dr. Göran Kauermann
Rigorosum:	03. Februar 2004

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Statistik der Ludwig–Maximilians–Universität München. Darüber hinaus wurde diese Arbeit durch Mittel des Sonderforschungsbereiches 386 gefördert.

Mein Dank gebührt in erster Linie meinem langjährigen Chef und Doktorvater Gerhard Tutz, gegen dessen Motivationsversuche ich mich lange Zeit resistent gezeigt habe. Seiner Hartnäckigkeit und aktiven Unterstützung, aber auch meiner Einsicht ist es letztlich zu verdanken, dass diese Arbeit dennoch zustande gekommen ist.

Helmut Küchenhoff, der als Koreferent fungierte und Göran Kauermann, der sich als externer Gutachter dieser Arbeit angenommen hat, sei ebenfalls ausdrücklich gedankt.

Einen nichtwissenschaftlichen, aber nicht minder wesentlichen Anteil an dieser Arbeit tragen zweifelsohne meine Freunde. Stellvertretend seien hier Angelika Blauth, Eva–Maria Fronk, Rüdiger Krause und Stefan Lang genannt. Sie haben es wiederholt verstanden, mich in schwierigen Phasen wieder aufzubauen – Ihnen und Wessinol sei gedankt.

Herzlich gedankt sei auch all meinen Kollegen am Institut, deren wertvoller Beitrag in der angenehmen Gestaltung der Zeit zwischen dem Schreiben der einzelnen Kapitel bestand und kaum zu bemessen ist.

Mein ganz besonderer Dank gilt meiner Familie, deren uneingeschränkte Unterstützung in den vergangenen Jahren ein wichtiger Rückhalt für mich war.

Zusammenfassung

Mehrkategoriale Regressionsmodelle stellen ein etabliertes Instrumentarium in der statistischen Datenanalyse dar. Die vorliegende Arbeit behandelt die Erweiterung klassischer parametrischer Regressionsmodelle für nominal– und ordinal-kategoriale abhängige Variablen um flexible nonparametrische Strukturen. Unspezifiziert funktionale Effekte stetiger Kovariablen werden dabei mit Polynom-Splines approximiert, deren Repräsentation in entsprechenden Spline–Basen auf rein parametrische Prädiktorstrukturen führt. Die damit im Rahmen multivariater generalisierter linearer Modelle mögliche Parameterschätzung wird über die Maximierung einer penalisierten Likelihood realisiert, in der diskrete Strafterme die Variation der geschätzten funktionalen Effekte regulieren.

Einleitende theoretische Betrachtungen zu penalisierten Basisfunktionsansätzen liefern Aussagen zur Äquivalenz der untersuchten Alternativen und qualifizieren P–Splines als die in diesem Kontext zu präferierende Kombination. Darauf basierend werden nonparametrische Erweiterungen des multinomialen Logit–Modells für nominalen und des kumulativen Logit–Modells für ordinalen Response analysiert. Im multinomialen Logit–Modell wird dabei explizit zwischen globalen Variablen und kategorienspezifischen Charakteristiken unterschieden, wobei Einflußgrößen beider Typen sowohl linear als auch unspezifiziert funktional berücksichtigt werden. Für kumulative Logit–Modelle mit nicht-globalen Effekten wird das den P–Splines entlehnte Penalisierungskonzept auf die kategorienspezifischen Parameter in benachbarten Responsekategorien übertragen. Kategorienübergreifende Penalties gewährleisten einerseits die Verfügbarkeit von Schätzungen in numerisch kritischen Fällen und ermöglichen damit die Durchführung von Tests auf das Vorliegen proportionaler Chancen. Darüber hinaus lassen sie sich als konzeptionelle Bestandteile in die Parameterschätzung in semiparametrischen Partial Proportional Odds Modellen integrieren.

Summary

Multicategorical regression models are an established tool in statistical data analysis. The present thesis extends common parametric regression models for nominal and ordinal responses to more flexible nonparametric approaches. In order to obtain a flexible form of the functional effects of metrically scaled covariates, expansions in basis functions are used. The resulting predictor allows parameter estimation within the framework of multivariate generalized linear models. Estimates are obtained by maximizing a penalized likelihood with discrete penalty terms restricting the variation of estimated smooth effects.

As a result of theoretical considerations, P–Splines seem to be the ideal alternative for applying penalized basis function approaches. Based on this result, nonparametric extensions of the multinomial logit model for nominal and the cumulative logit model for ordinal responses are derived. An important feature of the proposed multinomial logit model is the distinction between global and category-specific variables. Variables of both types may enter the model in a linear form or as unspecified smooth functions. For cumulative logit models the penalization concept adopted from P–Splines is used to restrict category-specific parameters in adjacent categories. Penalization across response categories ensures availability of estimates when common estimation procedures fail to converge, so that tests for proportional odds may be performed even for critical settings. Additionally, penalties across response categories are taken into account as fixed methodical parts when fitting semiparametric partial proportional odds models.

Inhaltsverzeichnis

Einleitung

Lassen sachlogische Überlegungen einen gerichteten funktionalen Zusammenhang zwischen verschiedenen Merkmalen eines Untersuchungsobjekts vermuten, können diese Abhängigkeitsbeziehungen durch Regressionsmodelle in einen geeigneten statistischen Rahmen gefasst werden. Verwertbare Informationen liefern derartige Modelle jedoch nur dann, wenn sie auf die Charakteristika der Untersuchungsmerkmale zugeschnitten sind. Die adäquate Formulierung, Untersuchung und Interpretation von Regressionsmodellen bedarf daher einer expliziten Berücksichtigung der Merkmalstypisierungen für die abhängigen Größen als auch auf Seiten der erklärenden Variablen.

Diesem Grundsatz folgend, führten Fragestellungen aus den verschiedensten Bereichen zu der Notwendigkeit, das auf der Normalverteilungsannahme basierende, klassische Regressionsmodell auch auf den Fall diskreter Responsevariablen zu erweitern. Das dafür erforderliche statistische Konzept schufen Nelder & Wedderburn (1972) mit der Einführung generalisierter linearer Modelle. Besitzt der diskrete Response nur endlich viele Ausprägungen, spricht man von kategorialer Regression. Zahlreiche Publikationen belegen die breitgefächerten Anwendungsmöglichkeiten dieser Regressionsform in der Medizin (Kay & Little, 1986; Armstrong & Sloan, 1989), den Wirtschaftswissenschaften (Amemiya, 1981; Maddala, 1983) und vielen anderen Gebieten. Im Kontext kategorialer Regression beschränkt sich die Problemadäquatheit jedoch häufig auf die Berücksichtigung der diskreten Responsestruktur. Eigenschaften der erklärenden Variablen werden hingegen nur unzureichend reflektiert, da die oft vorausgesetzte Parametrisierbarkeit der Kovariableneffekte zumindest für stetige Einflußgrößen eine zu rigide Annahme darstellt. Ziel der vorliegenden Arbeit ist es, die statistischen Methoden der kategorialen Regression um ein flexibles Instrument zur Modellierung nonparametrischer Kovariableneffekte zu erweitern.

Für Modelle mit stetigen Regressoren ist in den letzten Jahren ein umfangreiches Repertoire an nonparametrischen Schätzverfahren geschaffen worden. Den wohl größten Anteil bilden die sogenannten Glättungsverfahren, die auf der schwachen Voraussetzung beruhen, daß der Effekt einer metrischen Ein-

flußgröße auf den zu erwartenden Response als glatte, sprich hinreichend oft differenzierbare Funktion dargestellt werden kann. Aus der Vielzahl von Methoden seien hier exemplarisch der Kernregressionschätzer (Gasser & Müller, 1984; Staniswalis, 1989), Regressionssplines (Eubank, 1988; Friedman & Silverman, 1989), lokale (polynomiale) Ansätze (Hastie & Loader, 1993; Fan & Gijbels, 1996) und Smoothing Splines (Silverman, 1985; Green, 1987) aufgeführt. Ein geschlossener Überblick über Glättungsverfahren findet sich u.a. in Simonoff (1996).

Die in der vorliegenden Arbeit gewählte Modellierungsform für glatte Effekte kombiniert die verschiedenen Ansätze der Spline Regression. Den Grundbaustein bildet die aus einer Restgliedabschätzung der Taylorentwicklung resultierende Approximierbarkeit glatter Funktionen durch stückweise polynomiale Funktionen, sogenannte Polynom- bzw. Regressionssplines (Seber & Wild, 1989). Parameter eines Polynom–Splines sind der Grad seiner polynomialen Stücke sowie die diese Stücke definierende Zerlegung seines Definitionsbereiches durch endlich viele Knoten. Die zu einer Zerlegung gehörigen Polynom-Splines bilden einen Vektorraum. Für jedes Element dieses Raumes existiert demnach eine entsprechende Basisdarstellung, die wiederum eine Parametrisierung der approximierbaren glatten Effekte ermöglicht. Die damit gegebene Rückführbarkeit auf rein parametrische Strukturen stellt den entscheidenden Vorteil dieses Ansatzes dar.

Die Güte der Approximationen durch Polynom-Splines wird maßgeblich von der Zerlegung bestimmt. Wachsende Knotenzahlen erhöhen zwar die Anpassung, führen bei stark verrauschten Daten aber unweigerlich zu unruhig verlaufenden Schätzungen. Probate Mittel zur Variationskontrolle der geschätzten Effekte sind datenadaptive Verfahren zur Wahl der Knotenzahl und -positionen (Friedman, 1991; Stone, Hansen, Kooperberg & Truong, 1997). Ein alternativer Ansatz benutzt das den Smoothing Splines entlehnte Penalisierungskonzept. Im Zusammenspiel mit der Approximation der glatten Effekte durch Polynom–Splines operieren Strafterme dabei auf den Koeffizienten der entsprechenden Basisdarstellungen. Zwei Penalisierungsformen haben sich in diesem Zusammenhang als effektive Ergänzungen zur Spline–Approximation erwiesen – der von Ruppert & Carroll (2000) in Verbindung mit der Trunca-

ted-Power-Basis propagierte Ridge–Penalty und der Differenzenpenalty, den Eilers & Marx (1996) zusammen mit der B-Spline-Basis verwenden und dafür den Begriff des P–Splines prägen. Die Kombination der Spline–Approximation mit diesen diskreten Penalisierungsformen gestaltet sich in zweierlei Hinsicht als vorteilhaft. Einerseits kann die Parameterschätzung über die Maximierung einer penalisierten Likelihoodfunktion im Rahmen generalisierter linearer Modelle erfolgen, zum anderen wird das (hochdimensionale) Optimierungsproblem der Knotenwahl auf die Bestimmung nur eines Glättungsparameters pro nonparametrischer Komponente reduziert.

In der vorliegenden Arbeit soll das Konzept penalisierter Basisfunktionen auf die Besonderheiten im kategorialen Regressionsmodell erweitert werden. Das erste Kapitel gibt einen informellen Überblick über generalisierte lineare Modelle und den Prädiktor betreffende Verallgemeinerungsformen, die eine flexiblere Modellierung von Kovariableneffekten ermöglichen. Im zweiten Kapitel werden Basisdarstellungen von Polynom–Splines ausführlich untersucht. Neben der Definition von Truncated–Power- und B–Spline-Basis liefert das Kapitel Aussagen zur Äquivalenz dieser Basen. Darüber hinaus werden in einem weiteren Abschnitt die Penalisierungsformen Ridge- und Differenzenpenalty vorgestellt und deren Zusammenspiel mit den korrespondierenden Spline-Basen analysiert. Das Kapitel schließt mit einigen Ausführungen zu Eigenschaften von P-Splines, die sich als zu präferierende Kombination von Basisfunktionen und Penalties qualifizieren.

Kapitel 3 erläutert die Einsatzmöglichkeiten von penalisierten B–Splines zur Schätzung verschiedener nonparametrischer Komponenten im generalisierten additiven Modell. Ausgehend von der Modellierung univariater glatter Effekte wird die Berücksichtigung von Interaktionen in Form variierender Koeffizienten und bivariater Kovariablenfunktionen diskutiert. Neben der Umsetzung und Schätzung der einzelnen Komponenten steht die Behandlung von Identifikations- und Singularitätsproblemen im Mittelpunkt der Betrachtungen. Das vierte Kapitel ist der Wahl der Glättungsparameter gewidmet, die einen wichtigen konzeptionellen Bestandteil im Kontext penalisierter Schätzung darstellen. Optimale Glättungsparameter resultieren dabei aus der Minimierung eines entsprechenden Kriteriums, dessen Eignung zunächst moti-

viert wird. Darüber hinaus wird mit den genetischen Algorithmen ein praktikables Verfahren für die numerische Optimierung der Glättungsparameter vorgestellt.

Problemadäquatheit umfasst im kategorialen Regressionsmodell insbesondere die angemessene Berücksichtigung des Skalenniveaus der abhängigen Variablen. In diesem Sinne wird in der vorliegenden Arbeit explizit zwischen kategorial-nominalem sowie kategorial-ordinalem Response unterschieden. Kapitel 5 thematisiert Modelle für abhängige Variablen mit nominalem Skalenniveau. Einführende Kommentare zum prinzipiellen Vorgehen schaffen dabei die Grundlage für die anschließend ausgeführten nonparametrischen Erweiterungen. Die auf dem P–Spline Konzept beruhende Schätzung glatter Effekte von globalen Variablen und kategorienspezifischen Charakteristiken wird anhand eines Datenbeispiels bzw. in einer Simulationsstudie demonstriert.

McCullagh's kumulatives Logit–Modell stellt die wohl gebräuchlichste Form der Modellierung ordinaler Responsevariablen dar. Numerische Probleme bei der Schätzung kategorienspezifischer Parameter beschränken die Datenanalyse dabei häufig auf das Proportional Odds Modell. Kapitel 6 zeigt, wie kategorienübergreifende Strafterme in Modellen mit nicht-globalen Effekten stabilitätsfördernd wirksam werden. Die dadurch auch in kritischen Situationen gewährleistete Konvergenz iterativer Schätzverfahren ermöglicht die Anwendung von Teststatistiken, die auf kategorienspezifischen Parameterschätzern basieren. Untersucht werden in diesem Zusammenhang vornehmlich Signifikanztests auf das Vorliegen identischer Chancenverhältnisse. Nonparametrische Erweiterungen des kumulativen Logit–Modells bilden den Schwerpunkt der Betrachtungen in Kapitel 7. Das Hauptaugenmerk richtet sich dabei auf die Kombination von P–Splines und kategorienübergreifender Penalisierung. Darüber hinaus wird eine parameterökonomische Form der Modellierung kategorienspezifischer glatter Effekte vorgestellt.

Die Arbeit schließt mit einem Anhang, in den ein Großteil der zu führenden Beweise ausgelagert wurde. Ferner enthält dieser letzte Abschnitt kurze Anmerkungen zur verwendeten Software und gibt einen Überblick über die den Ausführungen zugrunde liegenden notationellen Vereinbarungen.

1 Generalisierte Modelle

Mit dem klassischen linearen Regressionsmodell ist eine einfache und effektive Möglichkeit gegeben, lineare Abhängigkeitsbeziehungen zwischen einer Menge von erklärenden Variablen und einer (zumindest approximativ) normalverteilten Zielgröße zu beschreiben. Die direkte, mit geringem numerischen Aufwand verbundene Möglichkeit, Parameterschätzungen zu erhalten, sowie die einfache Interpretierbarkeit der Ergebnisse machen das klassische lineare Regressionsmodell zu einem attraktiven und häufig angewandten Ansatz in der Analyse gerichteter Abhängigkeitsbeziehungen.

Für die Darstellung komplexerer, nicht-linearer Zusammenhänge und in Situationen mit diskreter Responsestruktur erweist sich das lineare Regressionsmodell jedoch als ungeeignet. Insbesondere die restriktive Verteilungsannahme für den Response geht dabei zu Lasten einer allgemeineren Gültigkeit des Modells.

Mit der Einführung generalisierter linearer Modelle durch Nelder & Wedderburn (1972) wurde erstmals ein einheitliches Werkzeug zur Behandlung nichtnormalverteilter Responsevariablen geschaffen. Zwar bleibt im Kontext dieser Modelle die grundlegende strukturelle Annahme linearer Zusammenhänge erhalten, die stark einschränkende Normalverteilungsannahme wird jedoch zugunsten einer allgemeiner gefassten Zugehörigkeit der Responsedichte zu einer exponentiellen Familie aufgegeben.

Generalisierte lineare Modelle bildeten die Basis für weiterführende Verallgemeinerungen. Insbesondere die von Hastie & Tibshirani (1990) propagierten generalisierten additiven Modelle (GAM's) sowie die Modelle mit variierenden Koeffizienten (VCM's) (Hastie & Tibshirani, 1993) waren weitere entscheidende Schritte auf dem Weg zu einer höheren Modellflexibilität.

Die folgenden Abschnitte sind einem informativen Überblick über die genannten Modellierungsformen gewidmet. Die Darstellungen beschränken sich dabei auf die grundlegenden Definitionen und Problemstellungen sowie die Behandlung der jeweiligen Modelle im Regressionskontext. Für weitergehende Ausführungen sei auf die zitierten Quellen verwiesen.

1.1 Generalisierte lineare Modelle

Es bezeichne y eine univariate Zufallsgröße (Response) und $\boldsymbol{x} := (x_1, \ldots, x_p)'$ den Vektor der Kovariablen (Einflußgrößen). Im folgenden wird in der Notation nicht explizit zwischen stochastischen und deterministischen Größen unterschieden. Der Charakter einer Variablen geht aus dem jeweiligen Kontext hervor. Ausgehend von den individuenspezifischen Paaren $(y_i, \boldsymbol{x}_i)_{i=1,\ldots,N}$ liegt dem klassischen linearen Ansatz folgende Modellannahme zugrunde

$$y_i \;=\; \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \;\; i = 1, \ldots, N,$$

mit unabhängigen und identisch verteilten Störtermen $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Vereinbart man $\boldsymbol{z}_i := (1, \boldsymbol{x}_i')'$ und $\boldsymbol{\beta} := (\beta_0, \ldots, \beta_p)'$, so sind die y_i (bedingt) unabhängig und identisch verteilt gemäß

$$y_i | \boldsymbol{x}_i \sim \mathcal{N}(\mu_i, \sigma^2), \;\; i = 1, \ldots, N, \tag{1.1}$$

wobei

$$\mu_i := E(y_i | \boldsymbol{x}_i) = \boldsymbol{z}_i' \boldsymbol{\beta}. \tag{1.2}$$

Die bedingten Schreibweisen und Formulierungen sind lediglich beim Vorliegen stochastischer Regressoren relevant. Nachstehende Verallgemeinerungen von (1.1) und (1.2) führen nun in natürlicher Weise zur Definition generalisierter linearer Modelle (GLM's) (vgl. Fahrmeir & Tutz (2001))

(i) Die y_i sind (bedingt) unabhängig mit (bedingten) Dichten

$$f(y_i \,|\, \theta_i, \phi, \omega_i) \;=\; \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi}\, \omega_i + c(y_i, \phi, \omega_i) \right\} \tag{1.3}$$

aus einer einfachen Exponentialfamilie und (bedingten) Erwartungswerten $E(y_i | \boldsymbol{x}_i) = \mu_i$. Die Größe $\theta_i = \theta(\mu_i)$ bezeichnet den sogenannten natürlichen Parameter, ϕ ist ein Dispersionparameter, die Funktionen $b(.)$, $c(.)$ sind abhängig vom konkreten Typ der Exponentialfamilie, und die ω_i stellen Gewichte für den Fall gruppierter Daten dar.

(ii) Der Kovariablenvektor $\boldsymbol{x}_i$ beeinflußt den Response y_i über den linearen Prädiktor $\eta_i = \boldsymbol{z}_i' \boldsymbol{\beta}$.

(iii) Der lineare Prädiktor η_i wird vermöge der sogenannten Responsefunktion $h : I\!R \longrightarrow I\!R$ zum Erwartungswert μ_i in Beziehung gesetzt

$$\mu_i = h(\eta_i) = h(\boldsymbol{z}_i'\boldsymbol{\beta}).$$

Im Falle ihrer Existenz wird die inverse Funktion $g = h^{-1}$ als Linkfunktion bezeichnet, so daß $\eta_i = g(\mu_i)$.

Die Erwartungswerte μ_i sind über die Beziehung $\mu_i = b'(\theta_i) = \partial b(\theta_i)/\partial\theta_i$ eindeutig durch den Typ der Exponentialfamilie bestimmt. Ferner gilt

$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = \sigma^2(\mu_i) = \phi v(\mu_i)/\omega_i$$

mit der ebenfalls durch den Exponentialfamilientyp bestimmten Varianzfunktion $v(\mu_i) = \partial^2 b(\theta_i)/\partial\theta_i^2$. Zugehörigkeitsnachweise ausgewählter Verteilungstypen zu einer Exponentialfamilie werden in Fahrmeir & Tutz (2001) geführt. Als Beispiel sei an dieser Stelle lediglich die $\mathcal{N}(\mu_i, \sigma^2)$-verteilung erwähnt, deren Dichte mit $\theta_i = \mu_i$, $\phi = \sigma^2$, $w_i = 1$, $c(y_i, \phi, \omega_i) = -y_i^2/2\phi - \log\sqrt{2\pi\phi}$ und $b(\theta_i) = \theta_i^2/2$ in der Form (1.3) geschrieben werden kann. Setzt man $g(\mu_i) = \mu_i$ resultiert daraus das klassische lineare Regressionsmodell als spezielles GLM.

1.1.1 Maximum–Likelihood–Schätzung

Die nachstehenden Ausführungen geben einen kurzen Einblick in die Methodik der Maximum–Likelihood–Schätzung im GLM. Ausgehend von der Log–Likelihood $l(\theta_1, \ldots, \theta_N) = \sum_i l_i(\theta_i)$ gilt für den Log–Likelihood Beitrag $l_i(\theta_i)$ der i-ten Beobachtung gemäß Verteilungsannahme (i) folgende Proportionalitätsaussage

$$l_i(\theta_i) \;=\; \log f(y_i|\theta_i, \phi, \omega_i) \;\propto\; \omega_i\phi^{-1}\big(y_i\theta_i - b(\theta_i)\big)$$

bzw. unter Berücksichtigung von $\theta_i = \theta(\mu_i) = \theta\big(h(\boldsymbol{z}_i'\boldsymbol{\beta})\big)$

$$l_i(\boldsymbol{\beta}) \;\propto\; \omega_i\phi^{-1}\Big\{y_i\theta\big(h(\boldsymbol{z}_i'\boldsymbol{\beta})\big) - b\big[\theta\big(h(\boldsymbol{z}_i'\boldsymbol{\beta})\big)\big]\Big\}$$

für die in Abhängigkeit vom Parametervektor $\boldsymbol{\beta}$ formulierten Log–Likelihood Beiträge $l_i(\boldsymbol{\beta})$ der Log–Likelihood $l(\boldsymbol{\beta}) = \sum_i l_i(\boldsymbol{\beta})$.

Maximum–Likelihood–Schätzer für den unbekannten Parametervektor $\boldsymbol{\beta}$ re-

sultieren dann aus den $p+1$ Schätzgleichungen $s(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$, wobei die Funktion $s(\boldsymbol{\beta})$ als Score–Funktion bezeichnet wird. Dabei ist

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \omega_i \phi^{-1} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial h(\eta_i)}{\partial \eta_i} \left[y_i - \frac{\partial b(\theta_i)}{\partial \theta_i} \right] \frac{\partial \boldsymbol{z}_i' \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}.$$

Mit $\partial \boldsymbol{z}_i'\boldsymbol{\beta}/\partial\boldsymbol{\beta} = \boldsymbol{z}_i$,

$$\frac{\partial}{\partial \theta_i}\mu_i = \frac{\partial^2}{\partial \theta_i^2} b(\theta_i) = v(\mu_i) = v\big(h(\boldsymbol{z}_i'\boldsymbol{\beta})\big), \quad \frac{\partial h(\eta_i)}{\partial \eta_i} = \frac{\partial h(\boldsymbol{z}_i'\boldsymbol{\beta})}{\partial \eta_i} =: D_i(\boldsymbol{\beta}),$$

und $\sigma_i^2(\boldsymbol{\beta}) := v\big(h(\boldsymbol{z}_i'\boldsymbol{\beta})\big)\phi\omega_i^{-1}$ ergibt sich die Score–Funktion zu

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{z}_i \, \sigma_i^{-2}(\boldsymbol{\beta}) \, D_i(\boldsymbol{\beta}) \, (y_i - \mu_i).$$

Mit den Deklarationen $\boldsymbol{y} := (y_1, \ldots, y_N)'$, $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_N)'$,

$\operatorname{cov}(\boldsymbol{y}) = \operatorname{diag}\big(\sigma_1^2(\boldsymbol{\beta}), \ldots, \sigma_N^2(\boldsymbol{\beta})\big) =: \Sigma(\boldsymbol{\beta})$, $D(\boldsymbol{\beta}) := \operatorname{diag}\big(D_1(\boldsymbol{\beta}), \ldots, D_N(\boldsymbol{\beta})\big)$

und der Designmatrix $Z := \big[\, \boldsymbol{z}_1 \,|\, \ldots \,|\, \boldsymbol{z}_N \,\big]'$ läßt sich die Darstellung der Score-Funktion auch in eine Matrizenschreibweise überführen

$$s(\boldsymbol{\beta}) = Z' D(\boldsymbol{\beta}) \, \Sigma(\boldsymbol{\beta})^{-1} (\boldsymbol{y} - \boldsymbol{\mu}). \tag{1.4}$$

Die Schätzgleichungen $s(\boldsymbol{\beta}) = \mathbf{0}$ sind im allgemeinen nicht-linear und können nur iterativ gelöst werden. Ein in solchen Fällen häufig angewandtes Lösungsverfahren ist der Newton-Raphson-Algorithmus. Ausgehend von einem Startvektor $\hat{\boldsymbol{\beta}}^{(0)}$ werden neue Iterierte in der vorliegenden Situation über die Vorschrift

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - \left[\frac{\partial}{\partial \boldsymbol{\beta}'} s\Big(\hat{\boldsymbol{\beta}}^{(k)}\Big) \right]^{-1} \cdot s\Big(\hat{\boldsymbol{\beta}}^{(k)}\Big), \quad k = 0, 1, \ldots, \tag{1.5}$$

berechnet, bis ein vordefiniertes Abbruchkriterium diese Iterationen beendet. Die Matrix $-\partial s(\boldsymbol{\beta})/\partial\boldsymbol{\beta}' = -\partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'$ ist die beobachtete Fisher'sche Informationsmatrix und wird mit $F_{obs}(\boldsymbol{\beta})$ bezeichnet.

Ersetzt man in (1.5) F_{obs} durch die erwartete Fisher'sche Informationsmatrix $F = E[F_{obs}]$, erhält man das sogenannte Fisher–Scoring als Modifikation des Newton-Raphson-Verfahrens. Mit Bartlett's erster und zweiter Identität ist

$$F(\boldsymbol{\beta}) = E\big[F_{obs}(\boldsymbol{\beta})\big] = E\left[-\frac{\partial s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right] = E\big[s(\boldsymbol{\beta})\, s'(\boldsymbol{\beta})\big] = \operatorname{cov}\big(s(\boldsymbol{\beta})\big),$$

wobei mit (1.4) und $\mathrm{cov}(\boldsymbol{y}) = \Sigma(\boldsymbol{\beta})$ weiterhin folgt

$$\mathrm{cov}\big(s(\boldsymbol{\beta})\big) = Z' D(\boldsymbol{\beta})\, \Sigma(\boldsymbol{\beta})^{-1} \mathrm{cov}(\boldsymbol{y})\, \Sigma(\boldsymbol{\beta})^{-1}\, D'(\boldsymbol{\beta}) Z = Z'\, W(\boldsymbol{\beta})\, Z,$$

für $W(\boldsymbol{\beta}) := D(\boldsymbol{\beta})\, \Sigma(\boldsymbol{\beta})^{-1} D'(\boldsymbol{\beta})$.

Da die beobachtete Fisher'sche Informationsmatrix $F_{obs}(\boldsymbol{\beta})$ nur für natürliche Linkfunktionen deterministisch ist, findet das Fisher-Scoring in der Praxis eine breitere Anwendung als das Newton-Raphson-Verfahren.

Eine abschließende Formalisierung beschreibt das Fisher-Scoring als iterative gewichtete Kleinste-Quadrate-Schätzung (Fahrmeir & Tutz, 2001)

Step 1: Initialisiere $k = 0$ und $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$.

Step 2: Berechne die *adjustierten Prädiktoren*

$$\tilde{\boldsymbol{\eta}}^{(k)} = \big(\tilde{\eta}_1^{(k)}, \ldots, \tilde{\eta}_N^{(k)}\big), \text{ mit } \tilde{\eta}_i^{(k)} = \boldsymbol{z}_i' \hat{\boldsymbol{\beta}}^{(k)} + D_i\Big(\hat{\boldsymbol{\beta}}^{(k)}\Big)^{-1}\big(y_i - \mu_i^{(k)}\big).$$

Step 3: Setze $\hat{\boldsymbol{\beta}}^{(k+1)} = F\Big(\hat{\boldsymbol{\beta}}^{(k)}\Big)^{-1} Z'\, W\Big(\hat{\boldsymbol{\beta}}^{(k)}\Big)\, \tilde{\boldsymbol{\eta}}^{(k)}$.

Step 4: Berechne $\Delta = ||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}|| \,/\, ||\hat{\boldsymbol{\beta}}^{(k)}||$.

Step 5: Falls $\Delta \leq \varepsilon$ für ein $\varepsilon > 0$, beende den Algorithmus. Andernfalls setze $k := k + 1$ und gehe zu *Step 2*.

1.2 Generalisierte additive Modelle

Eine weitere Verallgemeinerungsstufe des klassischen linearen Regressionsmodells wendet sich der Prädiktorstruktur zu. Hastie & Tibshirani (1990) untersuchen generalisierte additive Modelle (GAM's), in denen die bis dato unterstellte lineare Parametrisierbarkeit jedes Kovariableneffekts unspezifizierten, funktionalen Zusammenhängen weicht

$$\eta_i = \beta_0 + f_{(1)}(x_{i1}) + \ldots + f_{(p)}(x_{ip}), \;\; i = 1, \ldots, N. \tag{1.6}$$

Zur Reduktion der Dimension und aus Gründen einer besseren Interpretierbarkeit wird in diesem Modell die Annahme eines additiven Zusammenwir-

kens der einzelnen Effekte aufrecht erhalten. Ansätze, in denen die Kovariableneinflüsse via

$$\eta_i = \beta_0 + f(x_{i1}, \ldots, x_{ip})$$

in multivariat funktionaler Form im Prädiktor Berücksichtigung finden, werden u.a. in O'Sullivan, Yandell & Raynor (1986), Friedman (1991) und Ruppert & Wand (1994) vorgestellt. Spezifiziert man in (1.6) ausgewählte Kovariableneffekte als linear, $f_{(j)}(x_{ij}) := \beta_j x_{ij}$, $j \in \mathcal{D} \subset \{1, \ldots, p\}$, so resultieren semiparametrische Modelle als spezielle GAM's (Green & Yandell, 1985) .

Aus der Vielzahl an Methoden zur Behandlung nonparametrischer Regressionsmodelle der Form (1.6) werden hier lediglich *Glättungssplines* (Smoothing Splines) kurz diskutiert. Ausgangspunkt bildet dabei die Betrachtung der penalisierten Log–Likelihood

$$pl(\boldsymbol{\eta}) \;=\; pl(\eta_1, \ldots, \eta_N) \;=\; \sum_{i=1}^{N} l_i(\eta_i) \;-\; \frac{1}{2} \cdot \sum_{j=1}^{p} \lambda_j \int \big(f''_{(j)}(u)\big)^2 \, du, \qquad (1.7)$$

in der die Log–Likelihood der Stichprobe um zusätzliche Strafterme (Penalties) ergänzt wird. Das Integral der quadrierten zweiten Ableitung als globale Maßzahl für die Krümmung einer Funktion dient der Kontrolle der Variabilität in den zu schätzenden Regressionsfunktionen. Die Penalties in (1.7) definieren mit der vorausgesetzten Existenz und Quadratintegrierbarkeit der Ableitungen $f''_{(j)}$ gleichzeitig die Minimalanforderungen für die unbekannten Regressionsfunktionen.

Die Glättungsparameter $\lambda_j \geq 0$ regulieren den Einfluss der Strafterme. Sehr kleine Werte von λ_j reduzieren die Wirkung des korrespondierenden Penalties auf ein Minimum und führen (im Extremfall) zu einer Interpolation der Datenpunkte. Wachsende Werte von λ_j korrespondieren hingegen mit einer zunehmenden Glattheit des penalisierten Maximum–Likelihood–Schätzers von $f_{(j)}$. Für $\lambda_j \longrightarrow +\infty$ liegen die geschätzten Werte $\hat{f}_{(j)}(x_{1j}), \ldots, \hat{f}_{(j)}(x_{Nj})$ auf einer Geraden.

Die Maximierung von (1.7) führt auf sogenannte kubische Smoothing Splines als Lösung für die unbekannten Regressionsfunktionen (Reinsch, 1967). Mit den Vereinbarungen $\boldsymbol{f}_j = \big(f_{(j)}(x_{1j}), \ldots, f_{(j)}(x_{Nj})\big)'$, $j = 1, \ldots, p$, ist die Maxi-

mierung von (1.7) äquivalent zum Optimierungsproblem

$$pl(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_p) \;=\; \sum_{i=1}^{N} l_i(\eta_i) - \frac{1}{2}\cdot\sum_{j=1}^{p}\lambda_j\,\boldsymbol{f}_j' K_j \boldsymbol{f}_j \;\longrightarrow\; \max\,, \qquad (1.8)$$

mit speziell strukturierten Strafmatrizen K_j (Fahrmeir & Tutz, 2001). Bevor wir uns der Formulierung eines konkreten Maximierungsschemas für (1.8) zuwenden, sei noch auf die folgende Problematik hingewiesen: In additiven Modellen sind die Kovariableneffekte ohne zusätzliche Restriktionen nicht identifizierbar. Wie sich unmittelbar aus (1.6) erschließt, resultiert diese Nichtidentifizierbarkeit aus der möglichen Verschiebung additiver Konstanten. Die explizite Berücksichtigung bzw. Vermeidung dieses Problems läßt sich durch eine geeignete Zentrierung der geschätzten Kovariableneffekte realisieren.

Zur Ableitung eines Optimierungsverfahrens für (1.8) wird die Konstante β_0 (zunächst) außer Acht gelassen, d.h. $\boldsymbol{\eta} := \boldsymbol{\eta}(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_p)$. Unter Beibehaltung der bereits in Abschnitt 1.1 für GLM's formulierten Annahmen (i) und (iii) ergeben sich die partiellen penalisierten Score–Funktionen gemäß (1.8) zu

$$\boldsymbol{ps}_j(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_p) \;=\; \frac{\partial pl(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_p)}{\partial \boldsymbol{f}_j} \;=\; \boldsymbol{s} - \lambda_j K_j \boldsymbol{f}_j\,,\quad j=1,\ldots,p,$$

mit

$$\boldsymbol{s} \;=\; \sum_{i=1}^{N} \boldsymbol{e}_{i,N}\, w_i D_i(\eta_i)^{-1}(y_i-\mu_i)\,,$$

wobei

$$w_i \;:=\; \sigma^{-2}(\mu_i)\,D_i(\eta_i)^2 \quad\text{und}\quad D_i(\eta_i) \;=\; \frac{\partial h(\eta_i)}{\partial \eta_i}\,.$$

Eine Lösung der Np Schätzgleichungen $\boldsymbol{ps}_j(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_p)=\boldsymbol{0}$, $j=1,\ldots,p$, läßt sich wiederum nur iterativ via Fisher–Scoring gewinnen. Dazu definieren wir

$$W^{(k)} \;:=\; \mathrm{diag}\big(w_1^{(k)},\ldots,w_N^{(k)}\big), \qquad \tilde{\boldsymbol{\eta}}^{(k)} \;:=\; \boldsymbol{\eta}^{(k)} + \big(W^{(k)}\big)^{-1}\boldsymbol{s}^{(k)},$$

$$\text{und}\;\; S_j^{(k)} \;:=\; \big(W^{(k)} + \lambda_j K_j\big)^{-1} W^{(k)}.$$

Die Indizierung kennzeichnet dabei eine Auswertung der Größen an den Iterierten $\boldsymbol{\eta}^{(k)} = \boldsymbol{\eta}\big(\hat{\boldsymbol{f}}_1^{(k)},\ldots,\hat{\boldsymbol{f}}_p^{(k)}\big)$, $k=0,1,\ldots$. Mit obigen Festlegungen lassen sich die Fisher–Scoring Iterationen nun schreiben als

$$\begin{bmatrix} I & S_1^{(k)} & \dots & S_1^{(k)} \\ S_2^{(k)} & I & \dots & S_2^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ S_p^{(k)} & S_p^{(k)} & \dots & I \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{f}}_1^{(k+1)} \\ \hat{\boldsymbol{f}}_2^{(k+1)} \\ \vdots \\ \hat{\boldsymbol{f}}_p^{(k+1)} \end{bmatrix} = \begin{bmatrix} S_1^{(k)}\tilde{\boldsymbol{\eta}}^{(k)} \\ S_2^{(k)}\tilde{\boldsymbol{\eta}}^{(k)} \\ \vdots \\ S_p^{(k)}\tilde{\boldsymbol{\eta}}^{(k)} \end{bmatrix}.$$

Dieses System läßt sich in vielen Fällen nicht direkt lösen, es motiviert aber ein als Backfitting bekanntes iteratives Verfahren zur Bestimmung der Folge-Iterierten $\hat{\boldsymbol{f}}^{(k+1)}$. Die Einbettung einer zusätzlichen Backfitting–Schleife gestattet dann die Formulierung der folgenden Fisher–Scoring–Schritte:

Step 1: Initialisiere $k = 0$, $\hat{\beta}_0^{(0)} = 0$ und $\hat{\boldsymbol{f}}_1^{(0)} = \ldots = \hat{\boldsymbol{f}}_p^{(0)} = \mathbf{0}$.

Step 2: Berechne die *adjustierten Prädiktoren* $\tilde{\boldsymbol{\eta}}^{(k)} = \big(\tilde{\eta}_1^{(k)}, \ldots, \tilde{\eta}_N^{(k)}\big)'$, mit $\tilde{\eta}_i^{(k)} = \eta_i^{(k)} + D_i\big(\eta_i^{(k)}\big)^{-1}\big(y_i - \mu_i^{(k)}\big)$ und $\eta_i^{(k)} = \hat{\beta}_0^{(k)} + \sum_{j=1}^p \hat{f}_{(j)}^{(k)}(x_{ij})$.

Step 3: Berechne $\hat{\boldsymbol{f}}_1^{(k+1)}, \ldots, \hat{\boldsymbol{f}}_p^{(k+1)}$ via Backfitting:

Step 3.1: Initialisiere $\boldsymbol{\beta} = \mathbf{1}_N\hat{\beta}_0^{(k)}$ und $\boldsymbol{f}_j^0 = \hat{\boldsymbol{f}}_j^{(k)}$, $j = 1, \ldots, p$.

Step 3.2: Berechne Updates $\boldsymbol{f}_j^0 \longrightarrow \boldsymbol{f}_j^1$, $j = 1, \ldots, p$, via

$$\boldsymbol{f}_j^1 = S_j^{(k)}\Big(\tilde{\boldsymbol{\eta}}^{(k)} - \boldsymbol{\beta} - \sum_{l<j}\boldsymbol{f}_l^1 - \sum_{l>j}\boldsymbol{f}_l^0\Big),$$

$$\boldsymbol{f}_j^1 = \boldsymbol{f}_j^1 - \frac{1}{N}\sum_{i=1}^N f_{(j)}^1(x_{ij}) = \boldsymbol{f}_j^1 - \bar{\boldsymbol{f}}_j^1,\ \boldsymbol{\beta} = \boldsymbol{\beta} + \mathbf{1}_N\bar{\boldsymbol{f}}_j^1$$

Step 3.3: Berechne $\Delta_j = \big|\big|\boldsymbol{f}_j^0 - \boldsymbol{f}_j^1\big|\big|$, $j = 1, \ldots, p$. Falls $\Delta_j \leq \varepsilon$ für ein $\varepsilon > 0$ und alle j, setze $\hat{\boldsymbol{f}}_j^{(k+1)} := \boldsymbol{f}_j^1$, $\hat{\beta}_0^{(k+1)} := \boldsymbol{e}_{1,N}'\boldsymbol{\beta}$ und beende Backfitting. Andernfalls setze $\boldsymbol{f}_j^0 := \boldsymbol{f}_j^1$ und gehe zu *Step 3.2*.

Step 4: Berechne $\tilde{\Delta} = \sum_{j=1}^p \big|\big|\hat{\boldsymbol{f}}_j^{(k+1)} - \hat{\boldsymbol{f}}_j^{(k)}\big|\big| \,/\, \sum_{j=1}^p \big|\big|\hat{\boldsymbol{f}}_j^{(k)}\big|\big|$. Falls $\tilde{\Delta} \leq \tilde{\varepsilon}$ für ein $\tilde{\varepsilon} > 0$, beende Fisher-Scoring. Andernfalls setze $k := k + 1$ und gehe zu *Step 2*.

Die Konstante β_0 dient also lediglich dem Auffangen der beim Zentrieren entstandenen Skalierungsterme. Wegen der Zentriertheit der funktionalen Kom-

ponenten läßt sich $\hat{\beta}_0$ als mittlerer Effekt derjenigen Kovariablenkombination auffassen, für die $\hat{f}_{(1)}(x_1) = \ldots = \hat{f}_{(p)}(x_p) = 0$ gilt.

Bislang unbeantwortet geblieben ist die Frage nach der Wahl der Glättungsparameter λ_j, $j = 1, \ldots, p$. Der Behandlung dieser Fragestellung ist ein späteres Kapitel der vorliegenden Arbeit gewidmet. Die dort getroffenen Aussagen erfolgen zwar im Kontext einer alternativen Modellierungsform, sind aber hinsichtlich Kriterienwahl und möglicher Optimierungsstrategien auf die hier vorgestellten Glättungssplines übertragbar.

1.3 Modelle mit variierenden Koeffizienten

Die ebenfalls von Hastie & Tibshirani (1993) vorgestellten Modelle mit variierenden Koeffizienten (VCM's) modifizieren die Prädiktorstrukur des generalisierten linearen Modells durch eine Berücksichtigung spezieller Interaktionsterme. Ein variierendes Koeffizientenmodell ist nach wie vor linear in den Regressoren, deren Koeffizienten können sich aber in Abhängigkeit weiterer Kovariablen ändern

$$\eta_i \;=\; \beta_0 + f_{(1)}(w_{i1})\, x_{i1} + \ldots + f_{(p)}(w_{ip})\, x_{ip}\,, \;\; i = 1, \ldots, N. \qquad (1.9)$$

Generalisierte lineare bzw. additive Modelle folgen daraus als Spezialfälle mit $f_{(j)} \equiv \beta_j$ bzw. $x_{ij} \equiv 1$, $j = 1, \ldots, p$. Eine weitere Abwandlung des obigen Modells resultiert aus $w_{ij} \equiv w_i$ für $j \in \{1, \ldots, p\}$. In Erweiterung von (1.9) sind Mischformen denkbar, in denen die variierenden Koeffizienten $f_{(j)}(w_{ij})$ auf fixen Einflüssen β_j, $j = 1, \ldots, p$, aufsetzen, und die Effektmodifizierer w_{ij} über zusätzliche Haupteffekte $g_{(j)}(w_{ij})$, $j = 1, \ldots, p$, berücksichtigt werden

$$\eta_i \;=\; \beta_0 \,+\, \boldsymbol{x}_i'\boldsymbol{\beta} \,+\, \sum_{j=1}^{p} g_{(j)}(w_{ij}) \,+\, \sum_{j=1}^{p} f_{(j)}(w_{ij})\, x_{ij}\,, \;\; i = 1, \ldots, N. \qquad (1.10)$$

In diesem Ansatz treten jedoch, neben den bereits für GAM's beschriebenen, weitere Identifikationsprobleme auf. So können auch zwischen dem fixen und dem variierenden Effekt einer Kovariablen additive Konstanten ausgetauscht werden, ohne den Wert des Prädiktors zu verändern.

Die Schätzung der unbekannten Regressionsfunktionen in (1.9) kann wieder über die Maximierung einer penalisierten Log–Likelihood

$$pl(\boldsymbol{\eta}) \;=\; \sum_{i=1}^{N} l_i(\eta_i) \;-\; \frac{1}{2} \cdot \sum_{j=1}^{p} \lambda_j \int \left(f''_{(j)}(u)\right)^2 du \tag{1.11}$$

erfolgen. Einzig bei der partiellen Differentiation der Log–Likelihood Beiträge ist die veränderte Prädiktorstruktur zu beachten. Im Unterschied zu dem in Abschnitt 1.2 vorgestellten Verfahren sind die partiellen Score–Funktionen

$$\boldsymbol{s}_j \;=\; \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{f}_j}\, l_i(\eta_i) \;=\; \sum_{i=1}^{N} \boldsymbol{e}_{i,N}\, x_{ij}\, \sigma^{-2}(\mu_i)\, D_i(\eta_i)\, (y_i - \mu_i)\,, \;\; j = 1, \ldots, p,$$

jetzt für jedes $j \in \{1, \ldots, p\}$ verschieden. Hastie & Tibshirani (1993) zeigen, daß die Maximierung von (1.11), ähnlich wie in 1.2, iterativ über ein System von Normalengleichungen ausgeführt werden kann.

1.4 Surface smoother

Wie bereits erwähnt, kann der Einfluß mehrerer Kovariablen auch über funktionale Interaktionen mehrerer Veränderlicher modelliert werden

$$\eta_i \;=\; \beta_0 \;+\; f(x_{i1}, \ldots, x_{ip})\,, \;\; i = 1, \ldots, N, \tag{1.12}$$

wobei $f : I\!R^p \longrightarrow I\!R$. Die Erweiterung kubischer Smoothing Splines auf diesen mehrdimensionalen Fall wird in den Arbeiten von O'Sullivan, Yandell & Raynor (1986) und Gu (1990) thematisiert. Die durch eine Aufgabe der additiven Struktur der Prädiktoren η_i geschaffene Verallgemeinerung von GAM's geht jedoch zu Lasten einer überschaubaren Modellkomplexität. Aus diesem Grund beschränkt man sich in vielen Situationen auf die Modellierung zweidimensionaler Interaktionen, wahlweise unter Hinzufügen der entsprechenden Haupteffekte

$$\eta_i \;=\; \beta_0 \;+\; f_{(1)}(x_{i1}) \;+\; f_{(2)}(x_{i2}) \;+\; f_{(12)}(x_{i1}, x_{i2})\,, \;\; i = 1, \ldots, N. \tag{1.13}$$

Damit bleibt insbesondere die Visualisierung der geschätzten Effekte möglich.

Wie im Kapitel 3 gezeigt wird, ermöglicht der in der vorliegenden Arbeit gewählte Modellierungsansatz eine Rückführung der in den Abschnitten 1.2 bis 1.4 vorgestellten nonparametrischen Erweiterungen auf den Fall eines generalisierten linearen Modells, ohne die durch diese Erweiterungen erzielte höhere Flexibilität aufgeben zu müssen. Dabei erweist sich die Einsparung zusätzlicher Backfitting–Schleifen, wie für GAM's und VCM's nötig, als entscheidender Vorteil des im folgenden propagierten Modellierungsansatzes für nonparametrische Effekte.

2 Basisfunktionen

Für die Modellierung und Schätzung nonparametrischer Komponenten in generalisierten Modellen steht eine Vielzahl alternativer Ansätze zur Verfügung. Relativ schwache Voraussetzungen an die zu modellierenden unbekannten Effekte sichern dabei die weitgehende Allgemeingültigkeit der jeweiligen Methode. Das in Kapitel 1 vorgestellte Konzept der Glättungssplines setzt die zweimalige Differenzierbarkeit der Regressionsfunktionen $f_{(1)}, \ldots, f_{(p)}$ voraus und erfordert darüber hinaus die quadratische Integrierbarkeit der zweiten Ableitungen $f''_{(1)}, \ldots, f''_{(p)}$.

Ansätze zur Modellierung in Basisfunktionen beruhen auf der Annahme, daß die unbekannten, zu schätzenden Regressionsfunktionen durch die Elemente eines bekannten Funktionenraumes approximierbar sind. Existiert für diesen Raum eine analytisch sowie numerisch leicht handhabbare Basis, so kann die Schätzung nonparametrischer Effekte auf eine Schätzung der entsprechenden Basiskoeffizienten reduziert werden. Im Unterschied zu den Glättungssplines ist die Approximation von unbekannten Regressionsfunktionen also kein konzeptionelles Ergebnis, sondern bildet vielmehr den expliziten Ausgangspunkt für deren Modellierung.

Der in der vorliegenden Arbeit gewählte und in den folgenden Kapiteln näher erläuterte Ansatz unterstellt die Approximierbarkeit der Regressionsfunktionen durch sogenannte Polynom- bzw. Regressionssplines, die sich stückweise aus Polynomen niedrigen Grades zusammensetzen. Gegenstand dieses Kapitels sind ausführliche Betrachtungen zur Menge der Polynom–Splines auf einem abgeschlossenen Intervall $[a, b] \subset I\!R$. Von primärem Interesse sind dabei Aussagen zur Existenz und Gestalt von Basen dieses Funktionenraumes. Die abgeleiteten Eigenschaften der Basisfunktionen stützen sich im wesentlichen auf die Ausführungen in de Boor (1978) und Hämmerlin & Hoffmann (1992).

Neben grundlegenden Definitionen und detaillierten Betrachtungen zur Äquivalenz der vorgestellten Basen, wurde dieses Kapitel um die Darstellung diskreter Penalisierungskonzepte ergänzt, die als Approximationen der für Glättungssplines üblichen Bestrafungsterme angesehen werden können und deren Part bei der hier propagierten Schätzmethode übernehmen.

2.1 Polynom-Splines

Durch die Knotenmenge $\Omega_M := \{t_m\}_{m=0}^M$ sei eine Zerlegung des abgeschlossenen Intervalls $[a,b] \subset \mathbb{R}$ definiert, d.h. $a = t_0 < \ldots < t_M = b$. Eine Funktion $s\colon [a,b] \longrightarrow \mathbb{R}$ heißt *Polynom-Spline vom Grad* n, $n \in \mathbb{N}_0$, zur Zerlegung Ω_M, falls sie folgenden Bedingungen genügt

- s ist in $[a,b]$ $(n-1)$–mal stetig differenzierbar: $s \in C_{n-1}[a,b]$,
- $s \in \mathbb{P}_n := \left\{ p \in C(\mathbb{R}) \,\middle|\, p(x) = \sum_{r=0}^n \alpha_r x^r \right\}$ für $x \in [t_{m-1}, t_m)$, $1 \le m \le M$, d.h. s setzt sich stückweise aus Polynomen vom Grad n zusammen.

Unter $C_{-1}[a,b]$ ist dabei der Raum der auf $[a,b]$ stückweise stetigen Funktionen zu verstehen.

Den linearen Raum der Polynom-Splines vom Grad n zur Zerlegung Ω_M bezeichnen wir kurz mit $S_n(\Omega_M)$. Gesucht ist eine Basis dieses Raumes. Zu diesem Zweck definieren wir für jedes $l \in \mathbb{N}_0$ die bivariaten Funktionen

$$q_l(x,t) := (x-t)_+^l := \begin{cases} (x-t)^l, & \text{falls } x \ge t \\ 0, & \text{falls } x < t \end{cases}, \quad (x-t)_+^0 := 1, \text{ für } x \ge t.$$

Abbildung 2.1 zeigt Funktionen dieses Typ's für den Fall $l = 3$. Die abgebildeten Funktionen sind in dem Sinn lokalisiert, als daß ihr zweites Argument und damit der linke Randpunkt ihres Trägers jeweils durch ein Element der Knotenmenge Ω_M fest vorgegeben ist.

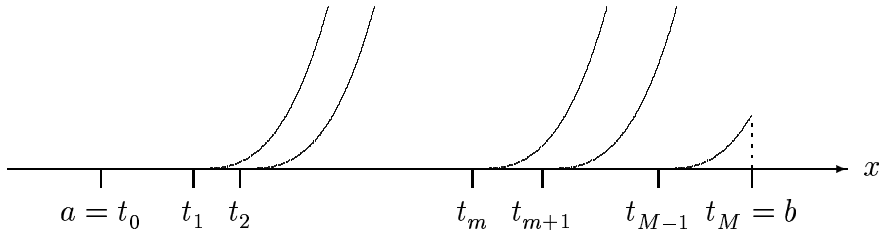

ABBILDUNG 2.1: *Basisfunktionen* $q_3(x,t_m)$, $m = 1,\ldots,M-1$, *der Truncated–Power-Basis von* $S_3(\Omega_M)$.

Mit den Vereinbarungen $p_r(x) := x^r$, $0 \le r \le n$, und $q_{nm}(x) := q_n(x,t_m)$ liefern Hämmerlin & Hoffmann (1992) mit nachstehendem Satz eine erste Aussage über eine mögliche Basisdarstellung von Elementen des Raumes $S_n(\Omega_M)$.

Satz 2.1. *$S_n(\Omega_M)$ bildet einen linearen Raum der Dimension $(n+M)$. Eine Basis dieses Raumes ist durch die Menge $\{p_0, \ldots, p_n, q_{n1}, \ldots, q_{n,M-1}\}$ gegeben. Diese Menge wird auch als T(runcated)–P(ower)-Basis bezeichnet.*

Nach Satz 2.1 besitzt jeder Polynom–Spline $s \in S_n(\Omega_M)$ eine Darstellung in der Truncated–Power-Basis

$$s(x) = \sum_{r=0}^{n} \tilde{\alpha}_r p_r(x) + \sum_{m=1}^{M-1} \alpha_m q_{nm}(x) \tag{2.1}$$

mit $x \in [a, b]$ und eindeutig festgelegten Basiskoeffizienten $\tilde{\alpha}_r$, $r = 0, \ldots, n$, und α_m, $m = 1, \ldots, M-1$.

2.2 B–Splines

Eine weitere Basis des Spline–Raumes $S_n(\Omega_M)$ läßt sich aus den sogenannten B(asic)–Spline–(Curves) gewinnen (de Boor, 1978, Hämmerlin & Hoffmann, 1992, Dierckx, 1993). Ausgehend von einer allgemeinen Definition, werden in diesem Abschnitt die wichtigsten Eigenschaften der B–Splines dargestellt und eine B–Spline-Basis für $S_n(\Omega_M)$ konstruiert.

Um die Bildungsvorschrift für einen B–Spline allgemein formulieren zu können, sind zunächst einige Begriffsbildungen notwendig. In notationeller Anlehnung an Hämmerlin & Hoffmann (1992) betrachten wir dazu die unendliche Knotenmenge $\Omega_\infty := \{t_\nu\}_{\nu \in \mathbb{Z}}$, $t_\nu < t_{\nu+1}$, mit $t_\nu \to -\infty$ für $\nu \to -\infty$ und $t_\nu \to \infty$ für $\nu \to \infty$. Die Indizierung der Knoten sei dabei so gewählt, daß $\Omega_M \subset \Omega_\infty$.

Für eine Funktion $f : \mathbb{R} \longrightarrow \mathbb{R}$ sowie Werte $z_j, \ldots, z_{j+k}$, $j \in \mathbb{Z}$, $k \in \mathbb{N}_0$, aus dem Definitionsbereich von f führen wir die *Steigung k-ter Ordnung* rekursiv ein als

$$\begin{aligned} [z_j]f &:= f(z_j) \\ [z_{j+k} \ldots z_j]f &:= \frac{[z_{j+k} \ldots z_{j+1}]f - [z_{j+k-1} \ldots z_j]f}{z_{j+k} - z_j}\,. \end{aligned} \tag{2.2}$$

Existiert für die Funktion f eine nichttriviale faktorielle Zerlegung $f = u \cdot v$, dann gilt für die Steigung k-ter Ordnung die sogenannte *Leibnizsche Regel*

$$[z_{j+k} \dots z_j]f \;=\; \sum_{i=j}^{j+k} \big([z_i \dots z_j]u\big) \cdot \big([z_{j+k} \dots z_i]v\big). \tag{2.3}$$

Ebenfalls rekursiv definieren wir die *Differenzen k-ter Ordnung*

$$\Delta^0 f(z_j) \;:=\; f(z_j) \quad \text{und} \quad \Delta^k f(z_j) \;:=\; \Delta^{k-1} f(z_{j+1}) - \Delta^{k-1} f(z_j)\,.$$

Ist speziell f die identische Abbildung, liefert obige Darstellung die rekursive Bildungsvorschrift für die k-te Differenz $\Delta^k z_j$ der Komponenten des Vektors $\boldsymbol{z} := (z_j, \dots, z_{j+k})'$.

Für multivariate Funktionen werden die Differenzen im Sinne einer eindeutigen Zuordnung mit einem zusätzlichen Index versehen. So beschreibt der Operator Δ_2^k die bezüglich des zweiten Arguments zu bildenden k-ten Differenzen einer mindestens bivariaten Funktion. Darüber hinaus gilt für k-te Differenzen die

Proposition 2.2. *Für eine reellwertige Funktion f und Werte $z_j, \dots, z_{j+k}$, $j \in \mathbb{Z}$, $k \in \mathbb{N}_0$, aus dem Definitionsbereich von f läßt sich folgende explizite Darstellung der Differenzen k-ter Ordnung ableiten*

$$\Delta^k f(z_j) \;=\; \sum_{l=0}^{k} (-1)^{k+l} \binom{k}{l} f(z_{j+l}).$$

Für den Beweis dieser Proposition wird auf den Anhang verwiesen.

Sind die Werte $z_j, \dots, z_{j+k}$ äquidistant mit $z_{j+s} = z_j + sh$, $s = 0, \dots, k$ und $h \in \mathbb{R}_+ \backslash \{0\}$, lassen sich Steigung und Differenz k-ter Ordnung über den Zusammenhang

$$k! h^k \cdot [z_{j+k} \dots z_j]f \;=\; \Delta^k f(z_j) \tag{2.4}$$

in Beziehung setzen (Hämmerlin & Hoffmann, 1992).

Der B–Spline vom Grad l zum Knoten t_ν aus der Menge Ω_∞ wird mit diesen Festlegungen durch die Vorschrift

$$B_{l\nu}(x) \;=\; (t_{\nu+l+1} - t_\nu)[t_{\nu+l+1} \dots t_\nu]\tilde{q}_l(\cdot, x) \tag{2.5}$$

erklärt. Die Funktionen

$$\tilde{q}_l(t,x) := (t-x)_+^l := \begin{cases} (t-x)^l, & \text{falls } x < t \\ 0, & \text{falls } x \geq t \end{cases}$$

ergeben sich dabei für $l \geq 1$ via $\tilde{q}_l(t,x) = q_l(2t-x,t)$ aus der Spiegelung der Funktionen $q_l(x,t)$ an der Achse $y \equiv t$. Für äquidistante Knoten mit der festen Schrittweite $h \in \mathbb{R}_+ \backslash \{0\}$ vereinfacht sich (2.5) wegen (2.4) zu

$$B_{l\nu}(x) \;=\; (l+1)h \cdot [t_{\nu+l+1} \ldots t_\nu]\tilde{q}_l(\cdot,x) \;=\; \frac{1}{l!h^l} \cdot \Delta_1^{l+1}\tilde{q}_l(t_\nu,x). \tag{2.6}$$

Ausgehend von der rekursiven Definition der Steigungen lassen sich folgende Rekursionsformeln für B–Splines ableiten (Hämmerlin & Hoffmann, 1992)

$$B_{l\nu}(x) = \frac{x-t_\nu}{t_{\nu+l}-t_\nu} B_{l-1,\nu}(x) \;+\; \frac{t_{\nu+l+1}-x}{t_{\nu+l+1}-t_{\nu+1}} B_{l-1,\nu+1}(x) \tag{2.7}$$

bzw.

$$B_{l\nu}(x) = (lh)^{-1} \cdot \big((x-t_\nu)B_{l-1,\nu}(x) + (t_{\nu+l+1}-x)B_{l-1,\nu+1}(x)\big) \tag{2.8}$$

für den Fall äquidistanter Knoten.

Mit $B_{0\nu}(x) = I_{[t_\nu,t_{\nu+1})}(x)$, $t_\nu \in \Omega_\infty$, gemäß Definition (2.5), liefert (2.7) eine einfache Möglichkeit, B–Splines beliebigen Grades zu berechnen. Die rekursiven Darstellungen sind insbesondere vor dem Hintergrund einer effizienten numerischen Implementation von B–Splines von Bedeutung.

Abbildung 2.2 stellt exemplarisch B–Splines ersten und zweiten Grades dar. B–Splines sind – ebenso wie die Funktionen $q_l(\cdot,t_\nu)$ – im Sinne einer eindeutigen Zuordnung zu den Knoten der Menge Ω_∞ lokalisiert. Im Gegensatz zu den abgeschnittenen Polynomen $q_l(\cdot,t_\nu)$ besitzen B–Splines aber einen durch Knoten beschränkten Träger: $supp\, B_{l\nu}(x) = (t_\nu, t_{\nu+l+1})$, $l \geq 1$.

Eine Eigenschaft der B–Splines, die im später betrachteten Modellierungskontext von Bedeutung sein wird, ist die sogenannte *Zerlegung der Einheit*: Für B–Splines $B_{l\nu}$ und jedes $x \in \mathbb{R}$ gilt

$$\sum_{\nu \in \mathbb{Z}} B_{l\nu}(x) \;=\; 1. \tag{2.9}$$

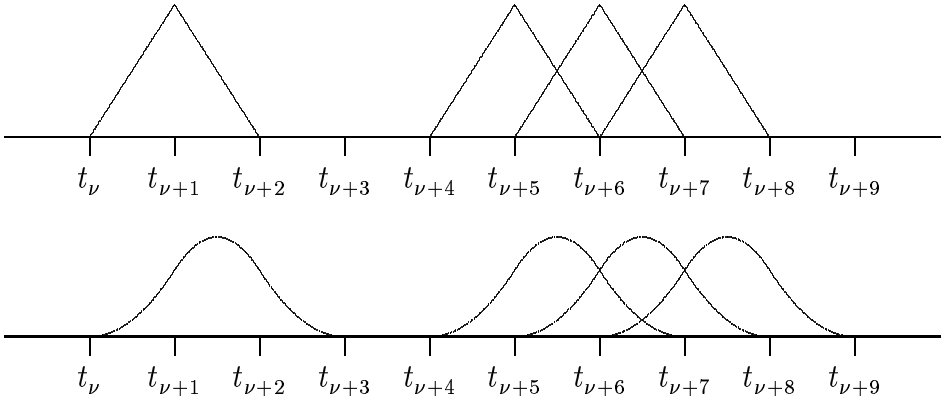

ABBILDUNG 2.2: *B–Splines ersten Grades (obere Abbildung) und zweiten Grades (untere Abbildung) für äquidistante Knoten aus* Ω_∞.

Weitere Charakteristika, die sich für lineare und quadratische B–Splines anhand der Abbildung 2.2 leicht verifizieren lassen, seien an dieser Stelle lediglich stichpunktartig notiert (Eilers & Marx, 1996):

Ein B–Spline vom Grad l, $l \in \mathbb{N}_0$, zum Knoten $t_\nu \in \Omega_\infty$

- besteht aus $l+1$ Polynom-Fragmenten l-ten Grades, die an l inneren Knoten miteinander verbunden sind,
- ist an diesen inneren Knoten $(l-1)$-mal stetig differenzierbar,
- überlappt sich mit $2l$ benachbarten B–Splines.

Darüber hinaus sind in jedem festen Punkt $x \in \mathbb{R}\backslash\Omega_\infty$ genau $l+1$ B–Splines l-ten Grades von Null verschieden.

Basierend auf diesen Eigenschaften liefern Hämmerlin & Hoffmann (1992) eine Aussage über die Gestalt einer B–Spline-Basis für den Raum $S_n(\Omega_M)$, die in folgendem Satz Ausdruck findet.

Satz 2.3. *Jeder Polynom-Spline* $s \in S_n(\Omega_M)$ *besitzt im Intervall* $[a,b]$ *eine eindeutige Darstellung durch B–Splines*

$$s(x) = \sum_{m=-n}^{M-1} \beta_m B_{nm}(x)\,, \quad x \in [a,b].$$

Zur Konstruktion einer B–Spline-Basis für $S_n(\Omega_M)$ ist die Knotenmenge Ω_M also um die Knoten $t_{-n}, \ldots, t_{-1}, t_{M+1}, \ldots, t_{M+n}$ aus Ω_∞ zu erweitern. Lediglich für $n = 0$ sind die zur Lokalisation der B–Spline-Basis notwendigen Knoten durch die Menge Ω_M gegeben.

Der Fall $n = 0$ bedarf einer weiteren Bemerkung: In Abwandlung zur Definition (2.5) gilt für die Basisdarstellung konstanter Polynom–Splines $s \in S_0(\Omega_M)$ durch B–Splines nullten Grades der Zusammenhang $B_{0,M-1}(t_M = b) = 1$.

2.3 Äquivalente Basisdarstellungen

Die Menge $\{t_{\nu-1}, \ldots, t_{\nu+l+2}\} =: \Omega_{l\nu} \subset \Omega_\infty$ stellt eine Zerlegung des Intervalls $[t_{\nu-1}, t_{\nu+l+2}]$ dar. Für den B–Spline $B_{l\nu} \in S_l(\Omega_{l\nu})$ existiert nach Satz 2.1 eine eindeutige Darstellung in der Truncated–Power-Basis von $S_l(\Omega_{l\nu})$. Von dieser Basisdarstellung ausgehend, soll im folgenden untersucht werden, wie sich die B–Spline–Repräsentation eines Polynom–Splines $s \in S_n(\Omega_M)$ in eine äquivalente Darstellung in der Truncated–Power-Basis überführen läßt.

Die Rekursionsformel (2.7) zur Berechnung von B–Splines resultiert aus einer Anwendung der bereits erwähnten Leibnizschen Regel (2.3) auf die Funktionen $\tilde{q}_l(t, x) = (t - x) \cdot \tilde{q}_{l-1}(t, x)$. Für $j \in \mathbb{Z}$ und $l \geq 1$ gilt demnach

$$\begin{aligned}[t_{j+l+1} \ldots t_j]\tilde{q}_l(., x) &= \frac{t_{j+l+1} - x}{t_{j+l+1} - t_j} \cdot [t_{j+l+1} \ldots t_{j+1}]\tilde{q}_{l-1}(., x) - \\ &\quad - \frac{t_j - x}{t_{j+l+1} - t_j} \cdot [t_{j+l} \ldots t_j]\tilde{q}_{l-1}(., x), \qquad (2.10)\end{aligned}$$

(Hämmerlin & Hoffmann, 1992). Eine völlig analoge Anwendung der Leibnizschen Regel (2.3) auf die Funktionen $q_l(x, t) = (x - t) \cdot q_{l-1}(x, t)$ liefert

$$\begin{aligned}-[t_{j+l+1} \ldots t_j]q_l(x, .) &= \frac{t_{j+l+1} - x}{t_{j+l+1} - t_j} \cdot [t_{j+l+1} \ldots t_{j+1}]q_{l-1}(x, .) - \\ &\quad - \frac{t_j - x}{t_{j+l+1} - t_j} \cdot [t_{j+l} \ldots t_j]q_{l-1}(x, .). \qquad (2.11)\end{aligned}$$

Die Rekursionsformeln (2.10) und (2.11) erlauben folgende Aussage über den Zusammenhang zwischen den Steigungen von $\tilde{q}_l(., x)$ und $q_l(x, .)$

Proposition 2.4. *Es gilt die folgende Beziehung zwischen den Steigungen der Funktionen $\tilde{q}_l(.,x)$ und $q_l(x,.)$, $l \in \mathbb{N}_0$,*

$$[t_{j+l+1} \dots t_j]\tilde{q}_l(.,x) = (-1)^{l+1} \cdot [t_{j+l+1} \dots t_j]q_l(x,.),\ j \in \mathbb{Z}.$$

Der Beweis dieser Proposition läßt sich mit den Gleichungen (2.10) und (2.11) induktiv über l führen. Die detaillierte Beweisführung wird im Anhang A gegeben.

Für äquidistante Knotenwahl reduziert sich die Aussage von Proposition 2.4 unter Berücksichtigung von (2.4) auf

$$\Delta_1^{l+1}\tilde{q}_l(t_j,x) = (-1)^{l+1} \cdot \Delta_2^{l+1}q_l(x,t_j). \tag{2.12}$$

Aus Proposition 2.4 kann folgende Darstellung des B–Splines $B_{l\nu}$ abgeleitet werden

Proposition 2.5. *Der B–Spline $B_{l\nu}$ vom Grad l zum Knoten $t_\nu \in \Omega_\infty$ kann äquivalent zu (2.5) auch definiert werden als*

$$B_{l\nu}(x) \ = \ (-1)^{l+1} \cdot (t_{\nu+l+1} - t_\nu)[t_{\nu+l+1} \dots t_\nu]q_l(x,\cdot),$$

bzw. $B_{l\nu}(x) = (-1)^{l+1} \cdot (l!h^l)^{-1} \cdot \Delta_2^{l+1}q_l(x,t_\nu)$ für den Fall äquidistanter Knoten.

Proposition 2.5 liefert die nach Satz 2.1 existierende Repräsentation von $B_{l\nu}$ in der Truncated–Power-Basis $\{p_0, \dots, p_l, q_{l\nu}, \dots, q_{l,\nu+l+1}\}$ von $S_l(\Omega_{l\nu})$. Ausgehend von (2.1) ist dabei $\tilde{\alpha}_r = 0$, $r = 0, \dots, l$, während sich die Koeffizienten α_m, $m = \nu, \dots, \nu+l+1$, aus Proposition 2.5 berechnen lassen.

Die mit Satz 2.3 gegebene B–Spline-Repräsentation von $s \in S_n(\Omega_M)$ soll nun unter Verwendung von Proposition 2.5 in eine äquivalente Darstellung in der entsprechenden Truncated–Power-Basis von $S_n(\Omega_M)$ überführt werden. Dazu beschränken wir uns im folgenden auf die Situation äquidistanter Knoten.

Setzt man $c_n := (-1)^{n+1} \cdot (n!h^n)^{-1}$ und $\tilde{c}_{n,k} := (-1)^{n+k} \cdot \binom{n}{k}$, so gilt für jedes $x \in [a,b]$ mit den Propositionen 2.2 und 2.5

$$c_n^{-1} \cdot s(x) \ = \ c_n^{-1} \cdot \sum_{m=-n}^{M-1} \beta_m B_{nm}(x) \ = \ \sum_{m=-n}^{M-1} \beta_m \Delta_2^{n+1} q_n(x,t_m)$$

$$= \sum_{k=0}^{n+1} \sum_{m=-n}^{M-1} \tilde{c}_{n+1,k}\, \beta_m q_n(x, t_{m+k}) \;= \sum_{k=0}^{n+1} \sum_{m=-n}^{M-k-1} \tilde{c}_{n+1,k}\, \beta_m q_n(x, t_{m+k}).$$

Letzteres gilt dabei wegen $q_n(x, t_{m+k}) = 0$ für $m+k \geq M$ und $n \geq 1$. Obiger Ausdruck läßt sich damit weiter äquivalent umformen zu

$$\begin{aligned} c_n^{-1} \cdot s(x) \;&= \sum_{k=0}^{n+1} \sum_{m=-n+k}^{M-1} \tilde{c}_{n+1,k}\, \beta_{m-k} q_n(x, t_m) \\ &= \underbrace{\sum_{k=0}^{n} \sum_{m=-n+k}^{0} \tilde{c}_{n+1,k}\, \beta_{m-k} q_n(x, t_m)}_{(a)} + \underbrace{\sum_{k=0}^{n+1} \sum_{m=1}^{M-1} \tilde{c}_{n+1,k}\, \beta_{m-k} q_n(x, t_m)}_{(b)}, \end{aligned}$$

wobei

$$(b) \;= \sum_{m=1}^{M-1} q_n(x, t_m) \sum_{k=0}^{n+1} \tilde{c}_{n+1,k}\, \beta_{m-k} = \sum_{m=1}^{M-1} q_{nm}(x) \sum_{k=0}^{n+1} (-1)^{n+1+k} \binom{n+1}{k} \beta_{m-k}$$

mit

$$\begin{aligned} \sum_{k=0}^{n+1} (-1)^{n+1+k} \binom{n+1}{k} \beta_{m-k} &= (-1)^{n+1} \sum_{k=0}^{n+1} (-1)^{n+1+k} \binom{n+1}{k} \beta_{m-n-1+k} \\ &= (-1)^{n+1} \Delta^{n+1} \beta_{m-n-1} \end{aligned}$$

nach Proposition 2.2 für die identische Abbildung.

Ferner ist

$$(a) = \sum_{m=-n}^{0} q_n(x, t_m) \sum_{k=0}^{n+m} \tilde{c}_{n+1,k}\, \beta_{m-k} \;= \sum_{m=-n}^{0} q_n(x, t_m) \sum_{k=-n}^{m} \tilde{c}_{n+1,m-k}\, \beta_k.$$

Setze $\tilde{\tilde{c}}_{m,n} := \sum_{k=-n}^{m} \tilde{c}_{n+1,m-k}\, \beta_k$. Da $x \in [a = t_0, b]$, gilt für $m \in \{-n, \ldots, 0\}$

$$q_n(x, t_m) = (x - t_m)^n = \sum_{r=0}^{n} (-1)^{n-r} \cdot \binom{n}{r} \cdot x^r t_m^{n-r}$$

gemäß Binomischem Lehrsatz und damit

$$(a) = \sum_{r=0}^{n} \sum_{m=-n}^{0} (-1)^{n-r} \binom{n}{r} x^r t_m^{n-r} \tilde{\tilde{c}}_{m,n} \;= \sum_{r=0}^{n} x^r (-1)^{n-r} \binom{n}{r} \sum_{m=-n}^{0} t_m^{n-r} \tilde{\tilde{c}}_{m,n}.$$

Somit kann jede B–Spline-Entwicklung eines Polynom-Splines $s \in S_n(\Omega_M)$ in

eine äquivalente Darstellung in der Truncated–Power-Basis überführt werden und es gilt die folgende

Proposition 2.6. *Für die mit den Sätzen 2.1 und 2.3 gegebenen Basisdarstellungen eines Polynom-Splines $s \in S_n(\Omega_M)$*

$$s(x) \;=\; \sum_{m=-n}^{M-1} \beta_m B_{nm}(x) \;=\; \sum_{r=0}^{n} \tilde{\alpha}_r p_r(x) \;+\; \sum_{m=1}^{M-1} \alpha_m q_{nm}(x)$$

gelten für $r = 0, \ldots, n$ und $m = 1, \ldots, M-1$ folgende Beziehungen zwischen den jeweiligen Basiskoeffizienten

$$\begin{aligned} \tilde{\alpha}_r(\boldsymbol{\beta}) \;&=\; c_n \cdot (-1)^{n-r} \cdot \binom{n}{r} \sum_{m=-n}^{0} t_m^{n-r} \tilde{\tilde{c}}_{m,n} \\ &= (n! h^n)^{-1} \binom{n}{r} \sum_{m=-n}^{0} t_m^{n-r} \sum_{k=-n}^{m} (-1)^{n-r+m-k} \cdot \binom{n+1}{m-k} \cdot \beta_k, \end{aligned} \tag{2.13}$$

und

$$\alpha_m(\boldsymbol{\beta}) \;=\; c_n \cdot (-1)^{n+1} \cdot \Delta^{n+1} \beta_{m-n-1} = (n! h^n)^{-1} \cdot \Delta^{n+1} \beta_{m-n-1}, \tag{2.14}$$

wobei $\boldsymbol{\beta} = (\beta_{-n}, \ldots, \beta_{M-1})'$ die B–Spline-Koeffizienten subsumiert.

Die bei der Herleitung von Proposition 2.6 für $x \in [a,b]$ benutzte Aussage

$$q_n(x, t_{m+k}) = 0\,, \;\; m+k \geq M,$$

gilt, wie angegeben, nur für $n \geq 1$. Für $n = 0$ ist sie ebenfalls zutreffend, falls $x \in [a,b)$. Ist $n = 0$ und $x = b$, gilt $q_n(x, t_{m+k}) = 0$ erst für $m+k > M$, wegen $q_0(b, t_M) = q_0(t_M, t_M) = 1$. Mit $B_{0,M-1}(b) = 1$ ist aber auch

$$\begin{aligned} \sum_{m=0}^{M-1} \beta_m B_{0m}(b) \;&=\; \beta_{M-1} \;=\; \beta_0 + \beta_{M-1} - \beta_0 \;=\; \beta_0 \;+ \sum_{m=1}^{M-1} (\beta_m - \beta_{m-1}) \\ &=\; \beta_0 b^0 + \sum_{m=1}^{M-1} \Delta^1 \beta_{m-1} \, (b - t_m)^0 \;=\; \sum_{r=0}^{0} \beta_r p_r(b) \;+ \sum_{m=1}^{M-1} \Delta^1 \beta_{m-1} \, q_{0m}(b)\,, \end{aligned}$$

so daß $\tilde{\alpha}_0(\boldsymbol{\beta}) = \beta_0$ und $\alpha_m(\boldsymbol{\beta}) = \Delta^1 \beta_{m-1}$, $m = 1, \ldots, M-1$, falls $n = 0$ und $x = b$. Gleiches folgt aus (2.13) und (2.14) für $n = 0$ und $x \in [a,b)$, d.h. Proposition 2.6 gilt generell für $n \in \mathbb{N}_0$ und $x \in [a,b]$.

2.4 Diskrete Penalisierungskonzepte

Wie in den einleitenden Bemerkungen zu diesem Kapitel bereits erwähnt, verfolgen die angestellten Untersuchungen das Ziel, im Regressionskontext unbekannte funktionale Kovariableneffekte durch Polynom-Splines bzw. deren Basisdarstellung zu approximieren.

Für ein stetiges Merkmal x, mit beobachteten Realisationen x_i, $i = 1, \ldots, N$, aus dem Intervall $[a, b] := [x_{\min}, x_{\max}]$ modellieren Ruppert & Carroll (2000), Ruppert (2002) dessen Effekt über eine Darstellung in der Truncated–Power-Basis des Spline-Raumes $S_n(\Omega_M)$

$$f(x_i) = \sum_{r=0}^{n} \tilde{\alpha}_r p_r(x_i) + \sum_{m=1}^{M-1} \alpha_m q_{nm}(x_i), \quad i = 1, \ldots, N. \tag{2.15}$$

Neben den Basiskoeffizienten $\tilde{\alpha}_r$, $r = 0, \ldots, n$ und α_m, $m = 1, \ldots, M-1$, bilden der Polynomgrad n sowie die Anzahl $M+1$ der Knoten und deren Lokalisation die unbekannten Modellparameter. Ruppert & Carroll (2000) empfehlen, die Lage der Knoten an den Stichprobenquantilen zu orientieren. Alternativ lassen sich die Knoten auch auf einem äquidistanten Gitter platzieren.

Die Knotenanzahl ist die wichtigste Determinante für die Modellkomplexität einerseits und die Qualität der resultierenden Schätzungen andererseits. Wenige Knoten sind gleichbedeutend mit einer geringen Anzahl zu schätzender Parameter, komplexere Strukturen in der unbekannten Funktion lassen sich damit aber nur ungenügend nachbilden. Wächst die Zahl der Knoten, steigt in gleichem Maße der Grad an Flexibilität des zugrunde liegenden Modells. Hohe Flexibilität geht aber oft mit einer unerwünschten Datennähe (Overfitting) des geschätzten Effekts einher, im Extremfall resultiert ein interpolierender Polynom-Spline als Schätzung.

Als Kompromiß schlagen Ruppert & Carroll (2000), Ruppert (2002) die Modellierung in vielen Basisfunktionen unter gleichzeitiger Verwendung eines sogenannten *Ridge-Penalties* zur Kontrolle der Variabilität vor

$$P_{ridge} = \sum_{m=1}^{M-1} \alpha_m^2. \tag{2.16}$$

Die unbekannten Basiskoeffizienten werden über die Maximierung einer penalisierten Log–Likelihood geschätzt

$$pl(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) \;=\; l(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) \;-\; \frac{1}{2}\,\lambda_{\boldsymbol{\alpha}} P_{ridge} \;\longrightarrow\; \max_{\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}} \tag{2.17}$$

mit $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_0, \ldots, \tilde{\alpha}_n)'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{M-1})'$ sowie einem Glättungsparameter $\lambda_{\boldsymbol{\alpha}} \geq 0$, der die Penalisierungsstärke bestimmt. Der Ridge-Penalty (2.16) läßt sich dabei wie folgt motivieren: das abgeschnittene Polynom $q_{nm}(x)$ ist in jedem Punkt $x \neq t_m$ beliebig oft differenzierbar. Insbesondere gilt

$$\frac{\partial^n}{\partial x^n}\, q_{nm}(x) \;=\; q_{nm}^{(n)}(x) \;=\; n! \cdot q_{0m}(x) \;=\; n! \cdot I_{[t_m, \infty)}(x)\,, \;\; x \neq t_m.$$

Für die n-te Ableitung $f^{(n)}$ des nach (2.15) modellierten Effekts folgt daraus

$$(n!)^{-1} f^{(n)}(x) \;=\; \tilde{\alpha}_n + \sum_{m=1}^{M-1} \alpha_m I_{[t_m, \infty)}(x)\,, \;\; x \in [a, b] \backslash \Omega_M.$$

Demnach hat $f^{(n)}$ die Gestalt einer Treppenfunktion mit den Stufen $n! \cdot \alpha_m$ an den inneren Knoten t_m, $m = 1, \ldots, M-1$. P_{ridge} fungiert daher als Strafterm für die Menge der Sprünge in $f^{(n)}$.

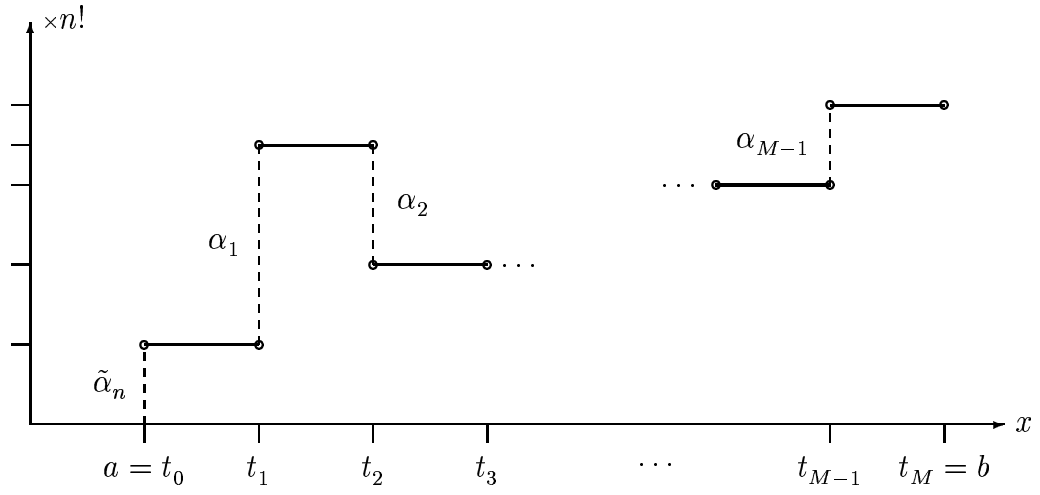

ABBILDUNG 2.3: *Graph von $f^{(n)}$ bei Modellierung in der TP-Basis.*

Ruppert & Carroll (2000) folgend, kann (2.16) als Strafterm für die $(n+1)$-te Ableitung von f betrachtet werden. Da $f^{(n)}$ als unstetige Funktion insbesondere nicht differenzierbar ist, faßt man $f^{(n+1)}$ dabei als *verallgemeinerte Funktion* (Gelfand & Schilow, 1967) auf.

Eilers & Marx (1996) arbeiten mit der B–Spline-Basis von $S_n(\Omega_M)$ und äquidistanten Knoten zur Modellierung der unbekannten Regressionsfunktion

$$f(x_i) \;=\; \sum_{m=-n}^{M-1} \beta_m B_{nm}(x_i) \;=\; Z(i,\cdot)\,\boldsymbol{\beta}\,, \quad i = 1,\ldots,N, \tag{2.18}$$

mit einer *Designmatrix* Z, deren Einträge durch $Z(i,m) := B_{nm}(x_i)$ bestimmt sind und einem Parametervektor $\boldsymbol{\beta} := (\beta_{-n},\ldots,\beta_{M-1})'$. Mit ähnlichen Argumenten motivieren die Autoren die Notwendigkeit einer Penalisierung der Basiskoeffizienten. Der von ihnen propagierte *Differenzenpenalty d-ter Ordnung*, $1 \le d \le M+n-1$, hat die Form

$$P_\Delta^d \;=\; \sum_{m=-n}^{M-1-d} \left(\Delta^d \beta_m\right)^2 \;=\; \sum_{m=1}^{M+n-d} \left(\Delta^d \beta_{m-n-1}\right)^2. \tag{2.19}$$

Definiert man die $(M+n-1)\times(M+n)$ Kontrastmatrix D^1_{M+n} als

$$D^1_{M+n} := \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}, \tag{2.20}$$

läßt sich der Vektor der Differenzen erster Ordnung auch schreiben als

$$\Delta^1\boldsymbol{\beta} \;:=\; \left(\Delta^1\beta_{-n},\ldots,\Delta^1\beta_{M-2}\right)' \;=\; D^1_{M+n}\boldsymbol{\beta}.$$

Der Differenzenpenalty erster Ordnung P^1_Δ reduziert sich damit auf die kompakte Form $P^1_\Delta = \boldsymbol{\beta}'\left(D^1_{M+n}\right)' D^1_{M+n}\boldsymbol{\beta}$. Differenzenpenalties beliebiger Ordnung $d > 1$ resultieren mit $D^d_{M+n} := D^1_{M+n-d+1} \cdot \ldots \cdot D^1_{M+n-1} \cdot D^1_{M+n}$ aus

$$\Delta^d\boldsymbol{\beta} \;=\; D^d_{M+n}\boldsymbol{\beta} \quad \text{und} \quad P^d_\Delta \;=\; \left(\Delta^d\boldsymbol{\beta}\right)'\Delta^d\boldsymbol{\beta} \;=\; \boldsymbol{\beta}'\left(D^d_{M+n}\right)' D^d_{M+n}\boldsymbol{\beta}. \tag{2.21}$$

Die unbekannten Basiskoeffizienten in (2.18) werden wiederum über die Maximierung einer penalisierten Log–Likelihood geschätzt

$$pl(\boldsymbol{\beta}) \;=\; l(\boldsymbol{\beta}) \;-\; \frac{1}{2}\,\lambda_{\boldsymbol{\beta}} P^d_\Delta \;\longrightarrow\; \max_{\boldsymbol{\beta}} \tag{2.22}$$

mit einem Glättungsparameter $\lambda_{\boldsymbol{\beta}} \ge 0$, der die Stärke der Penalisierung festlegt. Setzt man in (2.22) $d := n+1$, erhält man mit (2.19) sowie der Äquivalenzbeziehung (2.14) die

Proposition 2.7. *Die Existenz und Eindeutigkeit von Lösungen vorausgesetzt, sind die Optimierungsprobleme (2.17) und (2.22) unter den Prämissen äquidistanter Knoten, eines Differenzenpenalties $(n+1)$-ter Ordnung und der Glättungsparameterkonstellation $\lambda_{\boldsymbol{\alpha}} = (n!h^n)^2\,\lambda_{\boldsymbol{\beta}}$ äquivalent.*

Zum Beweis dieser Aussage wird auf den Anhang A verwiesen.

Die konkrete Modellierung in B–Splines nutzend, liefern Eilers & Marx (1996) Aussagen über die Beziehung zwischen den Differenzenpenalties und den klassischen Bestrafungstermen

$$\int \Big(f^{(d)}(x)\Big)^2\,dx = \int \Big(\sum_m \beta_m B_{nm}^{(d)}(x)\Big)^2\,dx. \tag{2.23}$$

Für jede Konstellation $d < n$ stellt der Bestrafungsterm P_Δ^d eine hinreichend gute Approximation für (2.23) dar, wobei die Güte dieser Approximation mit wachsender Differenz $n-d$ abnimmt. Prinzipiell möglich sind natürlich auch Kombinationen von B–Splines n–ten Grades und Differenzenpenalties P_Δ^d der Ordnung $d > n-1$. Für diese Situationen fehlen jedoch entsprechende Analogieaussagen zur klassischen Penalisierung des Integrals der quadrierten Ableitung $f^{(d)}$. Beliebig penalisierte B–Splines werden im folgenden auch kurz als *P–Splines* bezeichnet.

Die Proposition 2.7 und die freie Kombinierbarkeit von Splinegrad und Differenzenordnung definieren P–Splines im Vergleich zur Truncated–Power-Basis mit Ridge-Penalty als das allgemeinere Konzept im Kontext penalisierter Basisfunktionsansätze.

2.5 Nützliche Eigenschaften penalisierter B–Splines

Im folgenden sollen einige Eigenschaften von P–Splines und der aus der Maximierung der penalisierten Log–Likelihood (2.22) resultierenden Schätzung $\hat{f}$ dargestellt werden. Nach Hämmerlin & Hoffmann (1992) gilt für äquidistante Knotenwahl folgende Rekursionsformel für die erste Ableitung des B–Splines $B_{l\nu}$, $l \geq 1$, zum Knoten t_ν der Knotenmenge Ω_∞

$$B'_{l\nu}(x) \;=\; h^{-1}\big\{B_{l-1,\nu}(x) - B_{l-1,\nu+1}(x)\big\}. \tag{2.24}$$

Mit $B_l(x,t_\nu) := B_{l\nu}(x)$, läßt sich (2.24) auch als

$$B'_{l\nu}(x) \;=\; \frac{\partial}{\partial x}\, B_l(x,t_\nu) \;=\; (-h)^{-1}\cdot \Delta_2^1\, B_{l-1}(x,t_\nu) \tag{2.25}$$

schreiben. Für B–Splines ersten Grades beschränkt sich dabei die Gültigkeit von (2.25) auf Punkte $x \notin \{t_\nu,\ldots,t_{\nu+l+1}\}$. Die Rekursionsformel für die erste Ableitung läßt sich induktiv auf Ableitungen höherer Ordnung erweitern.

Proposition 2.8. *Für die d–te Ableitung des B–Splines $B_{l\nu}$ an einer beliebigen Stelle $x \in \mathbb{R}$ gilt die folgende Identität*

$$B^{(d)}_{l\nu}(x) \;=\; \frac{\partial^d}{\partial x^d}\, B_l(x,t_\nu) \;=\; (-h)^{-d}\cdot \Delta_2^d\, B_{l-d}(x,t_\nu)\,,\quad 0 \le d < l.$$

Für den Beweis dieser Proposition sei wieder auf den Anhang A verwiesen.

Die nachstehenden Betrachtungen liefern einen Zusammenhang zwischen der d–ten Ableitung der gefitteten Funktion

$$\hat f(x) \;=\; \sum_{m=-n}^{M-1} \hat\beta_m B_{nm}(x)\,,\quad x\in[a,b]$$

und den d–ten Differenzen der aus (2.22) hervorgehenden, geschätzten Basiskoeffizienten $\hat\beta_{-n},\ldots,\hat\beta_{M-1}$. Mit den Propositionen 2.2 und 2.8 gilt für $d<n$, $\tilde c_{d,j} := (-1)^{d+j}\cdot\binom{d}{j}$ und jedes $x\in[a,b]$

$$(-h)^d \sum_{m=-n}^{M-1} \beta_m B^{(d)}_{nm}(x) = \sum_{m=-n}^{M-1} \beta_m \Delta_2^d B_{n-d}(x,t_m) = \sum_{j=0}^{d} \tilde c_{d,j} \sum_{m=-n}^{M-1} \beta_m B_{n-d,m+j}(x)$$

Für $j=0,\ldots,d$ ist aufgrund des beschränkten Trägers der B–Splines

$$\sum_{m=-n}^{M-1} \beta_m B_{n-d,m+j}(x) \;=\; \sum_{m=-n-j+d}^{M-j-1} \beta_m B_{n-d,m+j}(x) \;=\; \sum_{m=-n+d}^{M-1} \beta_{m-j} B_{n-d,m}(x),$$

so daß

$$\sum_{j=0}^{d} \tilde c_{d,j} \sum_{m=-n}^{M-1} \beta_m B_{n-d,m+j}(x) \;= (-1)^d \sum_{m=-n+d}^{M-1} B_{n-d,m}(x) \sum_{j=0}^{d} \tilde c_{d,j}\beta_{m-d+j}\,,$$

also

$$\sum_{m=-n}^{M-1} \beta_m B^{(d)}_{nm}(x) \;=\; h^{-d} \sum_{m=-n+d}^{M-1} \Delta^d\beta_{m-d}\cdot B_{n-d,m}(x),$$

und damit

$$\hat{f}^{(d)}(x) \;=\; \sum_{m=-n}^{M-1} \hat{\beta}_m B_{nm}^{(d)}(x) \;=\; h^{-d} \sum_{m=-n}^{M-1-d} \Delta^d \hat{\beta}_m \cdot B_{n-d,m+d}(x). \qquad (2.26)$$

Wählt man in (2.22) einen unendlich großen Glättungsparameter $\lambda_{\boldsymbol{\beta}}$, sind alle Differenzen $\Delta^d \hat{\beta}_m$, $m = -n, \ldots, M-1-d$, gleich null, also auch $\hat{f}^{(d)}(x) = 0$ für jedes $x \in [a, b]$ nach (2.26), und man erhält die

Proposition 2.9. *Für einen unendlich großen Wert von $\lambda_{\boldsymbol{\beta}}$ und einen Differenzenpenalty der Ordnung d ist der Grenzwert eines P–Spline–Fits ein Polynom vom Grad $d-1$, falls die B–Splines mindestens $(d+1)$–ten Grades sind.*

Da für B–Splines vom Grad $n \geq 1$ globale Differenzierbarkeit lediglich bis zur Ordnung $n-1$ gegeben ist, beschränkt sich die Gültigkeit obiger Proposition auf die Situationen $n > d$. Eilers & Marx (1996) formulieren die Aussage von Proposition 2.9 für $n \geq d$, jedoch ohne die Angabe eines Beweises.

Unabhängig von einer Betrachtung der Ableitungen ermöglicht der mit (2.14) gegebene Zusammenhang eine Erweiterung von Proposition 2.9 auf den Fall $n = d-1$. Aus $d = n+1$ und $\lambda_{\boldsymbol{\beta}} \longrightarrow \infty$ in (2.22) folgt für $m = 1, \ldots, M-1$ mit (2.19) $\Delta^{n+1} \hat{\beta}_{m-n-1} \longrightarrow 0$ und damit $\alpha_m(\hat{\boldsymbol{\beta}}) \longrightarrow 0$ wegen (2.14). Demnach gilt die

Proposition 2.10. *Wählt man in (2.22) $d = n+1$ und läßt den Glättungsparameter $\lambda_{\boldsymbol{\beta}}$ gegen unendlich streben, so entspricht der resultierende P-Spline-Fit einem Polynom n–ten Grades.*

Weitere Eigenschaften penalisierter B–Splines, wie die Fähigkeit polynomiale Beobachtungen exakt zu fitten und datenbasierte Momente zu erhalten seien an dieser Stelle lediglich kommentarlos erwähnt. Detailliertere Ausführungen zu diesen Punkten finden sich in Eilers & Marx (1996).

Vor dem Hintergrund der aufgeführten Eigenschaften und der Äquivalenzaussage von Proposition 2.7 können B–Splines und Differenzenpenalties als ideale Kombination im Kontext der penalisierten Spline–Approximation einer unbekannten Regressionsfunktion bezeichnet werden. Die nachfolgenden Untersuchungen zur Modellierung und Schätzung nonparametrischer Effekte in generalisierten Modellen beschränken sich daher auf die Betrachtung adäquater P–Spline-Ansätze.

3 P-Splines in generalisierten Modellen

Die Modellierung einer unbekannten Regressionsfunktion als Basisdarstellung eines Polynom-Splines und die Schätzung der zugehörigen Basiskoeffizienten über einen penalisierten Maximum-Likelihood-Ansatz werden in diesem Kapitel für die nonparametrischen Komponenten der Abschnitte 1.2-1.4 verallgemeinert. Als Basisfunktionen werden B–Splines betrachtet, die in Kombination mit den Differenzenpenalties das im Kontext penalisierter Basisfunktionsansätze allgemeinere Modellierungskonzept darstellen.

Die Zusammenfassung von Verallgemeinerungsstufen generalisierter linearer Modelle in einer universellen Prädiktorstruktur soll eine simultane Behandlung verschiedenster Ansätze ermöglichen. Neben linearen Prädiktoranteilen werden beliebige Kombinationen der Komponenten aus GAM's, VCM's und Modellen mit bivariaten Funktionen betrachtet.

Es wird gezeigt, daß Basisdarstellungen nonparametrischer Effekte auf eine rein lineare Form des universellen Prädiktors führen. Die somit erzielte Reduktion auf Strukturen generalisierter linearer Modelle gestattet eine Schätzung nonparametrischer Komponenten ohne zusätzliche Backfittingschritte. Matrixschreibweisen für die Penalties gewährleisten darüber hinaus eine unkomplizierte Integration der Strafterme in den Algorithmus zur Parameterschätzung.

3.1 Die universelle Prädiktorstruktur

Die Komponenten des Kovariablenvektors $\boldsymbol{x} := (x_1, \ldots, x_p)'$ seien o.B.d.A. gemäß $\{1, \ldots, p\} = \{\mathcal{D}, \mathcal{S}\}$ angeordnet, wobei x_j diskret, falls $j \in \mathcal{D}$ und x_j stetig, falls $j \in \mathcal{S}$. Ferner bezeichne $\tilde{p} := |\mathcal{D}| + 1$ den ersten Index in $\mathcal{S} \neq \emptyset$. Dann werden die universellen Prädiktoren $\eta_{i,U}$, $i = 1, \ldots, N$, definiert als

$$\begin{aligned} \eta_{i,U} &:= \beta_0 + \eta_{i,L} + \eta_{i,A} + \eta_{i,V} + \eta_{i,O} \qquad (3.1) \\ &= \beta_0 + \sum_{j \in \mathcal{D}} x_{ij}\beta_{j,L} + \sum_{j \in \mathcal{S}} f_{(j)}(x_{ij}) + \sum_{(j,k) \in \mathcal{D} \times \mathcal{S}} x_{ij} g_{(jk)}(x_{ik}) + \sum_{\substack{(j,k) \in \mathcal{S}^2 \\ j<k}} f_{(jk)}(x_{ij}, x_{ik}), \end{aligned}$$

wobei die Prädiktoranteile $\eta_{i,L}$, $\eta_{i,A}$, $\eta_{i,V}$ und $\eta_{i,O}$ die linearen (GLM-), die univariat funktionalen (GAM-), die variierende Koeffizienten (VCM-) sowie die bivariat funktionalen (Oberflächen-) Komponenten repräsentieren. Werden für $j \in \mathcal{D}$ darüber hinaus $f_{(j)}(x_{ij}) = x_{ij}\beta_{j,L}$ und

$$f_{(jk)}(x_{ij}, x_{ik}) = \begin{cases} 0, & \text{falls } k \in \mathcal{D}, \\ x_{ij} g_{(jk)}(x_{ik}), & \text{falls } k \in \mathcal{S}, \end{cases}$$

vereinbart, läßt sich (3.1) auch schreiben als

$$\eta_{i,U} = \beta_0 + \sum_{j=1}^{p} f_{(j)}(x_{ij}) + \sum_{j=1}^{p-1} \sum_{k>j} f_{(jk)}(x_{ij}, x_{ik}) := \beta_0 + \eta_{i,H} + \eta_{i,I}, \quad (3.2)$$

mit $\eta_{i,H} = \eta_{i,L} + \eta_{i,A}$ und $\eta_{i,I} = \eta_{i,V} + \eta_{i,O}$. Diese Schreibweise des universellen Prädiktors unterscheidet lediglich zwischen den Haupteffekten ($\eta_{i,H}$) und den funktionalen Interaktionen zwischen je zwei Kovariablen ($\eta_{i,I}$).

Die Schätzung der unbekannten Parameter und der funktionalen Komponenten in $\boldsymbol{\eta}_U' = (\eta_{1,U}, \ldots, \eta_{N,U})$ erfolgt über eine Maximierung der penalisierten Log-Likelihood

$$pl(\boldsymbol{\eta}_U) = l(\boldsymbol{\eta}_U) - \frac{1}{2} \cdot \left(P_A + P_V + P_O^{(1)} + P_O^{(2)} \right). \quad (3.3)$$

Eine Definition der einzelnen Größen, insbesondere der Summanden des Bestrafungsblocks, wird in den sich anschließenden Abschnitten gegeben, in denen für die Prädiktoranteile mit nonparametrischen Komponenten der propagierte Modellierungsansatz in Basisfunktionen eingehender untersucht wird.

3.2 Die GLM – Komponente

Der lineare Anteil der universellen Struktur (3.1) wird hier nur aus Gründen der Vollständigkeit nochmals aufgeführt. Setze $Z_L(i,j) := x_{ij}$, $j \in \mathcal{D}$, für eine Matrix $Z_L \in M_{I\!R}(N, |\mathcal{D}|)$ und definiere $\boldsymbol{\beta}_L := (\beta_{1,L}, \ldots, \beta_{|\mathcal{D}|,L})'$. Damit ist die globale GLM–Komponente $\boldsymbol{\eta}_L = (\eta_{1,L}, \ldots, \eta_{N,L})'$ als Produkt $Z_L\boldsymbol{\beta}_L$ darstellbar und erfüllt die im Abschnitt 1.1 definierte strukturelle Annahme (ii) eines generalisierten linearen Modells.

3.3 Die GAM – Komponente

Für $j \in \mathcal{S}$ definieren wir $[a_j, b_j] := [\min_i x_{ij}, \max_i x_{ij}]$. Ausgehend von einer äquidistanten Zerlegung Ω_{M_j} des Intervalls $[a_j, b_j]$, wird die unbekannte Regressionsfunktion $f_{(j)}$ analog zu (2.18) durch einen Polynom-Spline aus der Menge $S_{n_j}(\Omega_{M_j})$ approximiert

$$f_{(j)}(x_{ij}) \;=\; \sum_{m=-n_j}^{M_j-1} \beta_{jm,A} B_{jn_jm}(x_{ij}), \;\; i = 1, \ldots, N, \; j \in \mathcal{S}.$$

Auch wenn grundsätzlich für jede der nonparametrischen Komponenten die Anzahl der (inneren) Knoten und der Grad der B-Splines verschieden sein können, soll die Abhängigkeit dieser Größen von $j \in \mathcal{S}$ zugunsten einer besseren Lesbarkeit aufgegeben werden und global $M_j \equiv M$ und $n_j \equiv n$ gelten. Eine weitere notationelle Vereinfachung resultiert aus der Transformation der Indexmenge $\{-n, \ldots, M-1\}$ zu $\{1, \ldots, P := M+n\}$ und dem Verzicht auf die explizite Ausweisung des B-Spline-Grades

$$f_{(j)}(x_{ij}) \;=\; \sum_{s=1}^{P} \beta_{js,A} B_{js}(x_{ij}), \;\; i = 1, \ldots, N, \; j \in \mathcal{S}. \tag{3.4}$$

Werden für $j \in \mathcal{S}$ die Einträge der partiellen Designmatrizen $Z_j \in M_{I\!R}(N, P)$ via $Z_j(i,s) := B_{js}(x_{ij})$ definiert, und setzt man $\boldsymbol{\beta}_{j,A} := (\beta_{j1,A}, \ldots, \beta_{jP,A})'$, so läßt sich die GAM–Komponente auch schreiben als

$$\eta_{i,A} \;=\; \sum_{j \in \mathcal{S}} f_{(j)}(x_{ij}) \;=\; \sum_{j \in \mathcal{S}} Z_j(i, \cdot)\, \boldsymbol{\beta}_{j,A}\,, \;\; i = 1, \ldots, N,$$

und entspricht damit der für GLM's üblichen Prädiktorstruktur. Die globale GAM–Komponente $\boldsymbol{\eta}_A = (\eta_{1,A}, \ldots, \eta_{N,A})'$ ergibt sich zu $\boldsymbol{\eta}_A = Z_A \boldsymbol{\beta}_A$, wobei $Z_A := \left[Z_{\tilde{p}} \,|\, \ldots \,|\, Z_p \right]$ die (gesamte) Designmatrix und $\boldsymbol{\beta}'_A := (\boldsymbol{\beta}'_{\tilde{p},A}, \ldots, \boldsymbol{\beta}'_{p,A})$ den gesamten Parametervektor bezeichnen.

Differenzenpenalties zur Kontrolle der Variabilität in den geschätzten nonparametrischen Komponenten lassen sich analog zu Abschnitt 2.4 über die Kontrastmatrizen D_P^d definieren

$$\Delta^d \boldsymbol{\beta}_{j,A} \;=\; D_P^d\, \boldsymbol{\beta}_{j,A} \quad \text{und} \quad P^d_{\Delta_j,A} \;=\; \sum_{s=1}^{P-d} \left(\Delta^d \beta_{js,A} \right)^2 \;=\; \boldsymbol{\beta}'_{j,A} \left(D_P^d \right)' D_P^d\, \boldsymbol{\beta}_{j,A},$$

wobei auch hier auf eine effektspezifische Definition der Differenzenordnung verzichtet wurde. Bezeichnet $\lambda_{j,A}$ den zu $P^d_{\Delta_j,A}$ gehörigen Glättungsparameter, so ist mit $\Lambda_A = \mathrm{diag}(\lambda_{\tilde{p},A}\mathbf{1}'_P, \ldots, \lambda_{p,A}\mathbf{1}'_P)$ sowie

$$D_A = \mathrm{Diag}\big(D^d_P, \ldots, D^d_P\big) = I_{|\mathcal{S}|} \otimes D^d_P \quad \text{und} \quad K_A = \Lambda_A D'_A D_A$$

für die globale Penalty-Matrix, der erste Teil des Bestrafungsblocks in (3.3) durch $P_A = \boldsymbol{\beta}'_A K_A \boldsymbol{\beta}_A$ spezifiziert.

Das Vorliegen metrischer Kovariablen muß nicht zwangsläufig deren Berücksichtigung in der GAM-Komponente zur Folge haben. Ebenso denkbar sind Modelle, in denen die Effekte stetiger Einflußfaktoren rein parametrisch oder lediglich in Form von Interaktionen wirken. Der modulare Charakter der Designmatrix Z_A, des Parametervektors $\boldsymbol{\beta}_A$ sowie des Penalties K_A ermöglicht in diesen Situationen eine einfache Beschränkung der GAM-Komponente auf entsprechende Teilmengen von $\mathcal{S}$.

3.3.1 Identifikationsprobleme und Singularitäten

Die in Abschnitt 2.2 als „Zerlegung der Einheit“ bezeichnete Normierungseigenschaft einer B-Spline-Basis wird für $j \in \mathcal{S}$ mit den obigen Deklarationen durch die Aussage $Z_j\mathbf{1}_P = \mathbf{1}_N$ reflektiert. Für zwei Indizes $k, l \in \mathcal{S}$ und eine Konstante $c \neq 0$ ist damit

$$\boldsymbol{\eta}_A = \ldots + Z_k\boldsymbol{\beta}_{k,A} + Z_l\boldsymbol{\beta}_{l,A} + \ldots = \ldots + Z_k\boldsymbol{\beta}^*_{k,A} + Z_l\boldsymbol{\beta}^*_{l,A} + \ldots,$$

wobei $\boldsymbol{\beta}^*_{k,A} = \boldsymbol{\beta}_{k,A} + c \cdot \mathbf{1}_P$ und $\boldsymbol{\beta}^*_{l,A} = \boldsymbol{\beta}_{l,A} - c \cdot \mathbf{1}_P$. Die zu $f_{(j)}$, $j \in \mathcal{S}$, gehörigen Basiskoeffizienten sind also nur bis auf eine additive Konstante eindeutig bestimmbar.

Damit vererbt sich die bereits in Abschnitt 1.2 geschilderte Problematik der Nichtidentifizierbarkeit der Effekte im GAM auf die Koeffizienten der zugehörigen Basisdarstellungen. Für die Schätzer $\hat{f}_{(j)}$ heißt das, daß sie zwar in Form und Gestalt, nicht aber hinsichtlich ihrer vertikalen Position eindeutig bestimmt sind.

Weitaus schwerwiegender sind die Konsequenzen für die Existenz von Schätzern. Es ist

$$rg(Z_A) \;=\; \sum_{j\in\mathcal{S}} P - |\mathcal{S}| + 1 \;=\; |\mathcal{S}|\cdot P - \big(|\mathcal{S}| - 1\big),$$

d.h. für $|\mathcal{S}| \geq 2$ besitzt Z_A keinen vollen Spaltenrang, und dies führt unweigerlich zu Singularitäten im Schätzalgorithmus zur Bestimmung der Modellparameter.

Die Eindeutigkeit der Basiskoeffzienten (und der mit ihnen korrespondierenden Effekte) läßt sich über folgende Restriktion garantieren. Postuliert man

$$\sum_{s=1}^{P} \beta_{js,A} \;=\; 0\,,\; j \in \mathcal{S}, \tag{3.5}$$

wird das angesprochene Verschieben additiver Konstanten zwischen den Basiskoeffizienten unterbunden.

Mit obigem Postulat läßt sich je ein Basiskoeffizient durch die verbleibenden ausdrücken und wird damit für die Betrachtungen redundant. Löst man (3.5) o.B.d.A. nach dem letzten Koeffizienten auf,

$$\beta_{jP,A} \;=\; -\beta_{j1,A} - \ldots - \beta_{j,P-1,A}\,,\; j \in \mathcal{S},$$

beschränkt sich der Vektor $\tilde{\boldsymbol{\beta}}_{j,A}$ der relevanten Koeffizienten auf die Komponenten $\beta_{j1,A}, \ldots, \beta_{j,P-1,A}$, und es ist

$$Z_j\boldsymbol{\beta}_{j,A} \;=\; Z_j \begin{pmatrix} \tilde{\boldsymbol{\beta}}_{j,A} \\ -\mathbf{1}'_{P-1}\tilde{\boldsymbol{\beta}}_{j,A} \end{pmatrix} \;=\; Z_j \begin{bmatrix} I_{P-1} \\ -\mathbf{1}'_{P-1} \end{bmatrix} \tilde{\boldsymbol{\beta}}_{j,A} \;=:\; \tilde{Z}_j\tilde{\boldsymbol{\beta}}_{j,A}\,,$$

mit modifiziertem Design $\tilde{Z}_j$, das aus Z_j durch Subtraktion der letzten Spalte $Z_j(\cdot, P)$ von allen übrigen Spalten und der anschließenden Streichung von $Z_j(\cdot, P)$ hervorgeht. Die reduzierte Koeffizientenzahl bewirkt demzufolge eine entsprechende Dimensionsreduktion der partiellen Designmatrizen. Darüber hinaus unterliegen die $\tilde{Z}_j$ nicht der Normierungseigenschaft (2.9), so daß ihre Zeilensummen im allgemeinen nicht konstant sind. Für die gesamte Designmatrix $\tilde{Z}_A := \big[\tilde{Z}_{\tilde{p}} \mid \ldots \mid \tilde{Z}_p\big]$ läßt sich mit den Restriktionen (3.5) daher auch der modellierungsbedingte Rangabfall vermeiden.

Neben einer Berücksichtigung von (3.5) im Design erfordert die Reduktion der Basiskoeffizienten eine entsprechende Modifikation der Strafterme. Für

die Differenzen $D_P^d \boldsymbol{\beta}_{j,A}$, $j \in \mathcal{S}$, läßt sich analog folgendes ableiten

$$D_P^d \boldsymbol{\beta}_{j,A} \;=\; \tilde{D}_P^d \tilde{\boldsymbol{\beta}}_{j,A}\,, \text{ mit } \tilde{D}_P^d := D_P^d \cdot \left[I_{P-1} \,|\, -\mathbf{1}_{P-1}\right]'.$$

Die Ausführungen dieses Abschnittes basierten auf der Annahme des Vorliegens mindestens zweier nonparametrischer Komponenten. Im Falle nur eines Haupteffektes treten die angesprochenen Identifikationsprobleme zwar nicht mehr innerhalb der GAM–Komponente auf, wohl aber zwischen den Basiskoeffizienten von $f_{(\tilde{p})}$ und dem Intercept β_0. Da ferner $rg\big([\mathbf{1}_N \,|\, Z_{\tilde{p}}]\big) = P$ ist, gilt gleiches für die Aussagen zur Singularität, so daß die vorgestellten Modifikationen auch für $|\mathcal{S}| = 1$ relevant sind.

3.4 Die VCM – Komponente

In Analogie zum vorhergehenden Abschnitt werden für $(j,k) \in \mathcal{D} \times \mathcal{S}$ die effektmodifizierenden Funktionen $g_{(jk)}$ in $\eta_{i,V}$ als Basisrepräsentationen von Polynom-Splines modelliert. Wird für jedes $g_{(jk)}$ dieselbe äquidistante Zerlegung und derselbe B-Spline-Grad wie bei der Modellierung des entsprechenden Haupteffektes $f_{(k)}$ in der GAM-Komponente zugrunde gelegt, lassen sich die Deklarationen aus Abschnitt 3.3 übernehmen, und es ist

$$g_{(jk)}(x_{ik}) \;=\; \sum_{s=1}^{P} \beta_{jks,V} B_{ks}(x_{ik}) \quad \text{bzw.} \quad \eta_{i,V} \;=\; \sum_{(j,k)\in\mathcal{D}\times\mathcal{S}} x_{ij} Z_k(i,\cdot)\, \boldsymbol{\beta}_{jk,V} \tag{3.6}$$

für $i = 1,\ldots,N$ und $\boldsymbol{\beta}_{jk,V} := (\beta_{jk1,V},\ldots,\beta_{jkP,V})'$. Deklariert man darüber hinaus $X_j = \mathrm{diag}(x_{1j},\ldots,x_{Nj})$, $j \in \mathcal{D}$, so läßt sich die globale VCM-Komponente $\boldsymbol{\eta}_V = (\eta_{1,V},\ldots,\eta_{N,V})'$ in kompakter Form als $\boldsymbol{\eta}_V = Z_V \boldsymbol{\beta}_V$ schreiben, mit

$$Z_V \;=\; \Big[X_1 Z_{\tilde{p}} \,\Big|\, X_1 Z_{\tilde{p}+1} \,\Big|\, \ldots \,\Big|\, X_{\tilde{p}-1} Z_p\Big] \quad \text{und} \quad \boldsymbol{\beta}_V \;=\; \big(\boldsymbol{\beta}'_{1\tilde{p},V},\ldots,\boldsymbol{\beta}'_{\tilde{p}-1,p,V}\big)'.$$

Damit existiert also auch für die (globale) VCM–Komponente eine Darstellung in der für GLM's typischen Schreibweise. Die Vereinbarungen

$$D_V \;=\; \mathrm{Diag}\big(D_P^d,\ldots,D_P^d\big) \;=\; I_{|\mathcal{D}|\cdot|\mathcal{S}|} \otimes D_P^d \quad \text{und} \quad K_V \;=\; \Lambda_V D_V' D_V$$

für die Penalty-Matrix sowie $\Lambda_V = \mathrm{diag}(\lambda_{1\tilde{p},V}\mathbf{1}'_P,\ldots,\lambda_{\tilde{p}-1,p,V}\mathbf{1}'_P)$ für die zu-

gehörige Matrix der Glättungsparameter definieren mit $P_V = \boldsymbol{\beta}_V' K_V \boldsymbol{\beta}_V$ den zweiten Teil des Bestrafungsblocks von (3.3). Die Modularität der deklarierten Größen erlaubt auch hier wieder eine einfache Einschränkung der modellierten Interaktionen auf Teilmengen von $\mathcal{D} \times \mathcal{S}$.

3.4.1 Identifikationsprobleme und Singularitäten

Die Normierungseigenschaft (2.9) äußert sich für $(j,k) \in \mathcal{D} \times \mathcal{S}$ in der Tatsache, daß $X_j Z_k \mathbf{1}_P = X_j \mathbf{1}_N$. Für $j \in \mathcal{D}$ und $k,l \in \mathcal{S}$ können damit additive Konstanten zwischen den Parametervektoren $\boldsymbol{\beta}_{jk,V}$ und $\boldsymbol{\beta}_{jl,V}$ ausgetauscht werden, ohne den Wert der VCM–Komponente zu verändern. Berücksichtigt man die Kovariable x_j mit einem fixen Parameter $\beta_{j,L}$ sowie einem variierenden Koeffizienten, so ist ferner

$$\begin{aligned} \boldsymbol{\eta}_L + \boldsymbol{\eta}_V &= \ldots + X_j \beta_{j,L} \mathbf{1}_N + \ldots + X_j Z_k \boldsymbol{\beta}_{jk,V} + \ldots \\ &= \ldots + X_j \beta^*_{j,L} \mathbf{1}_N + \ldots + X_j Z_k \boldsymbol{\beta}^*_{jk,V} + \ldots , \end{aligned}$$

für $\beta^*_{j,L} = \beta_{j,L} + c$, $\boldsymbol{\beta}^*_{jk,V} = \boldsymbol{\beta}_{jk,V} - c \cdot \mathbf{1}_P$ und eine Konstante $c \neq 0$. Identifikationsprobleme treten somit innerhalb der VCM–Komponente als auch zwischen $\boldsymbol{\eta}_V$ und dem linearen Anteil $\boldsymbol{\eta}_L$ auf. Darüber hinaus verursacht Eigenschaft (2.9) Singularitäten im Schätzalgorithmus, da Z_V für $|\mathcal{S}| \geq 2$ und die Matrix $[Z_L \,|\, Z_V]$ keinen vollen Spaltenrang besitzen.

Identifikationsprobleme und Singularitäten lassen sich analog zu 3.3.1 durch entsprechende Restriktionen der Basiskoeffizienten umgehen

$$\sum_{s=1}^{P} \beta_{jks,V} = 0 \quad \text{bzw.} \quad \beta_{jkP,V} = -\sum_{s=1}^{P-1} \beta_{jks,V} \,, \; (j,k) \in \mathcal{D} \times \mathcal{S},$$

mit den bereits in Abschnitt 3.3.1 gegebenen Modifikationen für die partiellen Designmatrizen Z_k, die Parametervektoren $\boldsymbol{\beta}_{jk,V}$ und die Differenzenmatrizen D^d_P.

3.5 Die Oberflächen – Komponente

Die Modellierung von funktionalen Interaktionen stetiger Kovariablen, sogenannten Oberflächeneffekten, erfordert eine Erweiterung des Basisfunktions-

konzepts auf den mehrdimensionalen Fall. In dieser Arbeit beschränken sich die Betrachtungen auf funktionale Interaktionen $f_{(jk)}$ in zwei Argumenten x_j und x_k, $j, k \in \mathcal{S}$. Analog zum eindimensionalen Fall unterstellt der hier propagierte Modellierungsansatz die Approximierbarkeit des unbekannten Oberflächeneffekts durch eine bekannte Funktionenklasse.

Im einzelnen wird angenommen, daß sich der zu modellierende Oberflächeneffekt als bivariater Polynom–Spline bzw. genauer als Element des (Tensorprodukt-) Raumes $S_n(\Omega_M^j) \otimes S_n(\Omega_M^k)$ darstellen läßt

$$f_{(jk)}(x_{ij}, x_{ik}) = \sum_{s=1}^{P} \sum_{t=1}^{P} \beta_{jkst,O} B_{js}(x_{ij}) B_{kt}(x_{ik}), \quad i = 1, \ldots, N, \tag{3.7}$$

vgl. Hämmerlin & Hoffmann (1992) bzw. Dierckx (1993). Für jede Schnittfunktion von $f_{(jk)}$ durch einen Punkt $x_{0k} \in [a_k, b_k]$ parallel zur x_j–Achse gilt dann

$$f^{(j)}(\cdot) = f_{(jk)}(\cdot, x_{0k}) = \sum_{s=1}^{P} \beta_{js} B_{js}(\cdot) \quad \text{mit} \quad \beta_{js} = \sum_{t=1}^{P} \beta_{jkst,O} B_{kt}(x_{0k}).$$

Analog gilt für jede Schnittfunktion durch einen Punkt $x_{0j} \in [a_j, b_j]$ parallel zur x_k–Achse

$$f^{(k)}(\cdot) = f_{(jk)}(x_{0j}, \cdot) = \sum_{t=1}^{P} \beta_{kt} B_{kt}(\cdot) \quad \text{mit} \quad \beta_{kt} = \sum_{s=1}^{P} \beta_{jkst,O} B_{js}(x_{0j}),$$

d.h. die mit (3.7) getroffenen Annahmen sind gleichbedeutend mit der Aussage, daß sich jede Schnittfunktion von $f_{(jk)}$ parallel zur x_j– bzw. x_k–Achse als Basisrepräsentation eines (eindimensionalen) Polynom-Splines darstellen läßt. Werden in (3.7) die B-Spline-Produkte zu den bivariaten Funktionen

$$B_{jkst}(x_{ij}, x_{ik}) := B_{js}(x_{ij}) B_{kt}(x_{ik})$$

zusammengefasst, lassen sich zweidimensionale B-Splines als Basiselemente des Funktionenraumes $S_n(\Omega_M^j) \otimes S_n(\Omega_M^k)$ definieren. Abbildung 3.1 zeigt bilineare bzw. biquadratische B-Splines zusammen mit ihren eindimensionalen Pendants.

Unter Verwendung der partiellen Designmatrizen aus Abschnitt 3.3 gewinnt

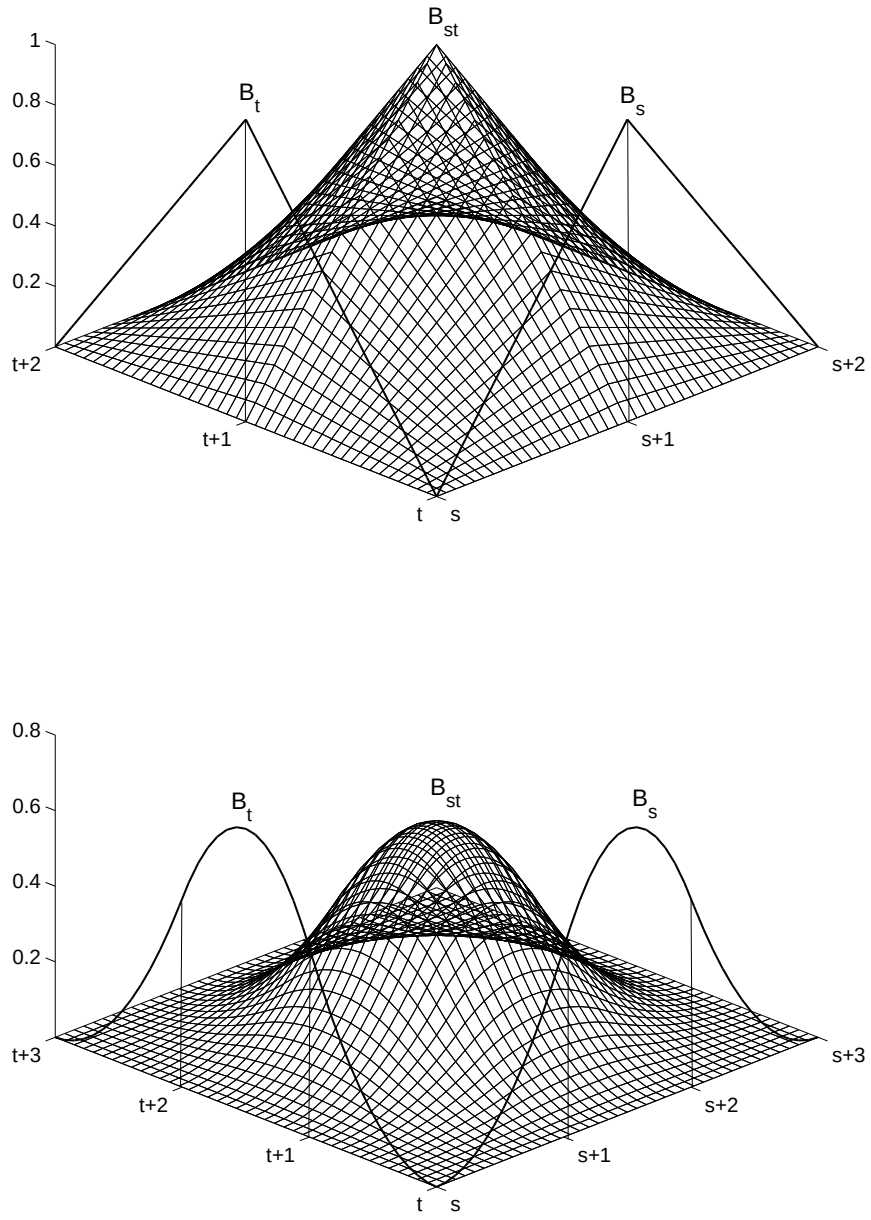

ABBILDUNG 3.1: *Geplottet sind die Graphen eines bilinearen und eines biquadratischen B-Splines über ihren jeweiligen Trägern. Zusätzlich dargestellt sind die definierendenden, eindimensionalen Faktoren.*

man ferner folgende kompakte (GLM-) Schreibweise für die (globale) Oberflächen–Komponente

$$\eta_{i,O} = \sum_{\substack{(j,k)\in\mathcal{S}^2 \\ j<k}} Z_j(i,\cdot) \otimes Z_k(i,\cdot)\, \boldsymbol{\beta}_{jk,O} \quad \text{bzw.} \quad \boldsymbol{\eta}_O = (\eta_{1,O},\ldots,\eta_{N,O})' = Z_O\, \boldsymbol{\beta}_O,$$

wobei

$$\boldsymbol{\beta}_{jk,O} = \left(\beta_{jk11,O}, \beta_{jk12,O}, \ldots, \beta_{jkPP,O}\right)', \quad \boldsymbol{\beta}'_O = \left(\boldsymbol{\beta}'_{\tilde{p},\tilde{p}+1,O}, \boldsymbol{\beta}'_{\tilde{p},\tilde{p}+2,O}, \ldots, \boldsymbol{\beta}'_{p-1,p,O}\right)'$$

und $Z_O = \left[\, Z_{\tilde{p}} \widetilde{\otimes} Z_{\tilde{p}+1} \mid Z_{\tilde{p}+1} \widetilde{\otimes} Z_{\tilde{p}+2} \mid \ldots \mid Z_{p-1} \widetilde{\otimes} Z_p \,\right]$. Der Operator $\widetilde{\otimes}$ definiert dabei das zeilenweise zu bildende Kronecker–Produkt (vgl. Anhang C). Vereinfachend wird im folgenden auch Z_{jk} statt $Z_j \widetilde{\otimes} Z_k$, $(j,k) \in \mathcal{S}^2$ geschrieben.

Analog zum eindimensionalen Fall soll die Variation in den geschätzten Oberflächeneffekten durch diskrete Bestrafungsterme kontrolliert werden. Zu diesem Zweck werden die Basiskoeffizienten in (3.7) gemäß ihrer Verlinkung mit den Knoten des quadratischen Gitters $\Omega_M^j \times \Omega_M^k$ in einer Matrix angeordnet

$$\begin{pmatrix} \beta_{jk1P,O} & \beta_{jk2P,O} & \cdots & \beta_{jkPP,O} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{jk12,O} & \beta_{jk22,O} & \cdots & \beta_{jkP2,O} \\ \beta_{jk11,O} & \beta_{jk21,O} & \cdots & \beta_{jkP1,O} \end{pmatrix}. \tag{3.8}$$

Formuliert man davon ausgehend die Glattheitsanforderungen an die Oberflächeneffekte als entsprechende Anforderungen an die axialen Schnittfunktionen, erhält man eine einfache Möglichkeit, die bekannten Penalisierungskonzepte auf die vorliegende Situation zu übertragen. Dazu werden separate Variationskontrollen jeweils entlang der achsenparallelen Linien des Gitters $\Omega_M^j \times \Omega_M^k$ durchgeführt. Aufgrund der speziellen Verknüpfung der Basiskoeffizienten mit den Gitterpunkten, ist dies äquivalent zum Wirken von Zeilen- und Spaltenpenalties in (3.8). Entsprechende Bestrafungsterme erster Ordnung sind durch die folgenden Ausdrücke gegeben

$$\sum_{t=1}^{P} \sum_{s=2}^{P} \left(\beta_{jkst,O} - \beta_{jk,s-1,t,O}\right)^2 = \boldsymbol{\beta}'_{jk,O} \left(D^1_{P,1}\right)' D^1_{P,1}\, \boldsymbol{\beta}_{jk,O}\,,$$

$$\sum_{s=1}^{P} \sum_{t=2}^{P} \left(\beta_{jkst,O} - \beta_{jks,t-1,O}\right)^2 = \boldsymbol{\beta}'_{jk,O} \left(D^1_{P,2}\right)' D^1_{P,2}\, \boldsymbol{\beta}_{jk,O}\,,$$

mit $D^1_{P,1} := D^1_P \otimes I_P$ und $D^1_{P,2} := I_P \otimes D^1_P$. Die Wirkungsweise dieser Penalties ist per Definition parallel zur x_j– bzw. x_k–Achse gerichtet und fungiert somit variationsbeschränkend für die axialen Schnitte.

Differenzenmatrizen beliebiger Ordnung lassen sich aus obigem über die Vorschriften $D^d_{P,1} = D^d_P \otimes I_P$ und $D^d_{P,2} = I_P \otimes D^d_P$ gewinnen. Definiert man ferner für Indizes $r \in \{1,2\}$ die richtungsgebundenen Strafmatrizen

$$D_{O,r} \;=\; \mathrm{Diag}\big(D^d_{P,r},\ldots,D^d_{P,r}\big) \;=\; I_{|\mathcal{S}|\cdot(|\mathcal{S}|-1)/2} \otimes D^d_{P,r} \;\text{ und }\; K^{(r)}_O \;=\; \Lambda^{(r)}_O D'_{O,r} D_{O,r}$$

mit $\Lambda^{(r)}_O = \mathrm{diag}\big(\lambda^{(r)}_{\tilde{p},\tilde{p}+1,O}\mathbf{1}'_{P^2},\ldots,\lambda^{(r)}_{p-1,p,O}\mathbf{1}'_{P^2}\big)$ als zugehörigen Matrizen der Glättungsparameter, so sind durch $P^{(r)}_O = \boldsymbol{\beta}'_O K^{(r)}_O \boldsymbol{\beta}_O$, $r \in \{1,2\}$, die verbliebenen zwei Strafterme in (3.3) erklärt. Wie schon für die anderen Prädiktorkomponenten gestattet die Streichung einzelner Module in den deklarierten Größen eine einfache Beschränkung der modellierten Oberflächeneffekte auf beliebige Teilmengen von $\{(j,k) \in \mathcal{S}^2 : j < k\}$.

3.5.1 Identifikationsprobleme und Singularitäten

Die Normierungseigenschaft (2.9) eindimensionaler B-Splines spielt auch bei der Modellierung von Oberflächeneffekten eine wichtige Rolle. Für jedes Paar $(j,k) \in \mathcal{S}^2$, $j < k$, ist

$$Z_{jk}\mathbf{1}_{P^2} \;=\; (Z_j \widetilde{\otimes} Z_k)\mathbf{1}_{P^2} \;=\; \mathbf{1}_N,$$

so daß auch die Basis des Tensorproduktraumes $S_n(\Omega^j_M) \otimes S_n(\Omega^k_M)$ eine Zerlegung der Einheit bildet. Die daraus erwachsenden Identifikationsprobleme und Singularitäten lassen sich analog zu Abschnitt 3.3.1 motivieren.

Bedingungen, die sowohl die Eindeutigkeit als auch die Existenz von Schätzern gewährleisten, können ebenso analog als Restriktionen an die Basiskoeffizienten formuliert werden

$$\sum_{s=1}^{P}\sum_{t=1}^{P} \beta_{jkst,O} \;=\; 0 \quad \text{bzw.} \quad \beta_{jkPP,O} \;=\; -\sum_{s=1}^{P-1}\sum_{t=1}^{P} \beta_{jkst,O} - \sum_{t=1}^{P-1} \beta_{jkPt,O}\,, \quad (3.9)$$

für alle $(j,k) \in \mathcal{S}^2$ mit $j < k$. Ähnlich wie im eindimensionalen Fall führen die Restriktionen zu einer Dimensionreduktion im Parametervektor, im De-

sign und in den richtungsgebundenen Strafmatrizen

$$\tilde{\boldsymbol{\beta}}_{jk,O} = \begin{pmatrix} \beta_{jk11,O} \\ \vdots \\ \beta_{jkP,P-1,O} \end{pmatrix} \quad , \quad \tilde{Z}_{jk} = Z_{jk} \cdot \begin{bmatrix} I_{P^2-1} \\ -\mathbf{1}'_{P^2-1} \end{bmatrix} \quad , \quad \tilde{D}^d_{P,r} = D^d_{P,r} \cdot \begin{bmatrix} I_{P^2-1} \\ -\mathbf{1}'_{P^2-1} \end{bmatrix}.$$

Die Restriktionen (3.9) verhindern zwar den Austausch globaler Konstanten zwischen $\boldsymbol{\eta}_A$ und $\boldsymbol{\eta}_O$, sie sind jedoch wirkungslos für Vektoren $\boldsymbol{c} \in I\!R^P \backslash \{\mathbf{0}\}$, die der Bedingung $\boldsymbol{c}'\mathbf{1}_P = 0$ genügen. Es ist

$$\begin{aligned} \boldsymbol{\eta}_A + \boldsymbol{\eta}_O &= \ldots + Z_j\boldsymbol{\beta}_{j,A} + \ldots + Z_k\boldsymbol{\beta}_{k,A} + \ldots + Z_{jk}\boldsymbol{\beta}_{jk,O} + \ldots \\ &= \ldots + Z_j\boldsymbol{\beta}^*_{j,A} + \ldots + Z_k\boldsymbol{\beta}_{k,A} + \ldots + Z_{jk}\boldsymbol{\beta}^*_{jk,O} + \ldots \\ &= \ldots + Z_j\boldsymbol{\beta}_{j,A} + \ldots + Z_k\boldsymbol{\beta}^{**}_{k,A} + \ldots + Z_{jk}\boldsymbol{\beta}^{**}_{jk,O} + \ldots , \end{aligned}$$

für Vektoren $\boldsymbol{\beta}^*_{j,A} = \boldsymbol{\beta}_{j,A} + \boldsymbol{c}$, $\boldsymbol{\beta}^{**}_{k,A} = \boldsymbol{\beta}_{k,A} + \boldsymbol{c}$, $\boldsymbol{\beta}^*_{jk,O} = \boldsymbol{\beta}_{jk,O} - \boldsymbol{c} \otimes \mathbf{1}_P$ sowie $\boldsymbol{\beta}^{**}_{jk,O} = \boldsymbol{\beta}_{jk,O} - \mathbf{1}_P \otimes \boldsymbol{c}$, da $Z_{jk}\,(\boldsymbol{c} \otimes \mathbf{1}_P) = Z_j\boldsymbol{c}$ und $Z_{jk}\,(\mathbf{1}_P \otimes \boldsymbol{c}) = Z_k\boldsymbol{c}$.

Wegen $\boldsymbol{c}'\mathbf{1}_P = 0$ sind (3.5) bzw. (3.9) auch für $\boldsymbol{\beta}^*_{j,A}$ und $\boldsymbol{\beta}^{**}_{k,A}$ bzw. $\boldsymbol{\beta}^*_{jk,O}$ und $\boldsymbol{\beta}^{**}_{jk,O}$ erfüllt, so daß sich obige Identifikationsprobleme mit diesen Restriktionen nicht abfangen lassen. Durch die Einführung von Restriktionen, die bezüglich (3.8) zeilen- bzw. spaltenspezifisch sind, ließe sich obige Konstantenverschiebung verhindern. Wird alternativ jeder Oberflächeneffekt mit je Q^2, $Q \neq P$, zweidimensionalen B-Splines modelliert, erzielt man mit deutlich geringerem Aufwand die gewünschte Eindeutigkeit und kann sich auf die Restriktionen (3.5) und (3.9) beschränken.

Mit den Überlegungen der letzten Abschnitte läßt sich das universelle Design als $Z_U := \left[\mathbf{1}_N \,|\, Z_L \,|\, \tilde{Z}_A \,|\, \tilde{Z}_V \,|\, \tilde{Z}_O\right]$ definieren. Für $\boldsymbol{\beta}' := \left(\beta_0, \boldsymbol{\beta}'_L, \tilde{\boldsymbol{\beta}}'_A, \tilde{\boldsymbol{\beta}}'_V, \tilde{\boldsymbol{\beta}}'_O\right)$ ist $\boldsymbol{\eta}_U = Z_U\boldsymbol{\beta}$ und damit $l(\boldsymbol{\eta}_U) = l(\boldsymbol{\beta})$. Werden die einzelnen Strafterme darüber hinaus zu einem globalen Penalty

$$K = \mathrm{Diag}\Big(0_{\tilde{p}\times\tilde{p}}, \tilde{K}_A, \tilde{K}_V, \tilde{K}^{(1)}_O + \tilde{K}^{(2)}_O\Big)$$

zusammengefaßt, läßt sich die penalisierte Log-Likelihood (3.3) schreiben als $pl(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - 1/2 \cdot \boldsymbol{\beta}' K \boldsymbol{\beta}$. Die Differentiation nach $\boldsymbol{\beta}$ liefert die penalisierte Score–Funktion

$$ps(\boldsymbol{\beta}) = s(\boldsymbol{\beta}) - K\boldsymbol{\beta} = Z'_U D(\boldsymbol{\beta})\,\Sigma(\boldsymbol{\beta})^{-1}\big(\boldsymbol{y} - \boldsymbol{\mu}_U\big) - K\boldsymbol{\beta},$$

wobei $D(\boldsymbol{\beta})$ und $\Sigma(\boldsymbol{\beta})$ analog zu Abschnitt 1.1.1 definiert sind. Die Schätzgleichungen $ps(\boldsymbol{\beta}) = \mathbf{0}$ lassen sich wiederum iterativ via Fisher-Scoring lösen

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \tilde{F}\big(\hat{\boldsymbol{\beta}}^{(k)}\big)^{-1} \cdot ps\big(\hat{\boldsymbol{\beta}}^{(k)}\big), \quad k = 0, 1, \ldots \tag{3.10}$$

mit der Pseudo-Fisher-Matrix

$$\tilde{F}(\boldsymbol{\beta}) = E\left[-\frac{\partial ps(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right] = Z_U' D(\boldsymbol{\beta})\, \Sigma(\boldsymbol{\beta})^{-1} D'(\boldsymbol{\beta})\, Z_U + K. \tag{3.11}$$

Die in Abschnitt 1.1 gegebene Formalisierung des Fisher-Scoring zum Schätzen in GLM's kann also auch für die hier definierte universelle Prädiktorstruktur herangezogen werden. Einzige Modifikation gegenüber dem in Paragraph 1.1.1 beschriebenen Algorithmus ist die Ersetzung der erwarteten Fisher-Matrix durch die in (3.11) gegebene Pseudo-Form. Als gewichtete KQ-Schätzung nimmt (3.10) dabei folgende Gestalt an

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \tilde{F}\big(\hat{\boldsymbol{\beta}}^{(k)}\big)^{-1} Z_U'\, W\big(\hat{\boldsymbol{\beta}}^{(k)}\big)\, \tilde{\boldsymbol{\eta}}_U^{(k)},$$

wobei die Spezifikation der einzelnen Größen, sofern hier nicht angegeben, in analoger Entsprechung dem Abschnitt 1.1.1 zu entnehmen ist. Im Zeitpunkt der Konvergenz stimmen aktuelle Schätzung und Nachfolgeiterierte überein, so daß Abhängigkeiten vom Iterationszähler unberücksichtigt bleiben können

$$\hat{\boldsymbol{\beta}} = \tilde{F}\big(\hat{\boldsymbol{\beta}}\big)^{-1} Z_U'\, W\big(\hat{\boldsymbol{\beta}}\big)\, \tilde{\boldsymbol{\eta}}_U\,.$$

Für den gefitteten universellen Prädiktor $\hat{\boldsymbol{\eta}}_U = Z_U \hat{\boldsymbol{\beta}}$ folgt daraus

$$\hat{\boldsymbol{\eta}}_U = Z_U\, \tilde{F}\big(\hat{\boldsymbol{\beta}}\big)^{-1} Z_U'\, W\big(\hat{\boldsymbol{\beta}}\big)\, \tilde{\boldsymbol{\eta}}_U =: H_U\, \tilde{\boldsymbol{\eta}}_U\,, \tag{3.12}$$

wobei die *Hatmatrix* $H_U = Z_U \tilde{F}\big(\hat{\boldsymbol{\beta}}\big)^{-1} Z_U' W\big(\hat{\boldsymbol{\beta}}\big)$ die Projektionsabbildung des adjustierten auf den gefitteten universellen Prädiktor repräsentiert. Der Hatmatrix kommt im sich anschließenden Kapitel eine wichtige Rolle als approximatives Maß für die Modellkomplexität zu.

Die in den einzelnen Abschnitten wiederholt angesprochenen Identifikationsprobleme und Singularitäten werden auch von Eilers & Marx (2002) thematisiert. Im Unterschied zur hier vorgestellten Restriktion der Basiskoeffizienten

ergänzen die genannten Autoren den globalen Bestrafungsterm um einen zusätzlichen Ridge-Penalty

$$K \;=\; \mathrm{Diag}\Big(0_{\tilde{p}\times\tilde{p}}, K_A, K_V, K_O^{(1)} + K_O^{(2)}\Big) \;+\; \lambda^{*}\cdot I,$$

für eine geeignet dimensionierte Einheitsmatrix und einen Glättungsparameter λ^* der Größenordnung 10^{-6}. Obgleich mit deutlich weniger Aufwand verbunden, ist der Erfolg dieses Ansatzes stark datenabhängig und damit nicht in jeder Situation gesichert – im Gegensatz zu der hier vorgeschlagenen analytischen Lösung des Problems.

Zusammenfassend läßt sich folgendes Fazit ziehen: Die abgeleitete Rückführbarkeit der universellen Prädiktorstruktur (3.1) auf ein generalisiertes lineares Modell verbunden mit der einfachen Integration diskreter Differenzenpenalties erlaubt die Schätzung der unbekannten Modellparameter im Rahmen eines modifizierten Fisher–Scoring. Damit können insbesondere die bei Präsenz nonparametrischer Komponenten sonst nicht vermeidbaren Backfitting-Schritte eingespart werden.

4 Wahl der Glättungsparameter

Nachdem das vorherige Kapitel die B–Spline basierte Modellierung und penalisierte Schätzung nonparametrischer Komponenten des universellen Prädiktors zum Gegenstand hatte, widmen sich die nun folgenden Ausführungen der noch offenen Frage nach der Wahl der Glättungsparameter. Für deren Beantwortung wird zunächst die Wirkungsweise des Glättungsparameters in (2.22) für die Extremfälle keiner bzw. unendlich starker Penalisierung rekapituliert.

Für $\lambda_{\boldsymbol{\beta}} = 0$ reduziert sich das Optimierungsproblem (2.22) zur Schätzung der via (2.18) modellierten Regressionsfunktion f auf die Maximierung der unpenalisierten Log–Likelihood $l(\boldsymbol{\beta})$. Die Schätzgleichungen $s(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$ werden nach (1.4) für $\hat{\boldsymbol{\beta}}$ mit

$$y_i \;=\; \hat{\mu}_i \;=\; h\big(Z(i,\cdot)\,\hat{\boldsymbol{\beta}}\big) \;=\; h\big(\hat{f}(x_i)\big) \;\text{ bzw. }\; \hat{f}(x_i) \;=\; g(y_i)\,,\; i = 1,\ldots,N,$$

gelöst. Die Maximierung der Log–Likelihood $l(\boldsymbol{\beta})$ führt demnach zu einer Interpolation der Datenpunkte $\big(x_i, g(y_i)\big)_{i=1,\ldots,N}$. Die damit erzielte Reproduktion der transformierten Responsebeobachtungen $g(y_i)$ bedeutet zwar ein Minimum an Informationsverlust, resultiert jedoch gleichzeitig in einem sehr unruhigen Verlauf der Schätzung $\hat{\boldsymbol{f}} := \big(\hat{f}(x_1), \ldots, \hat{f}(x_N)\big)' = Z\hat{\boldsymbol{\beta}}$.

Für das entgegengesetzte Extrem unendlich starker Penalisierung ($\lambda_{\boldsymbol{\beta}} \longrightarrow \infty$) liefern die Propositionen 2.9 und 2.10 entsprechende Aussagen über den polynomialen Charakter der resultierenden Schätzung für die hier betrachtete Beziehung $n \geq d-1$ zwischen B–Spline–Grad und Differenzenordnung. Im Unterschied zur Dateninterpolation bedienen polynomiale Fits – als beliebig oft differenzierbare Funktionen – in weitaus stärkerem Maße natürliche Anforderungen an die Glattheit der nonparametrischen Effekte. Die restritive Annahme polynomialer Strukturen wirkt sich jedoch nachteilig auf die Darstellbarkeit komplexerer Zusammenhänge aus und schränkt die Flexibilität des Modellierungsansatzes (2.18) unnötig ein.

Gesucht ist daher ein Wert für $\lambda_{\boldsymbol{\beta}}$, der einen Kompromiss zwischen Flexibilitätsansprüchen einerseits und Glattheitsanforderungen andererseits schließt. Es stellt sich also zunächst die Frage nach einem Zielkriterium, welches diese Kompromissfähigkeit widerspiegelt, und auf dessen Basis $\lambda_{\boldsymbol{\beta}}$ optimiert werden kann.

4.1 Akaike – Informations – Kriterium

Die grundlegende Idee des Akaike-Informations-Kriteriums (AIC) basiert auf der Verknüpfung der Log-Likelihood $l(\hat{\boldsymbol{\beta}})$ des gefitteten Modells und der effektiven Anzahl gefitteter Parameter $\dim(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}})$ (Akaike, 1973)

$$\mathrm{AIC}(\lambda_{\boldsymbol{\beta}}) \;=\; -2 \cdot l(\hat{\boldsymbol{\beta}}) \;+\; 2 \cdot \dim(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}}). \tag{4.1}$$

Die effektive, sprich tatsächliche Anzahl gefitteter Parameter als Maß für die Modellkomplexität wird maßgeblich vom Glättungsparameter $\lambda_{\boldsymbol{\beta}}$ beeinflusst und stimmt nur bei unpenalisierter Schätzung ($\lambda_{\boldsymbol{\beta}} = 0$) mit der Anzahl $M+n$ der Modellparameter ($\hat{=}$ Anzahl der Basiskoeffizienten in Darstellung (2.18)) überein. Mit zunehmender Penalisierung der Basiskoeffizienten nimmt die effektive Anzahl gefitteter Parameter ab. Für $\lambda_{\boldsymbol{\beta}} \longrightarrow \infty$ resultiert mit den Aussagen in Abschnitt 2.5 ein polynomialer Fit der Ordnung $d-1$, die tatsächliche Anzahl gefitteter Parameter reduziert sich damit auf $d \ll M+n$. Allgemein gilt für $\lambda_{\boldsymbol{\beta}} \in [0, \infty)$

$$d \;<\; \dim(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}}) \;\leq\; M + n.$$

Mit den anfangs wiederholten Aussagen zum Einfluß der Penalisierungsstärke auf die Gestalt der Schätzung $\hat{\boldsymbol{f}}$ lassen sich nunmehr die folgenden Parallelen ziehen: Große Werte von $\dim(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}})$ korrespondieren mit Schätzungen, die stark variieren, kleine Werte umgekehrt mit sehr glatten Fits. Die effektive Anzahl gefitteter Parameter kann daher als quantitative Größe zur Bewertung der Glattheit einer Schätzung herangezogen werden. Die zusätzliche Abhängigkeit von der Log-Likelihood $l(\hat{\boldsymbol{\beta}})$ als Indikator für die Anpassungsgüte des Fits weist (4.1) somit als Kriterium aus, in dem Flexibilitäts- und Glattheitsanforderungen vereint sind. Durch die unterschiedlichen Vorzeichen können die konträren Anforderungen in (4.1) additiv verknüpft werden, und eine Minimierung von $\mathrm{AIC}(\lambda_{\boldsymbol{\beta}})$ führt zur angestrebten Kompromisslösung für den Glättungsparameter.

Obige Definition des Akaike-Informations-Kriteriums ist ferner äquivalent zu

$$\mathrm{AIC}(\lambda_{\boldsymbol{\beta}}) \;=\; \mathrm{dev}(\boldsymbol{y}, \hat{\boldsymbol{\beta}}) \;+\; 2 \cdot \dim(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}}), \tag{4.2}$$

wobei die *Devianz* $\mathrm{dev}(\boldsymbol{y}, \hat{\boldsymbol{\beta}})$ definiert ist als

$$\mathrm{dev}\big(\boldsymbol{y}, \hat{\boldsymbol{\beta}}\big) \;=\; -2 \cdot \sum_{i=1}^{N} \big\{ l_i(\hat{\mu}_i) - l_i(y_i) \big\} \;=\; -2 \cdot \big\{ l(\hat{\boldsymbol{\mu}}) - l(\boldsymbol{y}) \big\}. \qquad (4.3)$$

Dabei kennzeichnen die Summanden $l_i(\hat{\mu}_i) = l_i\big(\mu_i(\hat{\boldsymbol{\beta}})\big)$ die Log–Likelihoodbeiträge der einzelnen Beobachtungen nach Maximierung der penalisierten Log-Likelihood (2.22), wohingegen die Summanden $l_i(y_i)$ den Log–Likelihoodbeiträgen der Beobachtungen nach Maximierung von $l(\boldsymbol{\mu})$ ohne eine zusätzliche Modellannahme, d.h. nur unter Verwendung der Verteilungsannahme für den Response entsprechen.

Die Log-Likelihood in (4.1) bzw. die Devianz in (4.2) sind leicht zu berechnen. Für Glättungsparameter $\lambda_{\boldsymbol{\beta}} \in (0, \infty)$ kann die effektive Anzahl gefitteter Parameter jedoch nicht ohne weiteres quantifiziert werden. Die folgenden Überlegungen liefern eine Lösung dieses Problems: Für normalverteilten Response resultiert $\hat{\boldsymbol{f}} = Z(Z'Z)^{-1}Z'\boldsymbol{y}$ als Schätzung aus der Maximierung der unpenalisierten Log–Likelihood $l(\boldsymbol{\beta})$. Dabei ist

$$tr\big(Z(Z'Z)^{-1}Z'\big) = tr\big(Z'Z(Z'Z)^{-1}\big) = tr(I_{M+n}) = M + n = \dim\big(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}} = 0\big),$$

d.h. die tatsächliche Anzahl gefitteter Parameter ist gleich der Spur der Hatmatrix $H := Z(Z'Z)^{-1}Z'$. Erfolgt die Maximierung penalisiert mit $\lambda_{\boldsymbol{\beta}} > 0$, so ist

$$\hat{\boldsymbol{f}} \;=\; Z\big(Z'Z + \lambda_{\boldsymbol{\beta}}(D^d_{M+n})'D^d_{M+n}\big)^{-1}Z'\boldsymbol{y} \;=:\; H_\lambda \boldsymbol{y},$$

mit der Hatmatrix $H_\lambda = Z\big(Z'Z + \lambda_{\boldsymbol{\beta}}(D^d_{M+n})'D^d_{M+n}\big)^{-1}Z'$. Hastie & Tibshirani (1990) schlagen in Anlehnung an den unpenalisierten Fall $tr(H_\lambda)$ zur Approximation von $\dim\big(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}}\big)$ vor. Dieser Idee folgend, soll auch für nicht-normalverteilten Response die Spur der korrespondierenden Hatmatrix zur Approximation der tatsächlichen Anzahl gefitteter Parameter verwandt werden. Analog zu (3.12) definiert man als Hatmatrix $H_\lambda := Z\,\tilde{F}\big(\hat{\boldsymbol{\beta}}\big)^{-1}Z'\,W\big(\hat{\boldsymbol{\beta}}\big)$ und nutzt deren Spur

$$tr(H_\lambda) \;=\; tr\Big(Z\,\tilde{F}\big(\hat{\boldsymbol{\beta}}\big)^{-1}Z'\,W\big(\hat{\boldsymbol{\beta}}\big)\Big) \;=\; tr\Big(\tilde{F}\big(\hat{\boldsymbol{\beta}}\big)^{-1}F\big(\hat{\boldsymbol{\beta}}\big)\Big) \qquad (4.4)$$

zur Approximation von $\dim\big(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}}\big)$. Ist $\lambda_{\boldsymbol{\beta}} = 0$, gilt $\tilde{F}\big(\hat{\boldsymbol{\beta}}\big) = F\big(\hat{\boldsymbol{\beta}}\big)$ und damit $tr(H_\lambda) = tr(I_{M+n}) = M+n$, so daß im Fall unpenalisierter Schätzung die Approximation von $\dim\big(\hat{\boldsymbol{\beta}}, \lambda_{\boldsymbol{\beta}}\big)$ via (4.4) exakt ist. Definition (4.2) läßt sich mit

diesen Überlegungen zu

$$\mathrm{AIC}(\lambda_{\boldsymbol{\beta}}) \;=\; \mathrm{dev}\big(\boldsymbol{y}, \hat{\boldsymbol{\beta}}\big) \;+\; 2 \cdot tr(H_\lambda) \tag{4.5}$$

modifizieren. In vielen praktischen Anwendungen zeigt das AIC allerdings eine starke Tendenz zum *Unterglätten*, d.h. der (4.5) minimierende Glättungsparameter korrespondiert mit einem (zu) stark variierenden Fit. Für normalverteilten Response liefern Hurvich, Simonoff & Tsai (1998) eine verbesserte Version des AIC, die dieser Tendenz entgegenwirkt

$$\mathrm{AIC}_C(\lambda_{\boldsymbol{\beta}}) \;=\; \log\big(N^{-1}(\boldsymbol{y}-\hat{\boldsymbol{f}})'(\boldsymbol{y}-\hat{\boldsymbol{f}})\big) \;+\; 1 \;+\; 2 \cdot \frac{tr(H_\lambda)+1}{N - tr(H_\lambda) - 2}\;.$$

Alternative Ansätze, das Unterglätten zu verhindern, beruhen auf einer stärkeren Penalisierung der tatsächlichen Anzahl gefitteter Parameter. Dazu verallgemeinert man die rechte Seite in (4.5) zur Kriterienklasse

$$\mathrm{ZF}(\lambda_{\boldsymbol{\beta}}) \;=\; \mathrm{dev}\big(\boldsymbol{y}, \hat{\boldsymbol{\beta}}\big) \;+\; \gamma \cdot tr(H_\lambda) \tag{4.6}$$

für $\gamma \geq 1$. Werte von γ im Bereich von $1-4$ werden im Zeitreihenkontext von Bhansali & Downham (1977) betrachtet; für Prädiktionszwecke optimale γ's untersucht Atkinson (1980). Setzt man in (4.6) $\gamma := \log(N)$, so resultiert das *Bayesianische Informations-Kriterium (BIC)* (Schwarz, 1978)

$$\mathrm{BIC}(\lambda_{\boldsymbol{\beta}}) \;=\; \mathrm{dev}\big(\boldsymbol{y}, \hat{\boldsymbol{\beta}}\big) \;+\; \log(N) \cdot tr(H), \tag{4.7}$$

dessen Verwendung das Unterglätten in vielen Situationen reduziert.

4.2 Genetische Algorithmen

Die analytische Minimierung der in (4.6) definierten Kriterien ist nicht möglich, da der Einfluß des Glättungsparameters $\lambda_{\boldsymbol{\beta}}$ auf die Devianz und die effektive Anzahl gefitteter Parameter zwar offensichtlich ist, eine explizite Abhängigkeit in der Form eines funktionalen Zusammenhangs jedoch nicht angegeben werden kann. Die Praktikabilität numerischer Optimierungsverfahren wird hingegen stark vom Krümmungsverhalten des Zielkriteriums beeinflußt. Da mangels einer expliziten Abhängigkeit vom Glättungsparameter $\lambda_{\boldsymbol{\beta}}$ keine Angaben zur Konvexität von (4.6) möglich sind, birgt eine Verwendung numerischer Optimierungsstrategien zur Bestimmung von $\lambda_{\boldsymbol{\beta}}$ die Gefahr, ge-

gen ein lokales Minimum von (4.6) zu konvergieren. Viele der klassischen numerischen Algorithmen verlieren darüber hinaus an Effizienz, sobald mehrere Glättungsparameter simultan zu optimieren sind. In der Situation von Kapitel 3, in der jeder nonparametrische Effekt im universellen Prädiktor $\eta_{i,U}$ mit mindestens einem Glättungsparameter korrespondiert, entwickelt sich deren Bestimmung zu einem hoch-dimensionalen Optimierungsproblem. In Analogie zu Abschnitt 4.1 lassen sich dabei die Devianz des gefitteten Modells und die Spur der Hatmatrix als quantitative Größen zur Bewertung von Flexibilität und Komplexität bzw. Glattheit motivieren, so daß mit

$$\mathrm{ZF}(\mathcal{L}) \;=\; \mathrm{dev}\big(\boldsymbol{y}, \hat{\boldsymbol{\beta}}\big) \;+\; \gamma \cdot tr(H_U) \tag{4.8}$$

eine naheliegende Verallgemeinerung von (4.6) auf die Situation einer universellen Prädiktorstruktur gegeben ist. In der Menge $\mathcal{L}$ sind die Glättungsparameter aller modellierten nonparametrischen Effekte subsumiert, $\hat{\boldsymbol{\beta}}$ bezeichnet die Maximum-Likelihood-Schätzung des globalen Parametervektors und H_U die in (3.12) definierte Hatmatrix.

Auf der Suche nach alternativen numerischen Verfahren, deren Effizienz nur bedingt von der Anzahl zu optimierender Parameter bestimmt wird, und deren Konvergenzverhalten weitestgehend unabhängig von der Existenz lokaler Optima ist, haben sich in den letzten Jahren vermehrt sogenannte *evolutionäre Algorithmen* durchgesetzt. Diese Verfahren basieren auf der Evolutionstheorie Darwin's und orientieren sich bei der Definition algorithmischer Operatoren an entsprechenden biologischen Vorbildern.

Genetische Algorithmen als spezielle evolutionäre Verfahren wurden von Holland (1975) und Goldberg (1989) entwickelt und von Michalewicz (1996) im Kontext allgemeiner Optimierungsprobleme betrachtet. Im additiven Modell untersuchen Krause & Tutz (2003) die Verwendung genetischer Algorithmen zur Bestimmung optimaler Glättungsparameter. Da der von diesen Autoren vorgestellte Algorithmus unabhängig vom Regressionsmodell operiert, kann das Verfahren auch für die Situation eines universellen Prädiktors herangezogen werden. Die folgenden Ausführungen zum Algorithmus beschränken sich daher auf vorwiegend verbale Beschreibungen inhaltlicher Zusammenhänge. Für detailliertere Darstellungen sei auf Krause & Tutz (2003) verwiesen.

Genetische Algorithmen haben ihren Ursprung in der Evolutionstheorie, nach der die Überlebenschance eines Individuums mit seiner Fähigkeit steigt, sich herrschenden Bedingungen anzupassen. Selektion, Crossover und Mutation, die bestimmenden Faktoren des Evolutionsprozesses, werden bei genetischen Algorithmen durch stochastische Äquivalente (Operatoren) modelliert. Das zu optimierende Zielkriterium wird als Fitnessfunktion bezeichnet und definiert den Status der Anpassung – je größer die Fitness eines Individuums (korrespondierender Wert der Fitnessfunktion), desto größer seine Chance, den simulierten Ausleseprozeß zu überleben. In diesem Sinne ist die Optimierung eines Zielkriteriums stets als Maximierung desselben zu verstehen.

In der Terminologie der Genetik verbleibend, identifizieren wir jedes Individuum im folgenden mit einer konkreten Glättungsparameterkonstellation $\mathcal{L}$. Jede dieser Konstellationen definiert einen bestimmten Wert für $\mathrm{ZF}(\mathcal{L})$. Da optimale Glättungsparameter aber aus der Minimierung von (4.8) resultieren, ist die Fitnessfunktion des genetischen Algorithmus mit obigen Aussagen zum Selektionsprinzip als

$$\mathrm{FIT}(\mathcal{L}) \;=\; \mathrm{ZF}(\mathcal{L})^{-1} \;=\; \left(\mathrm{dev}\left(\boldsymbol{y}, \hat{\boldsymbol{\beta}}\right) \;+\; \gamma \cdot tr(H_U)\right)^{-1} \tag{4.9}$$

zu formulieren. Die natürliche Auslese als ein Bestandteil des Evolutionsprozesses kommt einer Wettbewerbssituation gleich und setzt die Existenz mehrerer konkurrierender Individuen, einer sogenannten Population bzw. Generation, voraus. Daher werden zu jedem Zeitpunkt der Evolution, d.h. in jedem Iterationsschritt t des genetischen Algorithmus, mehrere Glättungsparameterkonstellationen $\mathcal{L}_{t1}, \ldots, \mathcal{L}_{tG}$ simultan betrachtet. Ausgehend von einer zufällig gewählten Startpopulation $\mathcal{L}_{01}, \ldots, \mathcal{L}_{0G}$, die den Beginn des Evolutionsprozesses repräsentieren möge, werden schwache Individuen (Konstellationen mit kleiner Fitness) zunächst selektiert. Aus den verbleibenden Individuen wird eine Zwischenpopulation generiert, die nach selektiven Crossover- und Mutationsschritten schließlich in die Nachfolgegeneration $\mathcal{L}_{11}, \ldots, \mathcal{L}_{1G}$ übergeht. Dieser Vorgang wird über mehrere Zeitpunkte (Iterationsschritte) wiederholt, bis ein vordefiniertes Abbruchkriterium erfüllt ist.

Nachstehend werden die aufgezählten Schritte kurz erläutert. Die dabei beschriebenen Operatoren erfordern eine für jede Konstellation gleichbleibende

Abfolge der Glättungsparameter. Aus diesem Grund wird jedes Individuum $\mathcal{L}_{tj}$, $j = 1, \ldots, G$, einer Population als Tupel bzw. String arrangiert, in dem die Positionierung der Glättungsparameter gemäß ihrer Zugehörigkeit zu den nonparametrischen Effekten erfolgt. Eine derartige Lokalisierung hat mit den festen Genloci der Chromosomen ein entsprechendes genetisches Pendant.

Natürliche Auslese / Selektion:
Für einen vorgegebenen Prozentsatz p_s werden die $p_s\%$ schwächsten Individuen der aktuellen Population $\mathcal{L}_{t1}, \ldots, \mathcal{L}_{tG}$ gelöscht.

Generieren einer Zwischenpopulation:
Aus den restlichen $G \cdot (1 - p_s/100)$ Individuen wird über einen Auswahlmechanismus die Zwischenpopulation $\mathcal{L}^*_{t1}, \ldots, \mathcal{L}^*_{tG}$ gezogen. Die Auswahl gestattet Wiederholungen und ist so angelegt, daß die Wahrscheinlichkeit, in die Zwischenpopulation zu gelangen mit der Fitness eines Individuums ansteigt.

Crossover / Vererbung:
$G \cdot p_c/100 =: r$ zufällig gewählte Paare $(\mathcal{L}^*_{ti}, \mathcal{L}^*_{tj})$, $i, j \in \{1, \ldots, G\}$, $i \neq j$, der Zwischenpopulation werden über arithmetische Operatoren zu drei neuen Individuen kombiniert, von denen die beiden fittesten in die Nachfolgegeneration $t + 1$ eingehen.

Mutation:
Zufällig gewählte Elemente mehrfach auftretender Individuen der Zwischenpopulation werden mittels arithmetischer Operatoren leicht modifiziert. Dieser Schritt erhöht die Variabilität (Vielfalt) in der Nachfolgegeneration.

Für die Anwendungen der vorliegenden Arbeit, in denen die Optimierung der involvierten Glättungsparameter über den hier vorgestellten genetischen Algorithmus erfolgt, werden die folgenden Eckdaten zugrunde gelegt: Populationsgröße $G = 48$, Selektionsrate $p_s = 60$ und Crossoveranteil $p_c = 62$. Das in Krause & Tutz (2003) propagierte *Improved Arithmetical Crossover* realisiert den beschriebenen Vererbungs–Schritt. Für die Mutationswahrscheinlichkeit eines Elements identischer Individuen wird ein Wert von 0.25 angesetzt.

In Abbildung 4.1 sind die Vorgänge eines Iterationsschrittes nochmals schematisch dargestellt. Die differenzierte Behandlung der aus der Aufteilung der

Zwischenpopulation resultierenden Stränge verdeutlicht die zwei wesentlichen Zielstellungen des genetischen Algorithmus. Das Entstehen neuer Individuen im Crossover–Schritt, wie auch die Mutation identischer Individuen fördern die Variabilität in den Populationen und gewährleisten die *Exploration* großer Teile des Suchraumes. Die damit gegebene Lösungsvielfalt ermöglicht es, eventuelle lokale Optima der Zielfunktion wieder zu verlassen.

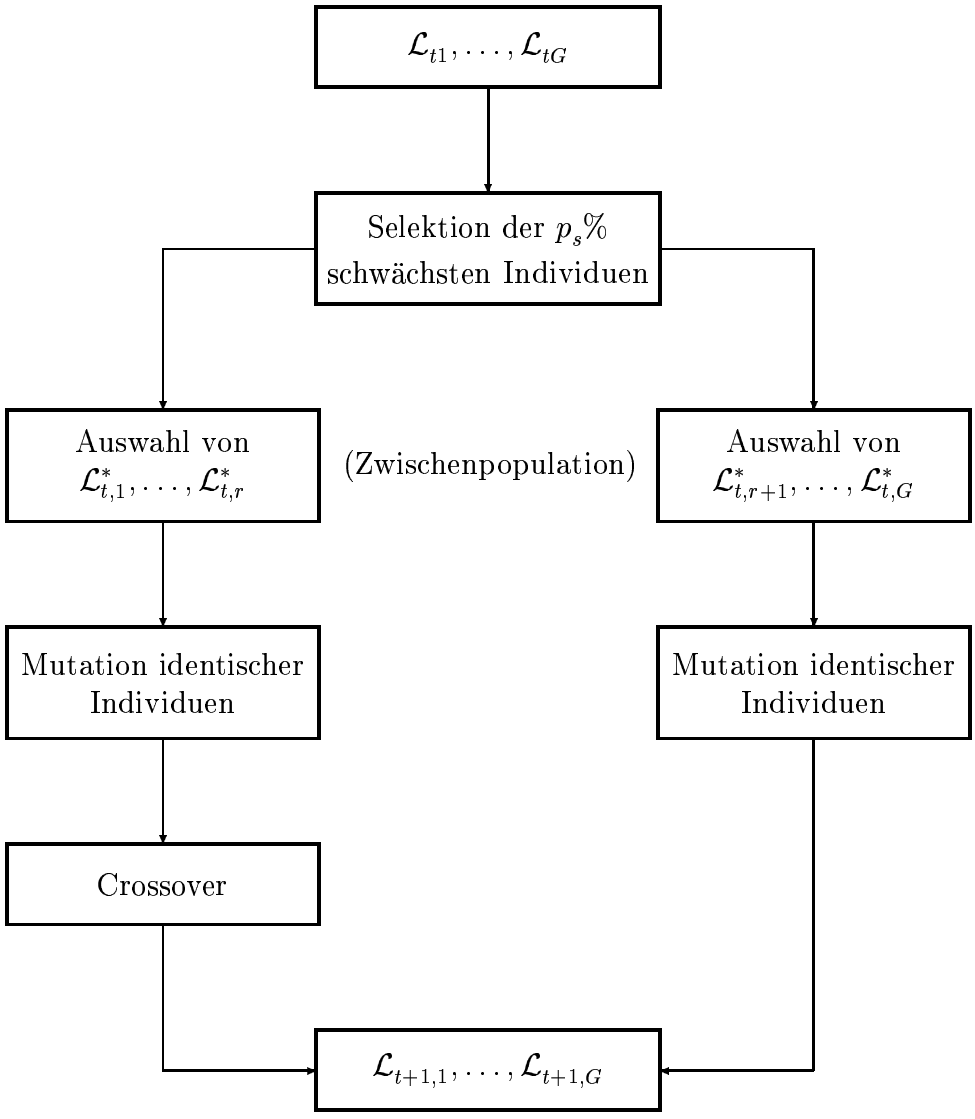

ABBILDUNG 4.1: *Schematischer Ablauf eines Iterationsschrittes im genetischen Algorithmus zur Optimierung der Glättungsparameter.*

Da lediglich identische Individuen mutiert werden, können Individuen einer Population über den rechten Strang unverändert in die Nachfolgegeneration eingehen. Die Chance, hierfür ausgewählt zu werden, steigt mit wachsender Fitness, so daß sich die Übernahmekandidaten mehrheitlich aus den fittesten Individuen rekrutieren. Beim Übergang zur Nachfolgegeneration werden somit stets auch Informationen über bereits durchsuchte, erfolgversprechende Regionen des Suchraumes vererbt, welche wiederum richtungsweisend für den Fortlauf der Suche sind. Die explizite Nutzung lokal vorhandener Informationen über die Fitnessfunktion, auch als *Exploitation* bezeichnet, repräsentiert die zweite wichtige Maßgabe genetischer Algorithmen.

Die entgegengesetzt gerichteten Zielstellungen Exploration und Exploitation werden in Abhängigkeit vom Iterationszähler unterschiedlich stark gewichtet. Während im Anfangsstadium der Schwerpunkt klar auf einem umfassenden Informationsgewinn liegt, verliert die Exploration von Iteration zu Iteration zunehmend an Gewicht. Gewährleistet wird dies durch eine iterationsabhängige Definition der Operatoren, deren abnehmende Wirkung die Variabilität in den Populationen mehr und mehr einschränkt. Aufeinanderfolgende Populationen werden sich immer ähnlicher, die Exploitation gewinnt an Bedeutung und erzwingt so die Konvergenz des Algorithmus in einer einheitlich fittesten Population.

Für die praktische Optimierung der Glättungsparameter werden die Iterationen jedoch aus Zeitgründen nicht bis zur endgültigen Konvergenz wiederholt. Der Iterationsprozeß wird abgebrochen, wenn die maximale Fitness in einer Population über eine vorgegebene Anzahl von Iterationen keine signifikanten Veränderungen aufweist.

5 Modelle mit nominalem Response

Während die bisherigen Betrachtungen ausschließlich Modellierungsaspekte auf Seiten des Prädiktors fokussierten, richtet sich das Augenmerk nunmehr auf eine Konkretisierung der Responsestruktur. In der vorliegenden Arbeit werden schwerpunktmäßig Modelle mit kategorialen Responsevariablen behandelt. Eine Variable wird dabei als kategorial bezeichnet, wenn die Anzahl ihrer möglichen Ausprägungen endlich ist. Abhängige Variablen dieses Typs finden sich in den verschiedensten Anwendungsbereichen. So wird in medizinischen Untersuchungen der Schweregrad einer Erkrankung häufig in den Abstufungen „leicht“, „mittel“ und „schwer“ erfaßt. Parteipräferenzen sowie die in den Kategorien „kurz-“, „mittel-“ und „langfristig“ gemessene Dauer von Arbeitslosigkeit sind nur zwei von zahllosen Beispielen aus Politik und Wirtschaft.

Die adäquate Modellierung kategorialer Responsevariablen erfordert eine explizite Unterscheidung hinsichtlich des zugrunde liegenden Skalenniveaus in nominal– und ordinal–kategoriale Variablen. Während die Ausprägungen einer ordinalen Variablen in eine sinnvolle Ordnung gebracht werden können, ist für nominalskalierte Merkmale eine derartige Anordnung nicht möglich. Im Fall einer nominalen Variablen führt die Anwendung von Modellen, die speziell für das Vorliegen geordneter Kategorien konzipiert wurden, im allgemeinen zu Artefakten, da die benötigten Voraussetzungen vom zugrunde liegenden Response nicht erfüllt werden. Umgekehrt lassen sich Modelle, die lediglich auf nominales Niveau zugeschnitten sind, zwar ohne weiters auf Fragestellungen mit ordinalem Response anwenden, ignorieren aber die mit der Ordnungsstruktur gegebene zusätzliche Information.

Gegenstand dieses Kapitels sind Modelle mit nominalem Response. Die möglichen Ausprägungen der Responsevariablen werden im folgenden mit $1, \ldots, k$ bezeichnet, wobei die Zahlwerte lediglich der Identifizierung bzw. Kodierung von Namen und Bezeichnungen der jeweiligen Responsekategorien dienen. Ausgehend vom einfachsten Fall eines binären Merkmals mit nur zwei möglichen Ausprägungen werden flexible Darstellungen für mehrkategoriale Responsevariablen entwickelt, die von den nonparametrischen Modellierungsfor-

men der universellen Prädiktoren aus Kapitel 3 expliziten Gebrauch machen. Den Schwerpunkt der Betrachtungen bilden die aus der kategorialen Struktur der abhängigen Variablen erwachsenden Probleme und Besonderheiten.

5.1 Das multinomiale Logit–Modell

Für den Fall einer abhängigen Variablen mit nur zwei möglichen Ausprägungen ist die wohl gebräuchlichste Modellierung der Wirkung eines Einflußgrößenvektors $\boldsymbol{x} = (x_1, \ldots, x_p)'$ auf den dichotomen Response $\tilde{y} \in \{1, 2\}$ das binäre Logit–Modell. Mit der Festlegung $\pi_r(\boldsymbol{x}) = P(\tilde{y} = r \mid \boldsymbol{x})$, $r = 1, 2$, hat es die Form

$$\pi_1(\boldsymbol{x}) \;=\; P(\tilde{y} = 1 \mid \boldsymbol{x}) \;=\; \frac{\exp(\eta(\boldsymbol{x}))}{1 + \exp(\eta(\boldsymbol{x}))}\,. \tag{5.1}$$

Die Responsewahrscheinlichkeit $\pi_2(\boldsymbol{x})$ bedarf keiner separaten Modellierung. Sie ergibt sich direkt aus der axiomatischen Bedingung $\pi_1(\boldsymbol{x}) + \pi_2(\boldsymbol{x}) = 1$. Die durch $\tilde{y} = 2$ charakterisierte zweite Kategorie fungiert demnach nur als Referenzkategorie.

Während (5.1) die Abhängigkeit der Auftretenswahrscheinlichkeit $\pi_1(\boldsymbol{x})$ von den Einflußgrößen verdeutlicht, liefert

$$\mathrm{Logit}\big(\pi_1(\boldsymbol{x})\big) \;:=\; \log\left(\frac{\pi_1(\boldsymbol{x})}{1 - \pi_1(\boldsymbol{x})}\right) \;=\; \log\left(\frac{P(\tilde{y} = 1 \mid \boldsymbol{x})}{P(\tilde{y} = 2 \mid \boldsymbol{x})}\right) \;=\; \eta(\boldsymbol{x}) \tag{5.2}$$

eine äquivalente Darstellung des binären Logit–Modells in den logarithmierten Chancen (Logits). Der Quotient $\pi_1/(1 - \pi_1)$ charakterisiert die Chancen für das Eintreten der modellierten ersten Kategorie gegenüber der Referenzkategorie. Aus der Betrachtung von Chancen bzw. Logits resultieren einfache Interpretationen für die im Prädiktor $\eta(\boldsymbol{x})$ modellierten Kovariableneffekte.

Der allgemeinere Fall $\tilde{y} \in \{1, \ldots, k\}$, $k \geq 2$, läßt sich durch die Betrachtung jeweils zweier Kategorien auf den binären Fall zurückführen. Deklariert man k als Referenzkategorie, werden Logits ausgehend von (5.2) modelliert durch

$$\log\left(\frac{P(\tilde{y} = r \mid \boldsymbol{x})}{P(\tilde{y} = k \mid \boldsymbol{x})}\right) \;=\; \eta_r(\boldsymbol{x}), \quad r = 1, \ldots, q := k - 1, \tag{5.3}$$

wobei der Prädiktor jetzt als spezifisch für die betrachtete Kategorie ausgewiesen wird. Für die entsprechenden Auftretens- bzw. Responsewahrscheinlichkeiten $\pi_r(\boldsymbol{x}) = P(\tilde{y} = r \mid \boldsymbol{x})$, $r = 1, \ldots, k$, folgt aus (5.3)

$$P(\tilde{y} = r \mid \boldsymbol{x}) \;=\; \frac{\exp\big(\eta_r(\boldsymbol{x})\big)}{1 + \sum_{j=1}^{q} \exp(\eta_j(\boldsymbol{x}))}\,, \quad r = 1, \ldots, q, \tag{5.4}$$

und

$$P(\tilde{y} = k \mid \boldsymbol{x}) \;=\; \frac{1}{1 + \sum_{j=1}^{q} \exp(\eta_j(\boldsymbol{x}))}\,. \tag{5.5}$$

Die äquivalenten Darstellungen (5.3) und (5.4) bzw. (5.5) definieren das sogenannte *multinomiale Logit–Modell mit Referenzkategorie k.*

5.1.1 Das Zufallsnutzen–Modell

Ein völlig anderer Zugang zum multinomialen Logit–Modell findet seinen Ursprung im Bereich probabilistischer Wahlmodelle. Die verschiedenen Ausprägungen $\{1, \ldots, k\}$ der abhängigen Variablen $\tilde{y}$ werden dabei als Alternativen betrachtet, unter denen einzelne Individuen auswählen können. Die Entscheidung fällt zu Gunsten derjenigen Alternative $r \in \{1, \ldots, k\}$, deren zufälliger latenter Nutzen $U_r(\boldsymbol{x})$ maximal für das durch den Kovariablenvektor $\boldsymbol{x}$ charakterisierte Individuum ist.

Der subjektive Nutzen, auch als Utility bezeichnet, ist zwar nicht meßbar, seine individuenspezifische Maximierung manifestiert sich jedoch in einer konkreten Wahl $\tilde{y} = r$. Das Prinzip des maximalen zufälligen Nutzens postuliert einen entsprechenden Zusammenhang zwischen beobachtbarer Entscheidung und latentem Nutzen (Block & Marschak, 1960)

$$\tilde{y} = r \mid \boldsymbol{x} \quad \Leftrightarrow \quad U_r(\boldsymbol{x}) \;=\; \max_{1 \le j \le k} U_j(\boldsymbol{x}). \tag{5.6}$$

Damit ein Zufallsnutzen–Modell vorliegt, sind darüber hinaus spezifische Bedingungen an die Nutzenfunktionen $U_r(\boldsymbol{x})$, $r = 1, \ldots, k$, zu formulieren. Insbesondere fordert man die Gültigkeit der Darstellungen

$$U_r(\boldsymbol{x}) \;=\; u_r(\boldsymbol{x}) + \varepsilon_r, \;\; r = 1, \ldots, k, \tag{5.7}$$

mit nichtstochastischen Funktionen $u_r(\cdot)$, die die systematischen Anteile der latenten Utilities beschreiben und zufälligen Störungen ε_r, $r = 1, \ldots, k$. Zur Verteilung der Störterme wird dabei zunächst kein Annahme getroffen.

Werden die $\varepsilon_1, \ldots, \varepsilon_k$ als unabhängige, doppelt–exponentialverteilte Zufallsgrößen spezifiziert, liefert McFadden (1973) eine Aussage zur Ableitung der Darstellungen (5.4) und (5.5) als Zufallsnutzen–Modell:

Lemma 5.1. *Nimmt man für die Störterme ε_r Unabhängigkeit an und spezifiziert deren Verteilung als die Maximum-Extremwertverteilung mit Verteilungsfunktion*

$$F_{\varepsilon_r}(z) \;=\; P(\varepsilon_r \le z) \;:=\; \exp(-\exp(-z)),\;\; r = 1, \ldots, k, \tag{5.8}$$

läßt sich das multinomiale Logit–Modell als Zufallsnutzen–Modell motivieren.

Beweis. Aus (5.6) folgt unmittelbar für die Responsewahrscheinlichkeiten

$$\pi_r(\boldsymbol{x}) \;=\; P(\tilde{y} = r \,|\, \boldsymbol{x}) \;=\; P\big\{U_r(\boldsymbol{x}) = \max_{1 \le j \le k} U_j(\boldsymbol{x})\big\}, \quad r = 1, \ldots, k.$$

In den folgenden Beweisschritten werden Abhängigkeiten vom Kovariablenvektor nicht mehr explizit kenntlich gemacht. Mit obigem ist dann

$$\begin{aligned}
\pi_r &= P\big(U_r = \max\nolimits_j U_j\big) \;=\; P(U_r \ge U_j\,,\; \forall j \ne r) \;=\; P\big(U_r \ge U_1, \ldots, U_r \ge U_k\big) \\
&= P\big(\varepsilon_1 - \varepsilon_r \le u_r - u_1 =: w_{r1}, \ldots, \varepsilon_k - \varepsilon_r \le u_r - u_k =: w_{rk}\big) \\
&= \int_{s_k=-\infty}^{s_r+w_{rk}} \cdots \int_{s_r=-\infty}^{+\infty} \cdots \int_{s_1=-\infty}^{s_r+w_{r1}} f(s_1, \ldots, s_r, \ldots, s_k)\, ds_1 \ldots ds_r \ldots ds_k,
\end{aligned}$$

wobei $f : \mathbb{R}^k \longrightarrow [0, \infty)$ die gemeinsame Dichte der Störterme $\varepsilon_1, \ldots, \varepsilon_k$ darstellt. Da $f \ge 0$, kann die Integrationsreihenfolge vertauscht werden, so daß

$$\begin{aligned}
\pi_r &= \int_{s_r=-\infty}^{+\infty} \int_{s_k=-\infty}^{s_r+w_{rk}} \cdots \int_{s_1=-\infty}^{s_r+w_{r1}} f(s_1, \ldots, s_r, \ldots, s_k)\, ds_1 \ldots ds_k\, ds_r \\
&= \int_{-\infty}^{+\infty} \partial_r F(s + w_{r1}, \ldots, s, \ldots, s + w_{rk})\, ds,
\end{aligned} \tag{5.9}$$

wobei F die zu f gehörige Verteilungsfunktion und $\partial_r F$ deren r-te partielle Ableitung bezeichnet.

Da die Störgrößen $\varepsilon_1, \ldots, \varepsilon_k$ unabhängig und identisch verteilt sind, läßt sich deren gemeinsame Verteilungsfunktion gemäß Definition (5.8) auch schreiben als

$$F(s_1, \ldots, s_k) \;=\; \prod_{j=1}^{k} F_{\varepsilon_j}(s_j) \;=\; \prod_{j=1}^{k} \exp\big(-\exp(-s_j)\big)$$

und damit gilt für die r-te partielle Ableitung

$$\begin{aligned} \partial_r F(s_1, \ldots, s_k) \;&=\; \frac{\partial}{\partial s_r}\big(-\exp(-s_r)\big) \cdot \prod_{j=1}^{k} \exp\big(-\exp(-s_j)\big) \\ &=\; \exp(-s_r) \cdot \prod_{j=1}^{k} \exp\big(-\exp(-s_j)\big), \end{aligned}$$

so daß

$$\begin{aligned} \partial_r F(s + w_{r1}, \ldots, s, \ldots, s + w_{rk}) \;&=\; \exp(-s) \cdot \prod_{j=1}^{k} \exp\big(-\exp(-s - w_{rj})\big) \\ &=\; e^{-s} \cdot \exp(-e^{-s} \cdot c), \end{aligned}$$

mit $c := \sum_{j=1}^{k} e^{-w_{rj}}$. Berücksichtigung in (5.9) ergibt mit $t := -e^{-s} \cdot c$ und $dt/ds = c \cdot e^{-s}$

$$\pi_r \;=\; c^{-1} \int_{-\infty}^{+\infty} c \cdot e^{-s} \cdot \exp(-e^{-s} \cdot c)\, ds \;=\; c^{-1} \int_{-\infty}^{0} e^t\, dt \;=\; c^{-1} \cdot \big[e^t\big]_{-\infty}^{0} \;=\; c^{-1}.$$

Für die Responsewahrscheinlichkeiten $\pi_1, \ldots, \pi_k$ erhält man damit letztlich

$$\pi_r \;=\; \frac{1}{\sum_{j=1}^{k} \exp(-w_{rj})} \;=\; \frac{1}{\sum_{j=1}^{k} \exp(u_j - u_r)} \;=\; \frac{\exp(u_r)}{\sum_{j=1}^{k} \exp(u_j)}\,. \qquad (5.10)$$

Offensichtlich sind die systematischen Anteile $u_1, \ldots, u_k$ in (5.10) nicht identifizierbar (Verschiebbarkeit additiver Konstanten zwischen den Exponenten in Zähler und Nenner). Dieses Problem läßt sich durch die Auszeichnung der Kategorie k als Referenzkategorie und den Übergang zur Betrachtung der Differenzen $u_j - u_k$, $j = 1, \ldots, k$, umgehen.

Unter Einbeziehung der Kovariablen wird (5.10) damit zu

$$\pi_r(\boldsymbol{x}) \;=\; \frac{\exp\big(u_r(\boldsymbol{x}) - u_k(\boldsymbol{x})\big)}{\sum\limits_{j=1}^{k} \exp\big(u_j(\boldsymbol{x}) - u_k(\boldsymbol{x})\big)} \;=:\; \frac{\exp\big(\eta_r(\boldsymbol{x})\big)}{\sum\limits_{j=1}^{k} \exp\big(\eta_j(\boldsymbol{x})\big)}\,,$$

mit $\eta_j(\boldsymbol{x}) := u_j(\boldsymbol{x}) - u_k(\boldsymbol{x})$, $j = 1, \ldots, k$. Da $\exp\big(\eta_k(\boldsymbol{x})\big) = \exp(0) = 1$, folgen die Schreibweisen (5.4) und (5.5) für das multinomiale Logit-Modell. □

5.1.2 Kategorienspezifische Charakteristiken

Die systematischen Anteile der Nutzenfunktionen $U_1, \ldots, U_k$ werden im rein parametrischen multinomialen Logit–Modell linear spezifiziert

$$u_r(\boldsymbol{x}) \;=\; \begin{pmatrix}1 & \boldsymbol{x}'\end{pmatrix} \begin{pmatrix}\tilde{\gamma}_{0r} \\ \tilde{\boldsymbol{\gamma}}_r\end{pmatrix} \;=\; \tilde{\gamma}_{0r} \;+\; \boldsymbol{x}'\tilde{\boldsymbol{\gamma}}_r, \quad r = 1, \ldots, k. \tag{5.11}$$

Dabei sind die Kovariablengewichte jeweils spezifisch für die modellierte Kategorie. Sind die verschiedenen Responsekategorien als Wahlalternativen interpretierbar, so umfaßt $\boldsymbol{x}$ ausschließlich Merkmale, die den Entscheidungsträger charakterisieren. In Wahlmodellen spielen jedoch häufig auch Einflußgrößen eine Rolle, die sowohl individuenspezifisch als auch charakteristisch für die gewählte Option sind. Ein klassisches Beispiel hierfür ist die Wahl des Transportmittels für den Weg zur Arbeit. Bus, Bahn oder Auto als denkbare Alternativen korrespondieren mit entsprechenden Fahrpreisen und Fahrdauern. Diese Merkmale sind also charakteristisch für das gewählte Transportmittel, hängen aber über die Entfernung zwischen Wohnort und Arbeitsstelle zusätzlich vom Entscheidungsträger ab.

Im folgenden sollen diese kategorienspezifischen Charakteristiken explizit berücksichtigt werden. Dazu seien mit $\boldsymbol{w}_r = (w_{1r}, \ldots, w_{mr})'$, $r = 1, \ldots, k$, diejenigen Merkmalsvektoren bezeichnet, die spezifisch für die jeweiligen Kategorien sind. Die Erweiterung von (5.11) zu

$$u_r\big(\boldsymbol{x}, \{\boldsymbol{w}_j\}\big) \;=\; \tilde{\gamma}_{0r} \;+\; \boldsymbol{x}'\tilde{\boldsymbol{\gamma}}_r \;+\; \boldsymbol{w}_r'\boldsymbol{\alpha}, \quad r = 1, \ldots, k, \tag{5.12}$$

erlaubt eine Einbeziehung kategorienspezifischer Charakteristiken in die Modellierung der Funktionen $u_1(\cdot), \ldots, u_k(\cdot)$.

Mit der aus Identifizierbarkeitsgründen erforderlichen Betrachtung der Differenzen $\eta_r(\boldsymbol{x}) := u_r(\boldsymbol{x}) - u_k(\boldsymbol{x})$, $r = 1, \ldots, k$, erhält man daraus

$$\eta_r\big(\boldsymbol{x}, \{\boldsymbol{w}_j\}\big) \;=\; \tilde{\gamma}_{0r} - \tilde{\gamma}_{0k} \;+\; \boldsymbol{x}'\big(\tilde{\boldsymbol{\gamma}}_r - \tilde{\boldsymbol{\gamma}}_k\big) \;+\; \big(\boldsymbol{w}_r - \boldsymbol{w}_k\big)'\boldsymbol{\alpha},$$

bzw. vereinfacht mit $\gamma_{0r} := \tilde{\gamma}_{0r} - \tilde{\gamma}_{0k}$ und $\boldsymbol{\gamma}_r := \tilde{\boldsymbol{\gamma}}_r - \tilde{\boldsymbol{\gamma}}_k$

$$\log(\pi_r/\pi_k) \;=\; \eta_r \;=\; \gamma_{0r} \;+\; \boldsymbol{x}'\boldsymbol{\gamma}_r \;+\; \big(\boldsymbol{w}_r - \boldsymbol{w}_k\big)'\boldsymbol{\alpha}, \quad r = 1, \ldots, q. \tag{5.13}$$

Der mit der Modellierung der Logits angestellte Vergleich zwischen Kategorie r und Referenzkategorie k spiegelt sich in der Differenz $\boldsymbol{w}_r - \boldsymbol{w}_k$ jetzt auch auf Seiten der Einflußgrößen wieder. Auf einer weiteren Verallgemeinerungsstufe können die Merkmalsvektoren $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ zudem Interaktionen enthalten (Tutz, 2000).

5.2 Das multinomiale Logit–Modell als GLM

In den bisherigen Untersuchungen blieb der eigentlich multivariate Charakter der Responsevariablen $\tilde{y} \in \{1, \ldots, k\}$ völlig unberücksichtigt. Dieser wird jedoch offensichtlich, wenn man statt $\tilde{y}$ den Vektor $\boldsymbol{y} := (y_1, \ldots, y_q)'$ mit

$$y_r := \begin{cases} 1, & \text{falls } \tilde{y} = r \\ 0, & \text{sonst} \end{cases}, \quad r = 1, \ldots, q \tag{5.14}$$

und der durch $\boldsymbol{y} = \boldsymbol{0}$ gekennzeichneten Referenzkategorie betrachtet. Übertragen auf die Responsewahrscheinlichkeiten heißt das

$$\pi_r \;=\; P(y_r = 1), \;\; r = 1, \ldots, q, \;\; \text{und} \;\; \pi_k \;=\; 1 - \sum_{r=1}^{q} \pi_r \;=\; P(\boldsymbol{y} = \boldsymbol{0}),$$

wobei hier und im folgenden Abhängigkeiten von den Kovariablen $\boldsymbol{x}$ und $\boldsymbol{w}_r$, $r = 1, \ldots, k$, nicht mehr explizit ausgewiesen werden. Der Responsevektor $\boldsymbol{y}$ besitzt eine Multinomialverteilung mit Wahrscheinlichkeitsfunktion

$$P_{\boldsymbol{y}}\big(\{(m_1, \ldots, m_q)'\}\big) \;=\; \frac{1}{m_1! \cdot \ldots \cdot m_k!} \; \pi_1^{m_1} \cdot \ldots \cdot \pi_k^{m_k} \;=\; \sum_{r=1}^{k} m_r \pi_r, \tag{5.15}$$

wobei $m_1, \ldots, m_q \in \{0, 1\}$, $\sum_{r=1}^{q} m_r \in \{0, 1\}$ und $m_k := 1 - \sum_{r=1}^{q} m_r$. Setzt man $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_q)'$, schreibt man oft auch kurz $\boldsymbol{y} \sim \mathcal{M}(1, \boldsymbol{\pi})$.

Eine Einbettung des multinomialen Logit–Modells (5.13) in den Rahmen generalisierter linearer Modelle erfordert zunächst die Erweiterung des Begriffs der Exponentialfamilie auf den Fall multivariater Verteilungen. Diese ist mit der Forderung

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}, \phi, \omega) \;=\; \exp\Big\{\big(\boldsymbol{y}'\boldsymbol{\theta} - b(\boldsymbol{\theta})\big)\,\omega\phi^{-1} \;+\; c(\boldsymbol{y}, \phi, \omega)\Big\} \tag{5.16}$$

als naheliegende Verallgemeinerung von (1.3) gegeben. Mit $\phi = \omega = 1$,

$$\boldsymbol{\theta} \;=\; (\theta_1, \ldots, \theta_q)' \;=\; \big(\log(\pi_1/\pi_k), \ldots, \log(\pi_q/\pi_k)\big)'$$

und

$$b(\boldsymbol{\theta}) \;=\; \log\left\{1 + \sum_{r=1}^{q} \exp(\theta_r)\right\} \;=\; -\log(1 - \pi_1 - \ldots - \pi_q) \;=\; -\log(\pi_k),$$

sowie $c(\boldsymbol{y}, \phi, \omega) = 0$, läßt sich die Multinomialverteilung $P_{\boldsymbol{y}}$ als Mitglied der Exponentialfamilie (5.16) schreiben.

Für eine konkrete Datensituation $\big\{\boldsymbol{y}_i, \boldsymbol{x}_i, \{\boldsymbol{w}_{ij}\}\big\}_{i=1,\ldots,N}$ mit (bedingt) unabhängigen $\boldsymbol{y}_i \mid \boldsymbol{x}_i, \{\boldsymbol{w}_{ij}\} \sim \mathcal{M}(1, \boldsymbol{\pi}_i)$, können auch die strukturellen Annahmen (ii) und (iii) aus Abschnitt 1.1 leicht verifiziert werden. Zu diesem Zweck definiert man die Designmatrizen Z_i, $i = 1, \ldots, N$, als

$$Z_i \;=\; \begin{pmatrix} 1 & & 0 & \boldsymbol{x}_i' & & 0 & \boldsymbol{w}_{i1}' - \boldsymbol{w}_{ik}' \\ & \ddots & & & \ddots & & \vdots \\ 0 & & 1 & 0 & & \boldsymbol{x}_i' & \boldsymbol{w}_{iq}' - \boldsymbol{w}_{ik}' \end{pmatrix}.$$

Für die Vektoren $\boldsymbol{\eta}_i := (\eta_{i1}, \ldots, \eta_{iq})'$, $i = 1, \ldots, N$, erhält man $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$, wobei $\boldsymbol{\gamma}_0 := (\gamma_{01}, \ldots, \gamma_{0q})'$ und $\boldsymbol{\beta}' := (\boldsymbol{\gamma}_0', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_q', \boldsymbol{\alpha}')$ die unbekannten Parameter subsumieren. Als Linkfunktion $g = (g_1, \ldots, g_q) : \mathbb{R}^q \longrightarrow \mathbb{R}^q$ läßt sich

$$g_r(\boldsymbol{\pi}_i) \;:=\; \log(\pi_{ir}/\pi_{ik}) \;=\; \log\left(\frac{\pi_{ir}}{1 - \pi_{i1} - \ldots - \pi_{iq}}\right) \;=\; \eta_{ir}, \;\; r = 1, \ldots, q,$$

definieren, so daß $g(\boldsymbol{\pi}_i) = \boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$ gilt. Mit $h = g^{-1}$ ergeben sich die Komponenten der Responsefunktion $h : \mathbb{R}^q \longrightarrow \mathbb{R}^q$ daraus zu

$$h_r(\boldsymbol{\eta}_i) \;=\; \frac{\exp(\eta_{ir})}{1 + \sum_{j=1}^{q} \exp(\eta_{ij})} \;=\; \pi_{ir}, \;\; r = 1, \ldots, q.$$

Da $y_{ir} \sim \mathcal{B}(1, \pi_{ir})$, $r = 1, \ldots, q$, folgt zudem $\boldsymbol{\mu}_i := E(\boldsymbol{y}_i) = \boldsymbol{\pi}_i$, $i = 1, \ldots, N$, so daß auch im multinomialen Logit–Modell die für GLM's charakteristische Beziehung $\boldsymbol{\mu}_i = h(\boldsymbol{\eta}_i)$ zwischen Erwartungswert(vektor) $\boldsymbol{\mu}_i$ und Prädiktor $\boldsymbol{\eta}_i$ besteht.

5.2.1 Maximum–Likelihood–Schätzung

Ausgehend von konkreten Beobachtungen $\{\boldsymbol{y}_i, \boldsymbol{x}_i, \{\boldsymbol{w}_{ij}\}\}_{i=1,\ldots,N}$ zielt die Maximum–Likelihood–Schätzung analog zum univariaten Fall auf die Maximierung der Log–Likelihood $l(\boldsymbol{\beta}) := \sum_i \ln f\big(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_i(\boldsymbol{\beta}), \phi_i, \omega_i\big)$ ab. Aus (5.16) erhält man die Score–Funktion

$$s(\boldsymbol{\beta}) \;=\; \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \;=\; \sum_{i=1}^{N} Z_i' \, D_i(\boldsymbol{\beta}) \, \Sigma_i(\boldsymbol{\beta})^{-1} \big(\boldsymbol{y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta})\big), \tag{5.17}$$

mit den Jacobi–Matrizen $D_i(\boldsymbol{\beta}) = \partial h(\boldsymbol{\eta}_i)/\partial \boldsymbol{\eta}_i$ sowie den Kovarianzmatrizen $\Sigma_i(\boldsymbol{\beta}) = \mathrm{cov}(\boldsymbol{y}_i) = \mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i'$, $i = 1, \ldots, N$.

Deklariert man ferner

$$D(\boldsymbol{\beta}) := \mathrm{Diag}\big(D_1(\boldsymbol{\beta}), \ldots, D_N(\boldsymbol{\beta})\big), \quad \Sigma(\boldsymbol{\beta}) := \mathrm{Diag}\big(\Sigma_1(\boldsymbol{\beta}), \ldots, \Sigma_N(\boldsymbol{\beta})\big)$$

$$\text{und} \;\; Z := \big[\, Z_1' \mid \ldots \mid Z_N' \,\big]'$$

als totale Designmatrix, so läßt sich die Score–Funktion kompakter schreiben als $s(\boldsymbol{\beta}) = Z' D(\boldsymbol{\beta}) \Sigma^{-1}(\boldsymbol{\beta}) \big(\boldsymbol{y}^* - \boldsymbol{\pi}^*\big)$, wobei

$$\boldsymbol{y}^* := (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')' \quad \text{und} \quad \boldsymbol{\pi}^* := \big(\boldsymbol{\pi}_1(\boldsymbol{\beta})', \ldots, \boldsymbol{\pi}_N(\boldsymbol{\beta})'\big)'.$$

Eine Lösung der Schätzgleichungen $s(\boldsymbol{\beta}) = \mathbf{0}$ ist wiederum nur iterativ möglich. Die numerische Umsetzung der entsprechenden Fisher–Scoring–Schritte erfolgt in Anlehnung an den univariaten Fall als gewichtete KQ–Schätzung. Auf die vorliegende Situation adaptierend, sind dazu sämtliche Vektoren und Matrizen des in 1.1.1 formulierten Algorithmus durch ihre multivariaten Versionen zu ersetzen. So werden die adjustierten Prädiktoren zu

$$\tilde{\boldsymbol{\eta}}'(\boldsymbol{\beta}) = \big(\tilde{\boldsymbol{\eta}}_1'(\boldsymbol{\beta}), \ldots, \tilde{\boldsymbol{\eta}}_N'(\boldsymbol{\beta})\big), \;\text{ mit }\; \tilde{\boldsymbol{\eta}}_i(\boldsymbol{\beta}) = Z_i \boldsymbol{\beta} + D_i\big(\boldsymbol{\beta}\big)^{-1} \big(\boldsymbol{y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta})\big)$$

modifiziert.

Folgeiterierte berechnen sich gemäß der Vorschrift

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left[Z' W\left(\hat{\boldsymbol{\beta}}^{(k)}\right) Z \right]^{-1} Z' W\left(\hat{\boldsymbol{\beta}}^{(k)}\right) \tilde{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\beta}}^{(k)}\right), \; k = 0, 1, 2, \ldots,$$

mit $W(\boldsymbol{\beta}) := D(\boldsymbol{\beta})\, \Sigma^{-1}(\boldsymbol{\beta})\, D'(\boldsymbol{\beta})$ bzw. in der klassischen Notation des Fisher-Scoring

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + F\left(\hat{\boldsymbol{\beta}}^{(k)}\right)^{-1} s\left(\hat{\boldsymbol{\beta}}^{(k)}\right), \; k = 0, 1, 2, \ldots,$$

mit der erwarteten Fisher'schen Informationsmatrix $F(\boldsymbol{\beta}) = Z'W(\boldsymbol{\beta})Z$.

5.3 Semiparametrische Modellierung

Mehrkategoriale parametrische Regressionsmodelle gehören mittlerweile zum Standardwerkzeug in der statistischen Datenanalyse. Im Zuge der Formulierung flexiblerer Zusammenhangsstrukturen ist man jedoch auch im multivariaten Fall an einer Lockerung der rigiden Linearitätsannahmen interessiert. Während univariate GAM's und deren Erweiterungen weit verbreitet sind, gibt es bis dato nur wenige Ansätze zur nonparametrischen Modellierung kategorialer Responsevariablen. Yee & Wild (1996) fitten multivariate additive Modelle unter Verwendung von Glättungssplines und iterativem Backfitting. Die im Marketing interessierende Frage nach Präferenzen bei der Markenwahl diskutiert Abe (1999) über eben diesen erweiterten GAM–Ansatz. Ebenfalls im Bereich der Markenwahl ist die Arbeit von Hruschka (2002) angesiedelt. Der Autor stellt eine Erweiterung des parametrischen multinomialen Logit–Modells um ein künstliches neuronales Netz vor, das eine nichtlineare Modellierung des deterministischen Nutzens von Marken erlaubt.

Das hier vorgestellte semiparametrische Modell unterscheidet sich von obigen Ansätzen in zweierlei Hinsicht. Während Markenwahlmodelle ausschließlich kategorienspezifische Charakteristiken berücksichtigen und die Modellierung in Yee & Wild (1996) auf globalen, sprich rein individuenspezifischen, Merkmalen basiert, sind hier Variablen beiden Typs zugelassen. P–Spline Ansätze zur Modellierung nonparametrischer Komponenten ermöglichen darüber hinaus die Schätzung der unbekannten Modellparameter im Rahmen multivariater generalisierter linearer Modelle.

In Verallgemeinerung von (5.12) modellieren wir die deterministischen Anteile der Nutzenfunktionen in unspezifizierter additiver Form. Die Beobachtungen $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ für die globalen Variablen und $\boldsymbol{w}_{ir} = (w_{i1r}, \ldots, w_{imr})'$ für die kategorienspezifischen Charakteristiken bestimmen die Utility–Komponenten u_{ir}, $i = 1, \ldots, N$, jetzt in der Form

$$u_{ir} \;=\; \tilde{\gamma}_{0r} \;+\; \sum_{j=1}^{p} \tilde{\gamma}_{(j),r}(x_{ij}) \;+\; \sum_{j=1}^{m} \alpha_{(j)}(w_{ijr}), \;\; r = 1, \ldots, k, \tag{5.18}$$

mit der korrespondierenden Darstellung

$$\eta_{ir} \;=\; \gamma_{0r} + \sum_{j=1}^{p} \gamma_{(j),r}(x_{ij}) + \sum_{j=1}^{m} \alpha_{(j)}(w_{ijr}) - \alpha_{(j)}(w_{ijk}), \;\; r = 1, \ldots, q,$$

für die Prädiktoren $\eta_{ir} := u_{ir} - u_{ik}$. In $\gamma_{(j),r}(x_{ij}) := \tilde{\gamma}_{(j),r}(x_{ij}) - \tilde{\gamma}_{(j),k}(x_{ij})$ ist dabei der Effekt der j-ten globalen Variable zusammengefaßt. Ein einfacheres Modell resultiert aus der Vorstellung, daß nur die Differenzen $w_{ijr} - w_{ijk}$ einen Effekt auf die Logits haben

$$\eta_{ir} \;=\; \gamma_{0r} + \sum_{j=1}^{p} \gamma_{(j),r}(x_{ij}) + \sum_{j=1}^{m} \alpha_{(j)}(w_{ijr} - w_{ijk}), \;\; r = 1, \ldots, q.$$

Für diese Form der Darstellung läßt sich jedoch im allgemeinen kein direkter Bezug zum Zufallsnutzen–Modell herstellen, da die Spezifizierung der Kovariableneffekte auf der Stufe des erwarteten Nutzens, also vor der Differenzenbildung zu erfolgen hat.

In einem weiteren Verallgemeinerungsschritt verzichten wir auf die mit (5.18) implizierte, responseunabhängige Modellierung der Effekte kategorienspezifischer Charakteristiken

$$u_{ir} \;=\; \tilde{\gamma}_{0r} \;+\; \sum_{j=1}^{p} \tilde{\gamma}_{(j),r}(x_{ij}) \;+\; \sum_{j=1}^{m} \alpha_{(j),r}(w_{ijr}), \;\; r = 1, \ldots, k, \tag{5.19}$$

bzw.

$$\eta_{ir} = \gamma_{0r} + \sum_{j=1}^{p} \gamma_{(j),r}(x_{ij}) + \sum_{j=1}^{m} \alpha_{(j),r}(w_{ijr}) - \alpha_{(j),k}(w_{ijk}), \; r = 1, \ldots, q. \tag{5.20}$$

(5.20) steht dabei in enger Beziehung zum Wahlmodell der Nutzenmaximierung. Ersetzt man in (5.12) $\boldsymbol{\alpha}$ durch $\boldsymbol{\alpha}_r$, so resultiert (5.19) als direkte Folge eines nonparametrischen Ansatzes.

Enthält die Kovariablenmenge diskrete Merkmale, so ist es fragwürdig, deren Einfluß nicht-linear zu modellieren. Semiparametrische Ansätze, in denen lediglich stetige Kovariablen nonparametrisch modelliert werden, erscheinen in diesen Situationen plausibler. Ähnlich den Ausführungen von Kapitel 3 werden die Indexmengen $\{1,\ldots,p\}$ bzw. $\{1,\ldots,m\}$ daher gemäß $\{\mathcal{D}_{\boldsymbol{x}},\mathcal{S}_{\boldsymbol{x}}\}$ bzw. $\{\mathcal{D}_{\boldsymbol{w}},\mathcal{S}_{\boldsymbol{w}}\}$ arrangiert und die Prädiktorstruktur für $r=1,\ldots,q$ in

$$\eta_{ir} = \gamma_{0r} + \eta_{ir,L} + \eta_{ir,A} = \gamma_{0r} + \eta_{ir,L}(\boldsymbol{x}) + \eta_{ir,L}(\boldsymbol{w}) + \eta_{ir,A}(\boldsymbol{x}) + \eta_{ir,A}(\boldsymbol{w}),$$

abgewandelt mit den separaten GLM– und GAM–Komponenten

$$\eta_{ir,L}(\boldsymbol{x}) = \sum_{j\in\mathcal{D}_{\boldsymbol{x}}} x_{ij}\gamma_{jr}, \qquad \eta_{ir,L}(\boldsymbol{w}) = \sum_{j\in\mathcal{D}_{\boldsymbol{w}}} w_{ijr}\alpha_{jr} - w_{ijk}\alpha_{jk},$$
$$\eta_{ir,A}(\boldsymbol{x}) = \sum_{j\in\mathcal{S}_{\boldsymbol{x}}} \gamma_{(j),r}(x_{ij}), \qquad \eta_{ir,A}(\boldsymbol{w}) = \sum_{j\in\mathcal{S}_{\boldsymbol{w}}} \alpha_{(j),r}(w_{ijr}) - \alpha_{(j),k}(w_{ijk}).$$

Setzt man $\tilde{\boldsymbol{x}}_i' = (x_{i1},\ldots,x_{i|\mathcal{D}_{\boldsymbol{x}}|})$ und $\boldsymbol{\gamma}_r = (\gamma_{1r},\ldots,\gamma_{|\mathcal{D}_{\boldsymbol{x}}|r})'$, $r=1,\ldots,q$, sowie $\tilde{\boldsymbol{w}}_{ir}' = (w_{i1r},\ldots,w_{i|\mathcal{D}_{\boldsymbol{w}}|r})$ und $\boldsymbol{\alpha}_r = (\alpha_{1r},\ldots,\alpha_{|\mathcal{D}_{\boldsymbol{w}}|r})'$, $r=1,\ldots,k$, so folgen

$$\eta_{ir,L}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}_i'\boldsymbol{\gamma}_r \quad \text{und} \quad \eta_{ir,L}(\boldsymbol{w}) = \tilde{\boldsymbol{w}}_{ir}'\boldsymbol{\alpha}_r - \tilde{\boldsymbol{w}}_{ik}'\boldsymbol{\alpha}_k, \;\; r=1,\ldots,q.$$

Die Modellierung der nonparametrischen Komponenten in $\eta_{ir,A}$ erfolgt unter Verwendung entsprechender B-Spline-Basen. In Anlehnung an Abschnitt 3.3 resultieren für $r=1,\ldots,q$ bzw. $r=1,\ldots,k$ die Darstellungen

$$\gamma_{(j),r}(x_{ij}) = \sum_{s=1}^{P} \gamma_{jrs}G_{js}(x_{ij}) \quad \text{bzw.} \quad \alpha_{(j),r}(w_{ijr}) = \sum_{s=1}^{P} \alpha_{jrs}A_{js}(w_{ijr}), \quad (5.21)$$

wobei die Differenzierung in B-Splines G_{js} und A_{js} lediglich der besseren Unterscheidung in globale und kategorienspezifische Variablen dient. Ausschlaggebendes Argument für die Modellierung in Basisfunktionen ist wiederum die damit einhergehende Rückführbarkeit auf lineare Strukturen. Unter Berücksichtigung entsprechender Restriktionen an die Basiskoeffizienten (vgl. Kapitel 3) leiten sich aus (5.21) die Schreibweisen

$$\eta_{ir,A}(\boldsymbol{x}) = \sum_{j\in\mathcal{S}_{\boldsymbol{x}}} \boldsymbol{c}_{ij}'\tilde{\boldsymbol{\gamma}}_{jr} \quad \text{und} \quad \eta_{ir,A}(\boldsymbol{w}) = \sum_{j\in\mathcal{S}_{\boldsymbol{w}}} \boldsymbol{a}_{ijr}'\tilde{\boldsymbol{\alpha}}_{jr} - \boldsymbol{a}_{ijk}'\tilde{\boldsymbol{\alpha}}_{jk}$$

für die GAM–Komponenten ab, wobei $\tilde{\boldsymbol{\gamma}}_{jr} = (\gamma_{jr1},\ldots,\gamma_{jr,P-1})'$, $r=1,\ldots,q$,

$\boldsymbol{c}'_{ij} = \big(G_{j1}(x_{ij}) - G_{jP}(x_{ij}), \ldots, G_{j,P-1}(x_{ij}) - G_{jP}(x_{ij})\big)$ und

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}_{jr} &= (\alpha_{jr1}, \ldots, \alpha_{jr,P-1})',\\ \boldsymbol{a}'_{ijr} &= \big(A_{j1}(w_{ijr}) - A_{jP}(w_{ijr}), \ldots, A_{j,P-1}(w_{ijr}) - A_{jP}(w_{ijr})\big),\end{aligned}$$

$r = 1, \ldots, k$. Die Prädiktoren $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iq})'$ sind damit gegeben als

$$\boldsymbol{\eta}_i = \boldsymbol{\gamma}_0 + X_i\boldsymbol{\gamma} + W_i\boldsymbol{\alpha} - W_{ik}\boldsymbol{\alpha}_k + \sum_{j\in\mathcal{S}_{\boldsymbol{x}}} C_{ij}\tilde{\boldsymbol{\gamma}}_j + \sum_{j\in\mathcal{S}_{\boldsymbol{w}}} A_{ij}\tilde{\boldsymbol{\alpha}}_j - A_{ijk}\tilde{\boldsymbol{\alpha}}_{jk},$$

$i = 1, \ldots, N$, wobei

$$\begin{aligned} X_i &= I_q \otimes \tilde{\boldsymbol{x}}'_i, & \boldsymbol{\gamma}_0 &= (\gamma_{01}, \ldots, \gamma_{0q})', & \tilde{\boldsymbol{\alpha}}_j &= (\tilde{\boldsymbol{\alpha}}'_{j1}, \ldots, \tilde{\boldsymbol{\alpha}}'_{jq})',\\ C_{ij} &= I_q \otimes \boldsymbol{c}'_{ij}, & \boldsymbol{\gamma} &= (\boldsymbol{\gamma}'_1, \ldots, \boldsymbol{\gamma}'_q)', & W_i &= \mathrm{Diag}(\widetilde{\boldsymbol{w}}'_{i1}, \ldots, \widetilde{\boldsymbol{w}}'_{iq}),\\ W_{ik} &= \mathbf{1}_q \otimes \widetilde{\boldsymbol{w}}'_{ik}, & \boldsymbol{\alpha} &= (\boldsymbol{\alpha}'_1, \ldots, \boldsymbol{\alpha}'_q)', & A_{ij} &= \mathrm{Diag}(\boldsymbol{a}'_{ij1}, \ldots, \boldsymbol{a}'_{ijq}).\\ A_{ijk} &= \mathbf{1}_q \otimes \boldsymbol{a}'_{ijk}, & \tilde{\boldsymbol{\gamma}}_j &= (\tilde{\boldsymbol{\gamma}}'_{j1}, \ldots, \tilde{\boldsymbol{\gamma}}'_{jq})', \end{aligned}$$

Mit $\tilde{p} := |\mathcal{D}_{\boldsymbol{x}}| + 1$ und $\widetilde{m} := |\mathcal{D}_{\boldsymbol{w}}| + 1$ definieren wir die Designmatrizen

$$Z_i = \Big[\, I_q \mid X_i \mid W_i \mid -W_{ik} \mid C_{i\tilde{p}} \mid \ldots \mid C_{ip} \mid A_{i\widetilde{m}} \mid -A_{i\widetilde{m}k} \mid \ldots \mid A_{im} \mid -A_{imk} \,\Big],$$

$i = 1, \ldots, N$ und den Vektor der unbekannten Parameter

$$\boldsymbol{\beta} = \big(\boldsymbol{\gamma}'_0, \boldsymbol{\gamma}', \boldsymbol{\alpha}', \boldsymbol{\alpha}'_k, \tilde{\boldsymbol{\gamma}}'_{\tilde{p}}, \ldots, \tilde{\boldsymbol{\gamma}}'_p, \tilde{\boldsymbol{\alpha}}'_{\widetilde{m}}, \tilde{\boldsymbol{\alpha}}'_{\widetilde{m}k}, \ldots, \tilde{\boldsymbol{\alpha}}'_m, \tilde{\boldsymbol{\alpha}}'_{mk}\big)',$$

dessen Schätzung wegen $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$, $i = 1, \ldots, N$, nunmehr im Rahmen multivariater GLM's erfolgen kann (vgl. Abschnitt 5.2.1). Analog zum univariaten Fall erfordern die B-Spline Darstellungen der nonparametrischen Komponenten jedoch Strafterme für die Basiskoeffizienten. Betrachtet wird daher nicht die Log–Likelihood $l(\boldsymbol{\beta})$, sondern deren penalisierte Fassung

$$\begin{aligned} pl(\boldsymbol{\beta}) &= l(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j\in\mathcal{S}_{\boldsymbol{x}}}\sum_{r=1}^{q} \lambda^g_{jr}\tilde{\boldsymbol{\gamma}}'_{jr}\big(\tilde{D}^d_P\big)'\tilde{D}^d_P\,\tilde{\boldsymbol{\gamma}}_{jr} - \frac{1}{2}\sum_{j\in\mathcal{S}_{\boldsymbol{w}}}\sum_{r=1}^{k} \lambda^c_{jr}\tilde{\boldsymbol{\alpha}}'_{jr}\big(\tilde{D}^d_P\big)'\tilde{D}^d_P\,\tilde{\boldsymbol{\alpha}}_{jr}\\ &= l(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j\in\mathcal{S}_{\boldsymbol{x}}} \tilde{\boldsymbol{\gamma}}'_j\,\tilde{K}^g_j\,\tilde{\boldsymbol{\gamma}}_j - \frac{1}{2}\sum_{j\in\mathcal{S}_{\boldsymbol{w}}} \tilde{\boldsymbol{\alpha}}'_j\,\tilde{K}^c_j\,\tilde{\boldsymbol{\alpha}}_j + \tilde{\boldsymbol{\alpha}}'_{jk}\,\tilde{K}^c_{jk}\,\tilde{\boldsymbol{\alpha}}_{jk}\end{aligned}$$

mit den Glättungsparametern λ^g_{jr}, λ^c_{jr} für die globalen und die kategorienspezifischen Variablen sowie den Strafmatrizen

$$\tilde{K}^g_j = \Lambda^g_j\tilde{D}'\tilde{D}, \qquad \tilde{K}^c_j = \Lambda^c_j\tilde{D}'\tilde{D}, \qquad \tilde{K}^c_{jk} = \lambda^c_{jk}\big(\tilde{D}^d_P\big)'\tilde{D}^d_P.$$

In $\Lambda_j^g = \mathrm{diag}\big((\lambda_{j1}^g, \ldots, \lambda_{jq}^g) \otimes \mathbf{1}'_{P-1}\big)$ und $\Lambda_j^c = \mathrm{diag}\big((\lambda_{j1}^c, \ldots, \lambda_{jq}^c) \otimes \mathbf{1}'_{P-1}\big)$ sind die Glättungsparameter variablenweise gruppiert, $\tilde{D} = I_q \otimes \tilde{D}_P^d$ bezeichnet die zugehörige Differenzenmatrix. Wie schon in Kapitel 3 wird dabei auf eine variable Auszeichnung von Knotenzahlen, B–Spline- und Differenzenordnungen verzichtet.

Der Maximum–Likelihood–Schätzer für $\boldsymbol{\beta}$ ist Lösung der Schätzgleichungen

$$ps(\boldsymbol{\beta}) \;=\; \partial pl(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \;=\; s(\boldsymbol{\beta}) \;-\; K\boldsymbol{\beta} \;=\; \mathbf{0} \tag{5.22}$$

für die Penalty–Matrix $K = \mathrm{Diag}\big(0_{v\times v}, \tilde{K}_{\tilde{p}}^g, \ldots, \tilde{K}_p^g, \tilde{K}_{\tilde{m}}^c, \tilde{K}_{\tilde{m}k}^c, \ldots, \tilde{K}_m^c, \tilde{K}_{mk}^c\big)$ mit quadratischer Nullmatrix der Dimension $v = q \cdot (\tilde{p} + \tilde{m} - 1) + \tilde{m} - 1$.

Die unpenalisierte Score-Funktion berechnet sich unter Berücksichtigung der hier vorliegenden Designstruktur wie in Abschnitt 5.2.1 als

$$s(\boldsymbol{\beta}) \;=\; Z' D(\boldsymbol{\beta})\, \Sigma(\boldsymbol{\beta})^{-1} (\boldsymbol{y}^* - \boldsymbol{\pi}^*),$$

so daß die einzig notwendige Modifikation der dort beschriebenen, gewichteten KQ–Schätzung – analog zum univariaten Fall – im Ersetzen der Fisher–Matrix durch ihre Pseudo–Form $\tilde{F}(\boldsymbol{\beta}) = F(\boldsymbol{\beta}) + K$ besteht.

Um Signifikanzaussagen über den Maximum–Likelihood–Schätzer $\hat{\boldsymbol{\beta}}$ ableiten zu können, muß dessen Kovarianzmatrix bestimmt werden. Die Entwicklung von $ps(\hat{\boldsymbol{\beta}})$ in eine Taylor–Reihe um den wahren Wert $\boldsymbol{\beta}$ liefert in erster Näherung

$$ps(\hat{\boldsymbol{\beta}}) \;\approx\; ps(\boldsymbol{\beta}) \;+\; \big(-F_{obs}(\boldsymbol{\beta}) - K\big)\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big).$$

Mit $ps(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ folgt daraus

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;\approx\; \big(F_{obs}(\boldsymbol{\beta}) + K\big)^{-1} ps(\boldsymbol{\beta}) \;\approx\; \tilde{F}(\boldsymbol{\beta})^{-1} ps(\boldsymbol{\beta}),$$

letzteres resultiert aus dem Ersetzen der beobachteten Fisher'schen Informationsmatrix $F_{obs}(\boldsymbol{\beta})$ durch ihren Erwartungswert $F(\boldsymbol{\beta})$. Damit ist

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) \;\approx\; \tilde{F}(\boldsymbol{\beta})^{-1} \mathrm{cov}\big(ps(\boldsymbol{\beta})\big) \tilde{F}(\boldsymbol{\beta})^{-1} \;=\; \tilde{F}(\boldsymbol{\beta})^{-1} F(\boldsymbol{\beta})\, \tilde{F}(\boldsymbol{\beta})^{-1},$$

da $\mathrm{cov}\big(ps(\boldsymbol{\beta})\big) = \mathrm{cov}\big(s(\boldsymbol{\beta})\big) = F(\boldsymbol{\beta})$.

Die getätigten Approximationen sind (unter gewissen Regularitätsbedingungen) asymptotisch exakt, d.h. mit $\hat{\boldsymbol{\beta}} \stackrel{a}{=} \boldsymbol{\beta}$ gilt $\mathrm{cov}(\hat{\boldsymbol{\beta}}) \stackrel{a}{=} \tilde{F}(\hat{\boldsymbol{\beta}})^{-1} F(\hat{\boldsymbol{\beta}}) \tilde{F}(\hat{\boldsymbol{\beta}})^{-1}$,

so daß es legitim ist, den Sandwich–Schätzer

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) := \tilde{F}(\hat{\boldsymbol{\beta}})^{-1} F(\hat{\boldsymbol{\beta}}) \tilde{F}(\hat{\boldsymbol{\beta}})^{-1} \tag{5.23}$$

zur Approximation der Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ heranzuziehen.

Aus (5.23) lassen sich ferner approximative $(1-\alpha)$–Konfidenzbänder für die geschätzten Effekte der GAM–Komponenten gewinnen. So gilt beispielsweise für die Varianz von $\hat{\gamma}_{(j),r}(x_{ij}) = \boldsymbol{c}_{ij}' \widehat{\tilde{\boldsymbol{\gamma}}}_{jr}$, $j \in \mathcal{S}_{\boldsymbol{x}}$, $r \in \{1,\ldots,q\}$, $i \in \{1,\ldots,N\}$,

$$\sigma^2_{ijr} := \mathrm{Var}\big(\hat{\gamma}_{(j),r}(x_{ij})\big) = \boldsymbol{c}_{ij}' \,\mathrm{cov}\big(\widehat{\tilde{\boldsymbol{\gamma}}}_{jr}\big)\, \boldsymbol{c}_{ij} \quad \text{bzw.} \quad \widehat{\sigma^2_{ijr}} = \boldsymbol{c}_{ij}' \,\widehat{\mathrm{cov}}\big(\widehat{\tilde{\boldsymbol{\gamma}}}_{jr}\big)\, \boldsymbol{c}_{ij}.$$

Ein Schätzer für die Kovarianzmatrix $\mathrm{cov}\big(\widehat{\tilde{\boldsymbol{\gamma}}}_{jr}\big)$ ergibt sich aus (5.23) als entsprechende Teilmatrix von $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}})$. Approximative $(1-\alpha)$–Konfidenzbänder, die symmetrisch um den geschätzten Effekt liegen, liefern die Forderungen

$$P\big[\,\hat{\gamma}_{(j),r}(x_{ij}) - c \;\le\; \gamma_{(j),r}(x_{ij}) \;\le\; \hat{\gamma}_{(j),r}(x_{ij}) + c\,\big] = 1-\alpha\,, \;\; i = 1,\ldots,N.$$

Setzt man für den Parameterschätzer $\hat{\boldsymbol{\beta}}$ asymptotische Erwartungstreue und Normalität voraus, erhält man die punktweisen Konfidenzbänder

$$\hat{\gamma}_{(j),r}(x_{ij}) \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\sigma^2_{ijr}}}\;, \quad i = 1,\ldots,N, \tag{5.24}$$

mit dem $(1-\alpha/2)$–Quantil $z_{1-\alpha/2}$ der Standardnormalverteilung.

5.3.1 Glättungsparameterwahl

Da den Kriterien und Methoden zur Bestimmung der Glättungsparameter in Kapitel 4 keinerlei Annahmen über den Verteilungstyp der abhängigen Variablen zugrunde liegen, kann die Optimierung der λ^g_{jr}, $(j,r) \in \mathcal{S}_{\boldsymbol{x}} \times \{1,\ldots,q\}$ und λ^c_{jr}, $(j,r) \in \mathcal{S}_{\boldsymbol{w}} \times \{1,\ldots,k\}$ auf deren Basis erfolgen. Für

$$\mathrm{FIT}\big(\{\lambda^g_{jr}\},\{\lambda^c_{jr}\}\big)^{-1} = \mathrm{dev}\big(\boldsymbol{y}^*, \hat{\boldsymbol{\beta}}\big) + \gamma \cdot tr(H)$$

ergibt sich die Spur der Hatmatrix zu $tr(H) = tr\big(\tilde{F}\big(\hat{\boldsymbol{\beta}}\big)^{-1} F\big(\hat{\boldsymbol{\beta}}\big)\big)$ (vgl. Kapitel 4). Die Devianz resultiert aus der Betrachtung der Likelihood

$$L(\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_N) = \prod_{i=1}^{N} L_i(\boldsymbol{\pi}_i) = \prod_{i=1}^{N} P_{\boldsymbol{y}}\big(\{\boldsymbol{y}_i\}\big) = \prod_{i=1}^{N} \sum_{r=1}^{k} y_{ir}\pi_{ir}\,, \tag{5.25}$$

wobei $y_{ik} := 1 - \mathbf{1}_q' \boldsymbol{y}_i$ und $\pi_{ik} := 1 - \mathbf{1}_q' \boldsymbol{\pi}_i$. Damit sind

$$l_i(\boldsymbol{y}_i) \;=\; \log\big(L_i(\boldsymbol{y}_i)\big) \;=\; \log\left(\sum_{r=1}^{k} y_{ir}^2\right) \;=\; \log(1) \;=\; 0\,,\;\; i = 1, \ldots, N$$

und

$$l_i(\hat{\boldsymbol{\pi}}_i) \;=\; \log\big(L_i(\hat{\boldsymbol{\pi}}_i)\big) \;=\; \log\left(\sum_{r=1}^{k} y_{ir}\hat{\pi}_{ir}\right) \;=\; \sum_{r=1}^{k} y_{ir}\log(\hat{\pi}_{ir})\,,\;\; i = 1, \ldots, N,$$

mit $\hat{\boldsymbol{\pi}}_i = Z_i\hat{\boldsymbol{\beta}}$ und $\hat{\pi}_{ik} := 1 - \mathbf{1}_q'\hat{\boldsymbol{\pi}}_i$. Für die Devianz folgt somit nach (4.3)

$$\mathrm{dev}\big(\boldsymbol{y}^*, \hat{\boldsymbol{\beta}}\big) \;=\; -2 \cdot \sum_{i=1}^{N} \big\{l_i(\hat{\boldsymbol{\pi}}_i) - l_i(\boldsymbol{y}_i)\big\} \;=\; -2 \cdot \sum_{i=1}^{N}\sum_{r=1}^{k} y_{ir}\log(\hat{\pi}_{ir}).$$

Da die zu bestimmenden Glättungsparameter variablen- und kategorienweise variieren, rechtfertigt die Komplexität des Optimierungsproblems die Anwendung genetischer Algorithmen schon bei nur einem glatten Effekt.

5.3.2 Beispiel: Sichelzellenanämie

In manchen Gegenden Afrikas und Asiens hat ein beträchtlicher Teil der Bevölkerung halbmondförmige rote Blutzellen. Diese als Sichelzellenanämie bekannte Erbkrankheit ist im gesamten Äquatorgürtel des afrikanischen Kontinents, auf Madagaskar sowie in den Malariagebieten von Indien verbreitet. Die Form der roten Blutzellen wird als diploides Merkmal von zwei Erbfaktoren (Genen) mit den krankheitsindizierenden Allelen S und C kontrolliert. Personen, die an Sichelzellenanämie leiden – sogenannte Sichler – sind durch die Genotypen SS, SC und CC charakterisiert. Sichler zeigen stärkere Anfälligkeiten, häufig aber auch größere Resistenzen gegenüber anderen Krankheiten. Da diese Krankheiten von Genotyp zu Genotyp verschieden sind, lassen sich die drei Formen der Sichelzellenanämie nicht ordnen, ihr Erscheinungsbild ist dementsprechend nominalskaliert.

Wir betrachten einen Datensatz über das Auftreten von Sichelzellenanämie, den Adebayo (2001) im Bayesianischen Kontext untersucht. In einer Studie an der Universität von Ilorin (Nigeria) wurden von 85 Sichlern neben ihrem Genotyp die Merkmale Geschlecht (GENDER), Alter in Jahren (AGE) und

Sedimentationsrate (Absinkgeschwindigkeit) der roten Blutzellen in Millimeter je Stunde (ESR) erfasst. Der nominale Charakter der Responsevariablen Genotyp legt ein multinomiales Logit–Modell zur Darstellung der Kovariableneffekte nahe. Da AGE und ESR stetig sind, das Geschlecht der Patienten hingegen binärkodiert vorliegt, ist ein semiparametrischer Ansatz zur Modellierung der Logits angemessen. Deshalb betrachten wir das Modell

$$\log(\pi_r/\pi_k) = \gamma_{0r} + \text{GENDER} \cdot \gamma_{G,r} + \gamma_{(A),r}(\text{AGE}) + \gamma_{(E),r}(\text{ESR}), \; r = SC, CC,$$

mit Referenzkategorie SS und unspezifizierten, funktionalen Effekten für das Alter und die Sedimentationsrate. Diese glatten Komponenten werden mit je zwanzig äquidistanten B–Splines zweiten Grades approximiert. Differenzenpenalties erster Ordnung kontrollieren dabei die Variabilität in den zu schätzenden Effekten. Die Bestimmung der korrespondierenden Glättungsparameter $\lambda^g_{A,SC}$, $\lambda^g_{A,CC}$, $\lambda^g_{E,SC}$ und $\lambda^g_{E,CC}$ erfolgt unter Verwendung des genetischen Algorithmus aus Kapitel 4. Da die Anzahl der Beobachtungen im Vergleich zur Zahl der unbekannten Modellparameter relativ klein ausfällt, wird in der Fitnessfunktion (4.9) die Spur der Hatmatrix mit $\gamma = 4$ bestraft.

Abbildung 5.1 zeigt die maximale Fitness in der aktuellen Population sowie die zugehörigen Werte der Glättungsparameter in Abhängigkeit vom Iterationszähler.

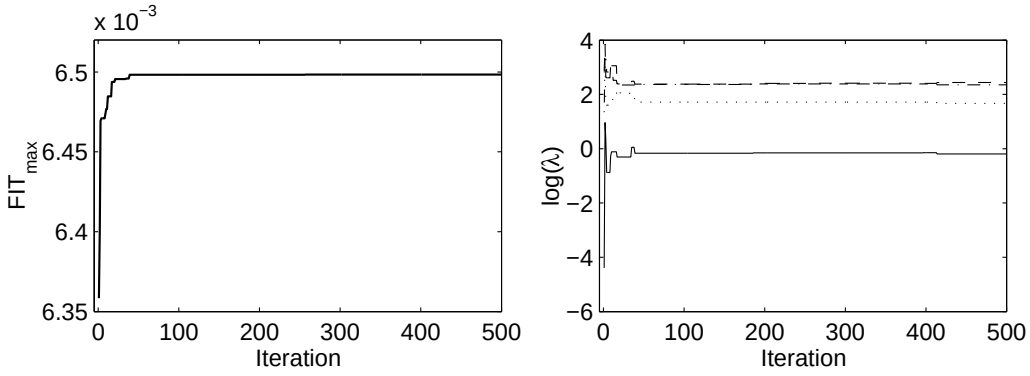

ABBILDUNG 5.1: *Verlauf der maximalen Fitnessfunktion (links) und der korrespondierenden, logarithmierten Glättungsparameter (rechts) im Datensatz zur Sichelzellenanämie.*

Offenbar stellt sich bereits nach wenigen Iterationen ein stabiler Zustand ein, der über die folgenden mehr als 400 Schritte nicht mehr verlassen wird. In der

Tabelle 5.1 sind die diesen Zustand charakterisierenden optimalen Glättungsparameter und der zugehörige Wert der Fitnessfunktion zusammengetragen.

FIT_{opt}	$\lambda^g_{A,SC}$	$\lambda^g_{A,CC}$	$\lambda^g_{E,SC}$	$\lambda^g_{E,CC}$
$6.5 \cdot 10^{-3}$	5.32	10.49	11.46	0.82

TABELLE 5.1: *Optimale Glättungsparameter und zugehöriger Wert der Fitnessfunktion im Datensatz zur Sichelzellenanämie.*

Für die konstanten Terme γ_{0r}, $r \in \{SC, CC\}$, und die kategoriale Kovariable GENDER finden sich die Parameterschätzungen und die mit (5.23) approximierten Standardabweichungen in Tabelle 5.2. Die Ergebnisse implizieren jedoch keinen signifikanten Effekt des Geschlechts auf die modellierten Logits der Genotypen.

	$\hat{\gamma}_{0r}$	$\hat{s}(\hat{\gamma}_{0r})$	$\hat{\gamma}_{G,r}$	$\hat{s}(\hat{\gamma}_{G,r})$
Genotyp SC	-0.945	0.550	0.352	0.564
Genotyp CC	-3.136	1.149	0.715	0.732

TABELLE 5.2: *Parameterschätzungen und geschätzte Standardabweichungen für die parametrischen Komponenten im Datensatz zur Sichelzellenanämie.*

In Abbildung 5.2 sind die mit den optimalen Glättungsparametern aus Tabelle 5.1 geschätzten glatten Effekte des Alters und der Sedimentationsrate auf die logarithmierten Chancen der Genotypen SC und CC gegenüber der Referenzkategorie SS geplottet. Die Darstellungen zeigen neben den Schätzungen approximative Konfidenzbänder gemäß (5.24) mit $\alpha = 0.05$.

Die Gestalt der Schätzung für $\gamma_{(A),SC}(\text{AGE})$ verweist auf einen deutlich nichtlinearen Effekt des Alters in der Kategorie SC. Einem merklichen Anstieg bis zum Alter von ca. 25 Jahren folgt ein allmählicher Abfall des Alterseffekts in dieser Kategorie. Die übrigen geschätzten Funktionen verlaufen monoton im Bereich der jeweiligen Kovariable. Während der Alterseffekt für den Genotyp CC monoton anwächst, zeigt die Sedimentationsrate in beiden modellierten Kategorien einen abfallenden Effekt.

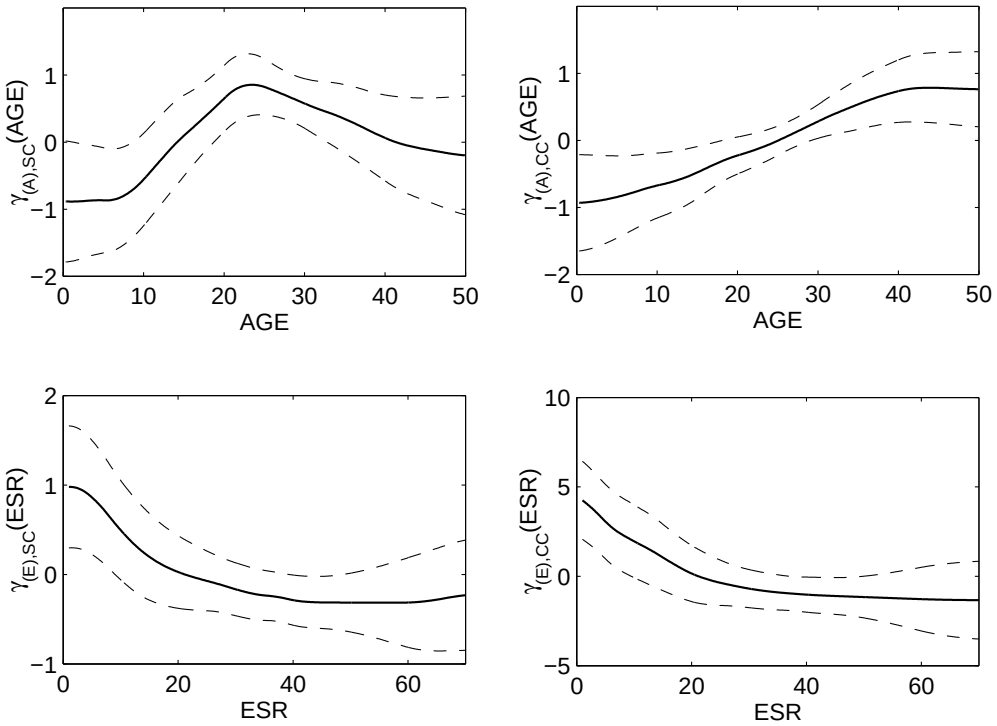

ABBILDUNG 5.2: *Geschätzte glatte Effekte des Alters (oben) und der Sedimentationsrate (unten) auf die Logits der Genotypen im Datensatz zur Sichelzellenanämie. Approximative Konfidenzbänder sind gestrichelt dargestellt.*

5.3.3 Simulation

Zur Untersuchung des multinomialen Logit-Modells mit kategorienspezifischen Charakteristiken simulieren wir Daten aus der nonparametrischen Prädiktorspezifikation

$$\log\left(\frac{\pi_{ir}}{\pi_{ik}}\right) = \eta_{ir} = \gamma_{0r} + \gamma_r(x_i) + \alpha_r(w_{ir}) - \alpha_k(w_{ik}), \quad r = 1, 2, \qquad (5.26)$$

für $i = 1, \ldots, N$. Die $N = 300$ Beobachtungen der globalen Variable x sowie der kategorienspezifischen Charakteristiken w_1, w_2 und w_3 werden jeweils als Realisationen einer auf $[0, 1]$ stetig gleichverteilten Zufallsgröße generiert.

Für die konstanten Terme setzen wir $\gamma_{01} = 0.5$ sowie $\gamma_{02} = -0.5$, als nonparametrische Komponenten werden die folgenden trigonometrischen Funktio-

nen zugrunde gelegt

$$\gamma_1(x) = -\cos(3\pi x), \quad \gamma_2(x) = -\sin(4\pi x),$$
$$\alpha_1(w_1) = \sin(2\pi w_1), \quad \alpha_2(w_2) = \cos(2\pi w_2), \quad \alpha_3(w_3) = \sin(4\pi w_3).$$

Wir ziehen $M = 100$ Stichproben $\left\{\boldsymbol{y}_1^{(m)}, \ldots, \boldsymbol{y}_N^{(m)}\right\}_{m=1,\ldots,M}$ mit je $N = 300$ unabhängigen, multinomialverteilten Responsevektoren, deren Auftretenswahrscheinlichkeiten durch die Logits (5.26) bestimmt sind. Für jede dieser Stichproben schätzen wir ein nonparametrisches, multinomiales Logit–Modell und optimieren die mit den B–Spline Darstellungen der glatten Effekte korrespondierenden fünf Glättungsparameter mit dem genetischen Algorithmus in Kapitel 4. Als Fitnessfunktion wird das reziproke AIC ((4.9) mit $\gamma = 2$) zugrunde gelegt. Die funktionalen Komponenten in (5.26) werden mit je 30 äquidistanten B-Splines dritten Grades approximiert. Variationsbeschränkungen der nonparametrischen Schätzungen gewährleisten Differenzenpenalties der Ordnung zwei.

In den Abbildungen 5.3 und 5.4 sind für sämtliche Kovariablen die aus der Mittelwertbildung der 100 AIC-optimalen Parametervektoren resultierenden geschätzten glatten Effekte zusammen mit der modellierten wahren Funktion dargestellt. Darüber hinaus zeigen die Darstellungen für jede Schätzung den aus der Menge der optimalen Simulationsergebnisse bestimmten Bereich zwischen den punktweise definierten empirischen 5% und 95% Quantilbändern.

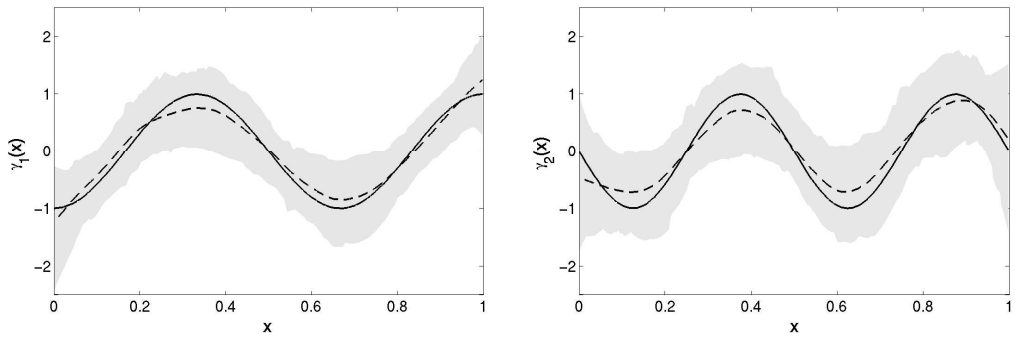

ABBILDUNG 5.3: *Wahre (durchgezogen) und geschätzte Kurve (gestrichelt) für die globale Variable in den Kategorien 1 (links) und 2 (rechts). Grau unterlegte Fläche kennzeichnet Bereich zwischen punktweisen empirischen 5% und 95% Quantilen.*

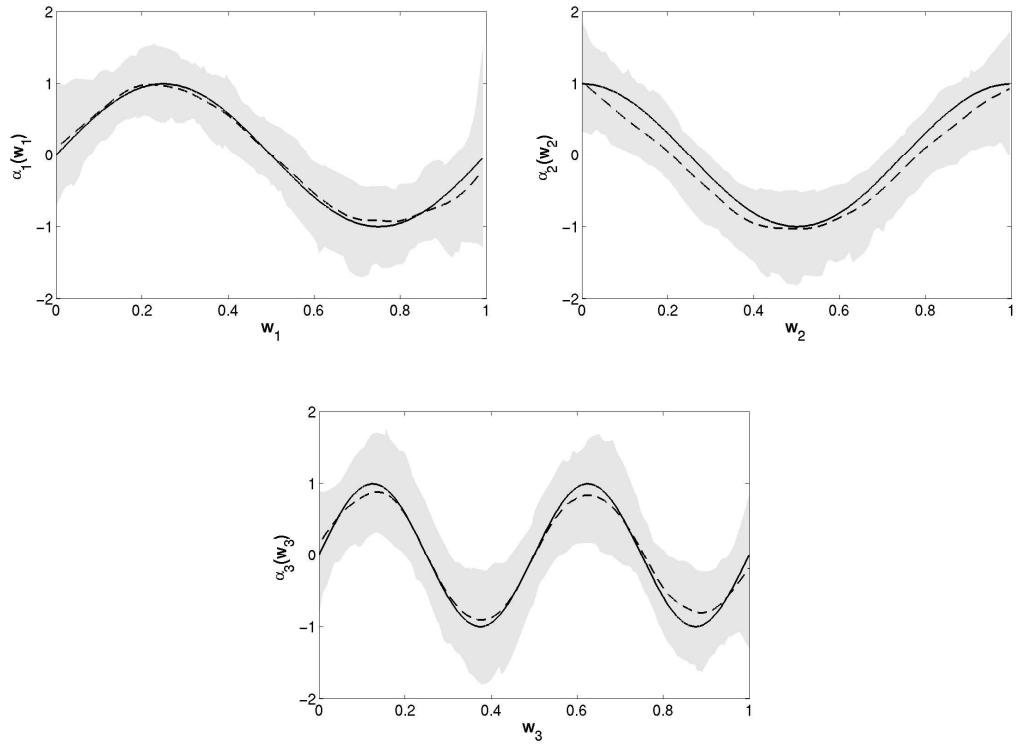

ABBILDUNG 5.4: *Wahre (durchgezogen) und geschätzte Kurve (gestrichelt) für die kategorienspezifischen Charakteristiken in jeder Kategorie. Grau unterlegte Fläche kennzeichnet Bereich zwischen punktweisen empirischen 5% und 95% Quantilen.*

Aus den Abbildungen 5.3 und 5.4 ist ersichtlich, daß die geschätzten glatten Effekte die zugrunde liegenden wahren Funktionen in allen Fällen formgetreu wiedergeben. Für die globale Variable als auch für die kategorienspezifischen Charakteristiken treten Abweichungen zwischen wahrer und gefitteter Kurve hauptsächlich an den Intervallrändern und in Bereichen stärkerer Krümmung auf. Trotz dieser Verzerrungen sind alle zugrunde liegenden wahren Funktionen in dem von den korrespondierenden Quantilbändern eingeschlossenen Bereich enthalten.

6 Modelle mit ordinalem Response

Während kategorial–nominale Variablen und deren flexible Modellierung im Mittelpunkt des letzten Kapitels standen, wenden sich die folgenden Betrachtungen Regressionsmodellen für abhängige Variablen mit ordinalem Skalenniveau zu. Im Kontext einer problemadäquaten Analyse konzentrieren sich die Ausführungen dabei auf Modelle, die die vorliegende Ordnung der Responsekategorien explizit voraussetzen und nutzen. Obwohl das multinomiale Logit-Modell (5.3) im Fall ordinaler Responsevariablen prinzipiell anwendbar ist, kommt es als Modellierungsform hier nicht in Betracht, da bei diesem Ansatz auf die in der Ordnungsstruktur enthaltene zusätzliche Information verzichtet wird.

Ordinale Modelle sind darüber hinaus zu der auf stärkeren Voraussetzungen basierenden metrischen Regression abzugrenzen. Die Ausprägungen ordinaler Variablen lassen sich zwar miteinander vergleichen, metrisches Skalenniveau, das zudem eine sinnvolle Interpretation von Abständen gestattet, wird jedoch nicht unterstellt. Resultieren die geordneten Kategorien aus der Diskretisierung eines ursprünglich stetigen Merkmals mit unbeschränktem Träger, ist die strikte Abgrenzung zum klassischen linearen Modell aus Abschnitt 1.1 besonders relevant. Da die erste und/oder letzte Kategorie in diesem Fall einem unbeschränkten Intervall entsprechen, ist – selbst bei feiner Intervallbildung – die Annahme eines metrischen Response nicht gerechtfertigt.

6.1 Das kumulative Logit–Modell

Die von McCullagh (1980) propagierten kumulativen Modelle gehören zu den wohl gebräuchlichsten ordinalen Regressionsmodellen. Für kategorial-ordinalen Response $\tilde{y} \in \{1, \ldots, k\}$ werden die kumulierten Responsewahrscheinlichkeiten dabei als

$$P(\tilde{y} \leq r \mid \boldsymbol{x}) \;=\; F\big(\eta_r(\boldsymbol{x})\big)\,, \;\; r = 0, \ldots, k, \tag{6.1}$$

modelliert, mit zunächst beliebiger, den konkreten Modelltyp bestimmender Verteilungsfunktion F, Kovariablenvektor $\boldsymbol{x} = (x_1, \ldots, x_p)'$ und den Festle-

gungen $\eta_0(\boldsymbol{x}) := -\infty$ sowie $\eta_k(\boldsymbol{x}) := \infty$. Wählt man F als logistische Funktion $F(x) := \exp(x)/\big(1+\exp(x)\big)$, erhält man das kumulative Logit–Modell, auf das sich die Ausführungen im folgenden beschränken werden.

Die allgemeinste (parametrische) Form des kumulativen Logit-Modells spezifiziert die Einflüsse der Kovariablen linear: $\eta_r(\boldsymbol{x}) = \gamma_{0r} + \boldsymbol{x}'\boldsymbol{\gamma}_r$, $r = 1, \ldots, q$, $q := k-1$, mit Effekten $\boldsymbol{\gamma}_r = (\gamma_{1r}, \ldots, \gamma_{pr})'$, die von der jeweiligen Responsekategorie r abhängen. Die daraus resultierende Schreibweise

$$\log\big(P(\tilde{y} \leq r \mid \boldsymbol{x}) \,/\, P(\tilde{y} > r \mid \boldsymbol{x})\big) \;=\; \gamma_{0r} + \boldsymbol{x}'\boldsymbol{\gamma}_r, \;\; r = 1, \ldots, q, \tag{6.2}$$

für die kumulierten logarithmierten Chancen motivierte die Bezeichnung des Modells und impliziert, daß jede Dichotomisierung der Responsevariablen in $\tilde{y} \leq r$ und $\tilde{y} > r$, $r = 1, \ldots, q$, durch eine Linearkombination der Kovariablen bestimmt ist, mit Koeffizienten, die spezifisch sind für die Kategorie, an der gesplittet wird.

Eine naheliegende Vereinfachung von (6.2) liefert die Annahme globaler, d.h. kategorienunabhängiger Kovariableneinflüsse

$$\log\big(P(\tilde{y} \leq r \mid \boldsymbol{x}) \,/\, P(\tilde{y} > r \mid \boldsymbol{x})\big) \;=\; \gamma_{0r} + \boldsymbol{x}'\boldsymbol{\gamma}. \tag{6.3}$$

Die globale Modellierung der Effekte zeichnet sich durch eine Reihe von Vorteilen aus. So wie sich das multinomiale Logit-Modell (5.3) als Zufallsnutzen-Modell interpretieren ließ, kann auch (6.3) durch die Existenz einer nicht beobachtbaren metrischen Größe motiviert werden (Tutz, 2000). Der beobachtbare Response wird dabei als kategorisierte Version einer dahinterstehenden latenten Variablen aufgefasst, wobei die Verknüpfung der beiden Größen nach dem sogenannten *Schwellenwertkonzept* erfolgt (Tutz, 2000).

Eine weitere Folge globaler Kovariableneffekte ist die sogenannte stochastische Ordnung der Kategorien (McCullagh, 1980), die sich für das kumulative Logit–Modell in Form proportionaler Chancen ausdrückt. Für zwei durch $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ gekennzeichnete Individuen ist die Differenz ihrer kumulierten Logits (6.3) demnach unabhängig von der Kategorie r, an der dichotomisiert wird

$$\log\left(\frac{P(\tilde{y} \leq r \mid \boldsymbol{x}_2)/P(\tilde{y} > r \mid \boldsymbol{x}_2)}{P(\tilde{y} \leq r \mid \boldsymbol{x}_1)/P(\tilde{y} > r \mid \boldsymbol{x}_1)}\right) = (\boldsymbol{x}_2 - \boldsymbol{x}_1)'\boldsymbol{\gamma}. \tag{6.4}$$

Das Vorliegen identischer (globaler) Chancenverhältnisse prägte die für (6.3) ebenfalls gebräuchliche Bezeichnung als *Proportional Odds Modell (POM).*

Darstellung (6.4) ermöglicht darüber hinaus einfache Interpretationen für die Kovariableneffekte. Unterscheiden sich $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ via $\boldsymbol{x}_2 - \boldsymbol{x}_1 = \boldsymbol{e}_{j,p}$ lediglich in der j-ten Komponente, so ist γ_j, der Effekt der j-ten Kovariable, als logarithmiertes Chancenverhältnis für den Übergang von x_j zum um eine Einheit größeren $x_j + 1$ gegeben.

In vielen Fällen erweist sich die Annahme proportionaler Chancen jedoch als nicht adäquat. Tabelle 6.1 zeigt Daten über das Auftreten von Erkrankungen der Herzkranzgefäße, die am Duke University Medical Center in Durham erhoben wurden (Quelle: Peterson & Harrell, 1990).

	Schweregrad der Herzkranzgefäßerkrankung				
	1	2	3	4	5
Nichtraucher	334	99	117	159	30
Raucher	350	307	345	481	67

TABELLE 6.1: *Kontingenztafel der Merkmale Raucherstatus und Herzkranzgefäßerkrankung. Der Schweregrad 1 steht für keine Erkrankung, die Kategorien 4 und 5 bezeichnen ernsthafte Erkrankungen der Herzkranzgefäße.*

Die Betrachtung einer ordinalen Maßzahl zur Bewertung des Schweregrades der Erkrankung und einer dichotomen Größe zur Kodierung des Raucherstatus (0: Nichtraucher, 1: Raucher) führt auf die vier empirischen Chancenverhältnisse 0.35, 0.52, 0.63 sowie 0.94. Für die Populationen der Nichtraucher und der Raucher steigt demzufolge die relative Chance, einen Erkrankungsgrad von maximal r gegenüber einem von mindestens $r+1$ aufzuweisen, mit zunehmendem Schweregrad r der Erkrankung. Dieses Wachstum spricht gegen einen globalen Effekt des Raucherstatus.

Obiges Beispiel verdeutlicht die Notwendigkeit, die Annahme proportionaler Chancen auf der Basis von Signifikanztests statistisch zu validieren. Kandidaten hierfür sind Likelihood–Quotienten– (LQ), Wald– sowie Score–Tests. Die

Ableitung entsprechender Prüfgrößen bzw. Test–Statistiken orientiert sich an der Darstellung der zu testenden Annahmen als lineares Hypothesenpaar. So ist für die hier relevante Fragestellung das Testproblem

$$H_0^* : \eta_r(\boldsymbol{x}) = \gamma_{0r} + \boldsymbol{x}'\boldsymbol{\gamma} \qquad \text{vs.} \qquad H_1^* : \eta_r(\boldsymbol{x}) = \gamma_{0r} + \boldsymbol{x}'\boldsymbol{\gamma}_r,$$

$r = 1, \ldots, q$, äquivalent zum Testen der linearen Hypothesen

$$H_0 : C\boldsymbol{\beta} = \mathbf{0} \qquad \text{vs.} \qquad H_1 : C\boldsymbol{\beta} \neq \mathbf{0}$$

im H_1^*–Modell mit Parametervektor $\boldsymbol{\beta} := \left(\gamma_{01}, \ldots, \gamma_{0q}, \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_q'\right)'$ und einer Matrix $C := \left[\, 0_{p(q-1)\times q} \mid D_q^1 \otimes I_p \,\right] \in M_{I\!R}\big(p(q-1), q(p+1)\big)$, wobei D_q^1 die mit (2.20) definierte $(q-1)\times q$ Kontrastmatrix bezeichnet. Die in einem späteren Abschnitt demonstrierte, mögliche Schätzung von (6.2) und (6.3) im Rahmen multivariater GLM's, gestattet dann die Definition der folgenden Statistiken:

Der *klassische Likelihood–Quotient*

$$LR := -2 \cdot \left\{ l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}) \right\}$$

stellt das unrestringierte Maximum $l(\hat{\boldsymbol{\beta}})$ der Log–Likelihood des H_1^*–Modells dem Maximum $l(\tilde{\boldsymbol{\beta}})$ gegenüber, das aus der Maximierung der Log–Likelihood des H_1^*–Modells unter zusätzlicher Beachtung der in H_0 formulierten Restriktion $C\boldsymbol{\beta} = \mathbf{0}$ resultiert. Da die restringierte Schätzung im H_1^*–Modell äquivalent zur Schätzung unter H_0^* ist, kann $l(\tilde{\boldsymbol{\beta}})$ auch durch das (identische) Maximum der Log–Likelihood des H_0^*–Modells ersetzt werden.

Die *klassische Wald–Statistik*

$$W := \hat{\boldsymbol{\beta}}' C' \big(C\, F(\hat{\boldsymbol{\beta}})^{-1} C'\big)^{-1} C\hat{\boldsymbol{\beta}}, \text{ mit } F(\boldsymbol{\beta}) = E_{H_1}\big[-\partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\big],$$

misst die gewichtete Distanz zwischen dem Schätzer $C\hat{\boldsymbol{\beta}}$ und dessen hypothetischem Wert $\mathbf{0}$ unter H_0. Das Gewicht wird durch die Inverse der asymptotischen Kovarianzmatrix $C\, F(\hat{\boldsymbol{\beta}})^{-1} C'$ von $C\hat{\boldsymbol{\beta}}$ definiert.

Die *klassische Score–Statistik*

$$s := s(\tilde{\boldsymbol{\beta}})' F(\tilde{\boldsymbol{\beta}})^{-1} s(\tilde{\boldsymbol{\beta}})$$

misst die Distanz zwischen der Score–Funktion $s(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ ausgewertet am Maximum–Likelihood–Schätzer $\tilde{\boldsymbol{\beta}}$ des unter H_0 geschätzten H_1^*–Modells und $s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. Die Inverse $F(\tilde{\boldsymbol{\beta}})^{-1}$, mit $F(\boldsymbol{\beta})$ wie oben definiert, dient bei dieser Distanzmessung als zusätzliche Gewichtungsmatrix.

Obwohl der Likelihood–Quotienten–Test attraktivere statistische Eigenschaften als der Wald– und der Score–Test besitzt, weist er dennoch zwei entscheidende Nachteile auf. Zum einen ist seine numerische Umsetzung aufwendiger, da die Likelihood des H_1^*–Modells zweimal zu maximieren ist. Weitaus schwerer wiegen jedoch die auch den Wald–Test betreffenden, numerischen Probleme, die bei der unrestringierten Schätzung des H_1^*–Modells auftreten können:

Für die mit (6.1) definierten kumulativen Modelle läßt sich folgende Darstellung der entsprechenden Responsewahrscheinlichkeiten ableiten

$$P(\tilde{y} = r \mid \boldsymbol{x}) \;=\; F\big(\eta_r(\boldsymbol{x})\big) - F\big(\eta_{r-1}(\boldsymbol{x})\big)\,, \quad r = 1, \ldots, k. \tag{6.5}$$

Mit der axiomatisch fixierten Nichtnegativität von Wahrscheinlichkeiten und der Monotonie von Verteilungsfunktionen folgt $\eta_1(\boldsymbol{x}) \leq \ldots \leq \eta_q(\boldsymbol{x})$ als Ordnungsrestriktion an die Prädiktoren. Während sich diese Restriktion für das Proportional Odds Modell auf $\gamma_{01} \leq \ldots \leq \gamma_{0q}$ reduziert, ist im allgemeineren Modell (6.2) mit kategorienspezifischen Parametern die Bedingung

$$\gamma_{01} + \boldsymbol{x}'\boldsymbol{\gamma}_1 \;\leq\; \ldots \;\leq\; \gamma_{0q} + \boldsymbol{x}'\boldsymbol{\gamma}_q \tag{6.6}$$

für jede Kovariablenkonstellation $\boldsymbol{x}$ zu gewährleisten. Bei iterativen Schätzverfahren wie dem Fisher–Scoring besteht die Gefahr der Divergenz bzw. des Schätzens von Artefakten, wenn (6.6) nicht in jedem Iterationsschritt sichergestellt wird. Eine einfache Möglichkeit zur Gewährleistung dieser Restriktion ist mit der wiederholten „Schrittweitenhalbierung“ zwischen aufeinanderfolgenden Iterierten gegeben. Häufig ist damit aber auch eine drastische Verschlechterung der Konditionierung des Schätzproblems verbunden. Beim Fisher–Scoring äußert sich dieser Umstand in sprunghaft anwachsenden, spektralen Konditionszahlen (Golub, 1989) der Fisher'schen Informationsmatrix. Deren damit einhergehende numerische Singularität kann oft auch durch die Skalierung von Beobachtungen nicht vermieden werden und führt zwangsläufig zum vorzeitigen Abbruch des Algorithmus, so daß Parameterschätzungen dann nicht verfügbar sind.

Aufgrund der genannten numerischen Probleme ist man in vielen Situationen auf inadäquate Alternativen angewiesen, um überhaupt Schätzwerte für kategorienspezifische Effekte zu erhalten. Zu erwähnen ist in diesem Zusammenhang – neben dem multinomialen Logit–Modell – die separate Betrachtung

der q binären Logit–Modelle, die aus den möglichen Dichotomisierungen der Responsevariablen resultieren. Obwohl dieser Ansatz asymptotische Kovarianzen für das allgemeine Modell (6.2) liefert (Brant, 1990), können die separaten Schätzungen nicht zur Konstruktion von Schätzungen für das geschlossene Modell herangezogen werden. Da für einen Likelihood–Quotienten-Test die Verfügbarkeit ebendieser aber zwingend erforderlich ist, kann dieser Test bei numerischen Problemen nicht zur Anwendung kommen.

In der Praxis wird das Vorliegen identischer Chancenverhältnisse daher meist auf der Basis des Score–Tests überprüft, der lediglich die – nahezu immer garantierte – Verfügbarkeit von Schätzungen im POM voraussetzt. Unabhängig von der gesicherten Durchführbarkeit stellt sich jedoch die Frage, wie im Falle einer Ablehnung der Hypothese proportionaler Chancen durch den Score–Test weiter zu verfahren ist. Eine problemadäquate Analyse erfordert die Bestimmung kategorienspezifischer Effekte, die dafür notwendige Maximierung der Likelihood ist jedoch unter Umständen numerisch nicht umsetzbar.

Die bisherigen Betrachtungen fokussierten ausschließlich die Reinformen globaler und kategorienspezifischer Modellierung. Denkbar sind jedoch auch Situationen, in denen lediglich ein Teil der Kovariablen globale Parameter aufweist, während für die verbleibenden ein über die Kategorien hinweg variabler Effekt unterstellt wird. Sei dazu $\mathcal{P} := \{1, \ldots, p\}$ die Indexmenge aller p Kovariablen und $\mathcal{G} \subseteq \mathcal{P}$ eine Teilmenge. Der allgemeinste hier betrachtete Fall ist mit (6.2) gegeben, das im folgenden auch als *Non-Proportional Odds Modell (NPOM)* bezeichnet wird. Als *Partial Proportional Odds Modell (PPOM($\mathcal{G}$))* sei das kumulative Logit–Modell definiert, in dem alle Kovariablen x_j, $j \in \mathcal{G}$, globale Effekte besitzen, d.h. die Gültigkeit der Hypothese

$$H_{\mathcal{G}}: \ \gamma_{j1} = \ldots = \gamma_{jq} \tag{6.7}$$

wird für alle $j \in \mathcal{G}$ angenommen. Mit diesen Definitionen sind das Proportional Odds Modell äquivalent zum PPOM($\mathcal{P}$) und das Non-Proportional Odds Modell seinerseits äquivalent zum PPOM($\emptyset$).

Die Modellfamilie $\{\text{PPOM}(\mathcal{G}) : \mathcal{G} \subseteq \mathcal{P}\}$ ist mit PPOM($\emptyset$) = NPOM als allgemeinstem und PPOM($\mathcal{P}$) = POM als einfachstem Modell nur teilweise hierarchisch. Für zwei nichtleere, echte Teilmengen $\mathcal{G}_1$, $\mathcal{G}_2$ von $\mathcal{P}$ läßt sich lediglich

im Fall $\mathcal{G}_2 \subseteq \mathcal{G}_1$ via PPOM($\mathcal{G}_1$) $\subseteq$ PPOM($\mathcal{G}_2$) eine Aussage hinsichtlich einer gewissen Ordnungsstruktur treffen. Die Modelle aller sonstigen Konstellationen können in keine Rangfolge gebracht werden. Im Bestreben, den Prädiktor möglichst einfach, hier also weitestgehend global zu gestalten, sind daher alle Hypothesenpaare $\{H_{\mathcal{G}}$ *versus* $H_{\emptyset}\}$, $\mathcal{G} \subseteq \mathcal{P}$, zu betrachten. Die Durchführung aller erforderlichen Tests ist zum einen sehr aufwendig, andererseits läßt sich das Schätzen kategorienspezifischer Gewichte – abgesehen von $\mathcal{G} = \mathcal{P}$ – selbst unter der Nullhypothese nicht umgehen. Da die angesprochenen numerischen Probleme ebenso für Partial Proportional Odds Modelle relevant sind, ist eine Verfügbarkeit von Schätzungen im PPOM($\mathcal{G}$) und damit selbst die Durchführbarkeit eines Score–Tests nicht gesichert.

Im folgenden wird eine Methode zur numerischen Stabilisierung des Fisher–Scoring für die Schätzung in Partial Proportional Odds Modellen vorgestellt, mit der Schätzwerte kategorienspezifischer Parameter auch in kritischen Fällen gewonnen werden können. Der vorgeschlagene Ansatz basiert auf dem bereits bekannten Konzept der penalisierten Likelihood. Im Unterschied zu den P–Splines wirken Differenzenpenalties dabei prädiktorübergreifend als Variationsbeschränkung der kategorienspezifischen Parameter.

6.2 Penalisierte Schätzungen im PPOM

Betrachtet werde eine konkrete Datensituation $\{\tilde{y}_i, \boldsymbol{x}_i\}_{i=1,\ldots,N}$. Im allgemeinen Partial Proportional Odds Modell PPOM($\mathcal{G}$), $\mathcal{G} \subseteq \mathcal{P}$, mit globalen Effekten für x_j, $j \in \mathcal{G}$, und kategorienspezifischen Gewichten für x_j, $j \in \bar{\mathcal{G}} := \mathcal{P} \backslash \mathcal{G}$, lassen sich die Prädiktoren η_{ir} gemäß

$$\eta_{ir} \;=\; \gamma_{0r} \;+\; \boldsymbol{x}'_{i,\mathcal{G}}\boldsymbol{\gamma}_{\mathcal{G}} \;+\; \boldsymbol{x}'_{i,\bar{\mathcal{G}}}\boldsymbol{\gamma}_{\bar{\mathcal{G}},r}\,, \quad r = 1,\ldots,q, \tag{6.8}$$

partitionieren, wobei im Vektor $\boldsymbol{x}_{i,\mathcal{G}}$ die Variablen x_{ij}, $j \in \mathcal{G}$, und in $\boldsymbol{x}_{i,\bar{\mathcal{G}}}$ die Variablen x_{ij}, $j \in \bar{\mathcal{G}}$, zusammengefaßt sind. Für $\boldsymbol{\gamma}_0 := (\gamma_{01},\ldots,\gamma_{0q})'$ und den Parametervektor $\boldsymbol{\beta} := (\boldsymbol{\gamma}'_0, \boldsymbol{\gamma}'_{\mathcal{G}}, \boldsymbol{\gamma}'_{\bar{\mathcal{G}},1},\ldots,\boldsymbol{\gamma}'_{\bar{\mathcal{G}},q})'$ sowie eine Designmatrix

$$Z_i \;:=\; \left[\, I_q \;\mid\; \mathbf{1}_q \otimes \boldsymbol{x}'_{i,\mathcal{G}} \;\mid\; I_q \otimes \boldsymbol{x}'_{i,\bar{\mathcal{G}}} \,\right],$$

kann $\boldsymbol{\eta}_i := (\eta_{i1},\ldots,\eta_{iq})'$ auch als $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$ geschrieben werden. Wird ferner

mit $\boldsymbol{\pi}_i := (\pi_{i1}, \ldots, \pi_{iq})'$ der Vektor der Responsewahrscheinlichkeiten mit den Komponenten $\pi_{ir} := P(\tilde{y}_i = r|\boldsymbol{x}_i)$, $r = 1, \ldots, q$, bezeichnet, so können via

$$h_r(\boldsymbol{\eta}_i) := \frac{\exp(\eta_{ir}) - \exp(\eta_{i,r-1})}{\big(1 + \exp(\eta_{ir})\big)\big(1 + \exp(\eta_{i,r-1})\big)}, \quad r = 1, \ldots, q,$$

für die Komponenten h_r der Responsefunktion $h : I\!R^q \longrightarrow I\!R^q$ und via

$$g_r(\boldsymbol{\pi}_i) := \log(\pi_{i1} + \ldots + \pi_{ir}) - \log(1 - \pi_{i1} - \ldots - \pi_{ir}), \quad r = 1, \ldots, q,$$

für die Komponenten g_r der Linkfunktion $g : I\!R^q \longrightarrow I\!R^q$ die Beziehungen

$$\boldsymbol{\pi}_i = h(\boldsymbol{\eta}_i) = h(Z_i\boldsymbol{\beta}) \quad \text{und} \quad g(\boldsymbol{\pi}_i) = \boldsymbol{\eta}_i = Z_i\boldsymbol{\beta},$$

für das PPOM($\mathcal{G}$) (6.8) abgeleitet werden. Mit der Multinomialverteilung als zugrunde liegender Responseverteilung (vgl. Abschnitt 5.2) besitzt also auch das PPOM($\mathcal{G}$) alle Merkmale (multivariater) generalisierter linearer Modelle. Die numerische Maximierung der zugehörigen Log–Likelihood $l_{\mathcal{G}}(\boldsymbol{\beta})$ über das klassische Fisher–Scoring ist jedoch aufgrund der erläuterten Probleme beim Vorliegen kategorienspezifischer Parameter nicht immer möglich.

Statt der Log–Likelihood $l_{\mathcal{G}}(\boldsymbol{\beta})$ wird hier die penalisierte Version

$$pl_{\mathcal{G}}(\boldsymbol{\beta}) = l_{\mathcal{G}}(\boldsymbol{\beta}) - \frac{1}{2} \cdot P_{\boldsymbol{\beta}} \tag{6.9}$$

betrachtet. Für $\bar{\mathcal{G}} := \{j_1, \ldots, j_{|\bar{\mathcal{G}}|}\} \subseteq \mathcal{P}$ ist der Penalty dabei gegeben als

$$P_{\boldsymbol{\beta}} = \sum_{j \in \{0\} \cup \bar{\mathcal{G}}} \lambda_{j,\mathcal{G}} \sum_{r=1}^{q-1} (\Delta^1 \gamma_{jr})^2 = \lambda_{0,\mathcal{G}} \sum_{r=1}^{q-1} (\Delta^1 \gamma_{0r})^2 + \sum_{s=1}^{|\bar{\mathcal{G}}|} \lambda_{j_s,\mathcal{G}} \sum_{r=1}^{q-1} (\Delta^1 \gamma_{j_s r})^2.$$

Der erste Teil repräsentiert einen Differenzenpenalty für die γ_{0r}, $r = 1, \ldots, q$, die im kumulativen Logit-Modell auch als *Schwellenwerte* bezeichnet werden. Durch die zweite Summe werden für jedes $s \in \{1, \ldots, \bar{\mathcal{G}}\}$ darüber hinaus die ersten Differenzen $\Delta^1 \gamma_{j_s r} := \gamma_{j_s,r+1} - \gamma_{j_s r}$ der s–ten Komponenten von $\boldsymbol{\gamma}_{\bar{\mathcal{G}},r+1}$ und $\boldsymbol{\gamma}_{\bar{\mathcal{G}},r}$, $r = 1, \ldots, q$, bestraft.

Penalisiert werden demnach die ersten Differenzen von kategorienspezifischen Parametern benachbarter Responsekategorien. Nachbarschaften ergeben sich dabei auf natürliche Weise aus der Ordnung der Kategorien.

Die Stärke der Penalisierung wird wie schon bei den P–Splines über die Glättungsparameter $\lambda_{j,\mathcal{G}}$, $j \in \{0\} \cup \bar{\mathcal{G}}$, gesteuert. Für $\lambda_{j,\mathcal{G}} = 0$, $j \in \{0\} \cup \bar{\mathcal{G}}$, entspricht die Maximierung von (6.9) einer Maximierung der üblichen, unpenalisierten Log-Likelihood und damit der gängigen Schätzung im PPOM($\mathcal{G}$), die mit den erläuterten (numerischen) Problemen behaftet ist. Im Spezialfall unendlich starker Bestrafung der kategorienspezifischen Parameter, $\lambda_{j,\mathcal{G}} \longrightarrow \infty$, $j \in \bar{\mathcal{G}}$, wird hingegen das Modell mit ausschließlich globalen Effekten gefittet, dessen Schätzung im allgemeinen unproblematisch ist. Gesucht ist daher eine Glättungsparameterkonstellation zwischen diesen Extremen, die die Konvergenz des Fisher–Scoring gewährleistet. Die Wahl der Glättungsparameter erfolgt dabei ausschließlich unter dem Gesichtspunkt einer numerischen Stabilisierung, nicht jedoch im Sinne eines zu optimierenden Kriteriums wie im Kapitel 4.

Der Penalty $P_{\boldsymbol{\beta}}$ umfasst zwar auch die Parameter $\gamma_{01}, \ldots, \gamma_{0q}$, für die Stabilisierung des Fisher–Scoring ist deren Penalisierung jedoch nicht erforderlich. Für $P_{\boldsymbol{\beta}}$ läßt sich mit der in (2.20) definierten Kontrastmatrix eine kompaktere Darstellung in Matrixschreibweise ableiten. Setzt man

$$\tilde{D} := D_q^1 \otimes \operatorname{diag}\big(\lambda_{j_1,\mathcal{G}}^{1/2}, \ldots, \lambda_{j_{|\bar{\mathcal{G}}|},\mathcal{G}}^{1/2}\big),$$

so ist

$$P_{\boldsymbol{\beta}} \;=\; \boldsymbol{\beta}' \operatorname{Diag}\big(\lambda_{0,\mathcal{G}}(D_q^1)'D_q^1, 0_{|\mathcal{G}|\times|\mathcal{G}|}, \tilde{D}'\tilde{D}\big)\,\boldsymbol{\beta} \;=:\; \boldsymbol{\beta}' K_{\mathcal{G}} \boldsymbol{\beta}.$$

Die Maximierung von $pl_{\mathcal{G}}(\boldsymbol{\beta})$ führt auf die Schätzgleichungen $ps_{\mathcal{G}}(\boldsymbol{\beta}) = \mathbf{0}$, wobei die penalisierte Score–Funktion $ps_{\mathcal{G}}(\boldsymbol{\beta}) = \partial pl_{\mathcal{G}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ die Form

$$ps_{\mathcal{G}}(\boldsymbol{\beta}) \;=\; s_{\mathcal{G}}(\boldsymbol{\beta}) \;-\; K_{\mathcal{G}}\boldsymbol{\beta} \;=\; \sum_{i=1}^{N} Z_i' D_i \Sigma_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) \;-\; K_{\mathcal{G}}\boldsymbol{\beta}$$

hat, mit $D_i = \partial h(\boldsymbol{\eta}_i)/\partial\boldsymbol{\eta}_i$, $\Sigma_i = \operatorname{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'$ und den in (5.14) definierten multivariaten Pendants $\boldsymbol{y}_i := (y_{i1}, \ldots, y_{iq})'$ von $\tilde{y}_i$, $i = 1, \ldots, N$. Als Schätzverfahren wird ein modifiziertes Fisher–Scoring betrachtet, bei dem sich, ausgehend von einem Startwert $\hat{\boldsymbol{\beta}}^{(0)}$, Parameter–Updates gemäß der Vorschrift

$$\hat{\boldsymbol{\beta}}^{(k+1)} \;=\; \hat{\boldsymbol{\beta}}^{(k)} \;+\; \tilde{F}_{\mathcal{G}}\Big(\hat{\boldsymbol{\beta}}^{(k)}\Big)^{-1} ps_{\mathcal{G}}\Big(\hat{\boldsymbol{\beta}}^{(k)}\Big), \;\; k = 0, 1, 2, \ldots, \tag{6.10}$$

berechnen, wobei $\tilde{F}_{\mathcal{G}}(\boldsymbol{\beta}) = F_{\mathcal{G}}(\boldsymbol{\beta}) + K_{\mathcal{G}} = \sum_{i=1}^{N} Z_i' D_i \Sigma_i^{-1} D_i' Z_i + K_{\mathcal{G}}$.

Die Einhaltung der Ordnungsrestriktion $\eta_{i1}^{(k)} \leq \ldots \leq \eta_{iq}^{(k)}$, $i = 1, \ldots, N$, wird in jeder Iteration überprüft und gegebenenfalls durch wiederholte Halbierung der Schrittweite via

$$\hat{\boldsymbol{\beta}}_{\text{neu}}^{(k+1)} := \frac{1}{2} \cdot \left(\hat{\boldsymbol{\beta}}_{\text{alt}}^{(k+1)} + \hat{\boldsymbol{\beta}}^{(k)}\right)$$

erzwungen. Daraus resultierenden numerischen Problemen wirken die Penalties auf den kategorienspezifischen Parametern entgegen. Da die penalisierten Schätzungen denen des numerisch stabileren POM mit zunehmender Bestrafung immer ähnlicher werden, existieren im allgemeinen Glättungsparameter $\lambda_{j,\mathcal{G}}^0 > 0$, so daß das Fisher–Scoring (6.10) für $\lambda_{j,\mathcal{G}} \geq \lambda_{j,\mathcal{G}}^0$, $j \in \bar{\mathcal{G}}$, konvergiert.

6.2.1 Simulation: Potential penalisierter Schätzungen

In einer kleinen Simulationsstudie soll das Potential von penalisierten Schätzungen bewertet werden. Dazu wird ein NPOM mit der Prädiktorspezifikation

$$\begin{aligned} \eta_{i1} &= -0.8 + 0.7 \cdot x_i \\ \eta_{i2} &= -0.4 + 0.4 \cdot x_i \\ \eta_{i3} &= 0.2 + 0.1 \cdot x_i \end{aligned}$$

als zugrunde liegendes, wahres Modell betrachtet, wobei die Kovariablen x_i, $i = 1, \ldots, 300$, einer Gleichverteilung auf $[-1, 1]$ entstammen. Abbildung 6.1 erlaubt einen direkten Vergleich obiger Prädiktoren. Im relevanten Kovariablenintervall genügt das angesetzte Non-Proportional Odds Modell offensichtlich der erforderlichen Ordnungsrestriktion.

Von 100 Ziehungen aus obiger Prädiktorspezifikation schlug das klassische Fisher–Scoring in 17 Fällen aufgrund numerischer Singularitäten bei dem Versuch fehl, ein Non-Proportional Odds Modell zu schätzen. Für diese 17 kritischen Ziehungen konnte die Konvergenz des Verfahrens jedoch über die Penalisierung des kategorienspezifischen Kovariableneffekts sichergestellt werden. Um die Qualität der penalisierten Schätzungen bewerten zu können, wurden diese einem Vergleich mit den (hier inadäquaten) Parameterschätzungen des Proportional Odds Modells unterzogen.

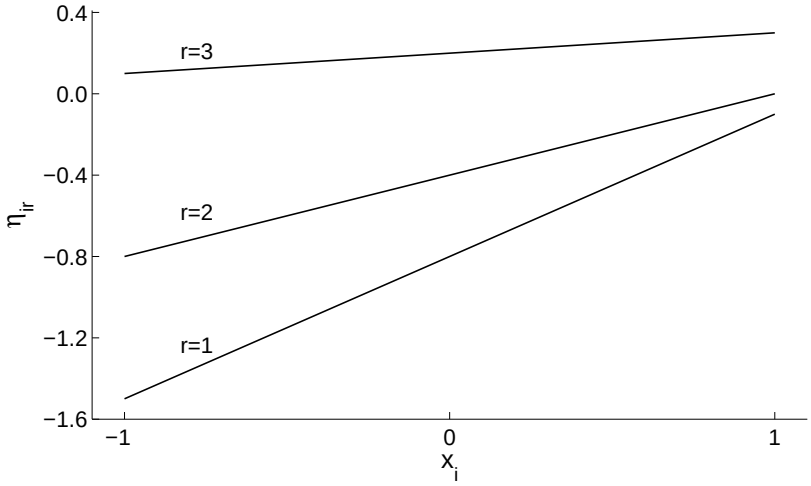

ABBILDUNG 6.1: *Verlauf der Prädiktorwerte in Abhängigkeit von den Kovariablenbeobachtungen für die Kategorien* $r = 1, 2, 3$.

Der Vergleich erfolgte auf der Basis verschiedener Verlustfunktionen. Im einzelnen waren dies (vgl. Santner & Duffy, 1989):

Mean Squared Error Loss

$$\text{MSEL} = \frac{1}{N} \sum_{i=1}^{N} \sum_{r=1}^{k} (\pi_{ir} - \hat{\pi}_{ir})^2,$$

Mean Relative Squared Error Loss

$$\text{MRSEL} = \frac{1}{N} \sum_{i=1}^{N} \sum_{r=1}^{k} \frac{(\pi_{ir} - \hat{\pi}_{ir})^2}{\pi_{ir}},$$

und *Mean Entropy* bzw. *Kullback-Leibler Loss*

$$\text{MEL} = \frac{1}{N} \sum_{i=1}^{N} \sum_{r=1}^{k} \pi_{ir} \log \left(\frac{\pi_{ir}}{\hat{\pi}_{ir}} \right).$$

Für die 17 fehlgeschlagenen Versuche zeigt Abbildung 6.2 die Verlustfunktionen der Modelle POM und NPOM im direkten Vergleich. Bei der Schätzung im Non-Proportional Odds Modell entspricht der Glättungsparameter dabei jeweils dem kleinsten Wert, für den das NPOM durch die penalisierte Version des Fisher–Scoring gefittet werden kann. Den Abbildungen ist zu entnehmen,

daß die penalisierten Schätzungen des NPOM in allen Fällen deutlich kleinere Werte der Verlustfunktionen MSEL und MRSEL aufweisen. Lediglich für zwei der 17 Fehlversuche erbringt die penalisierte Schätzung im NPOM keine Verbesserung des MEL gegenüber der Schätzung im POM.

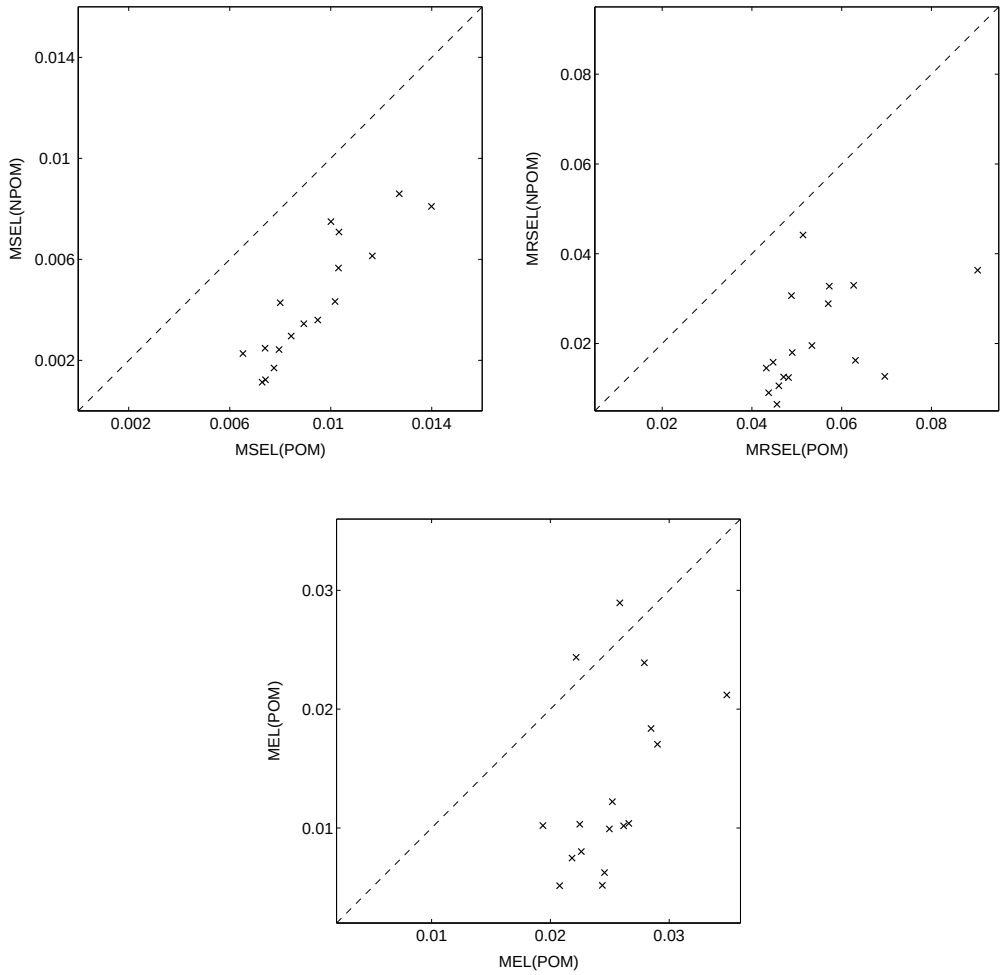

ABBILDUNG 6.2: *Verlustfunktionen des POM im Vergleich zu korrespondierenden Verlustfunktionen im penalisiert geschätzten NPOM für Ziehungen, in denen das Fisher–Scoring ohne Penalisierung fehlschlug. Die gestrichelten Linien repräsentieren die Situationen identischer Verluste.*

In Tabelle 6.2 sind die Ergebnisse für die 17 auftretenden Fehlversuche nochmals zusammengefaßt. Berechnet wurden die mittleren Verlustfunktionen im POM und im NPOM sowie die gemittelten Verhältnisse NPOM / POM. Die

mittleren Verluste des POM belaufen sich im Vergleich zu denen des penalisierten NPOM auf ungefähr das Doppelte. Mittelung der Verhältnisse liefert einen Faktor von mindestens 0.53, um den der Verlust reduziert werden kann, wenn anstelle des inadäquaten POM ein penalisiertes NPOM geschätzt wird.

	MSEL	MRSEL	MEL
POM	0.0093	0.0542	0.0251
NPOM	0.0043	0.0208	0.0135
NPOM/POM	0.4336	0.3801	0.5277

TABELLE 6.2: *Mittlere Verlustfunktionen und –verhältnisse, gemittelt über die Ziehungen, in denen Fisher–Scoring ohne Penalisierung fehlschlug.*

6.3 Penalisierte Test–Statistiken

Die mit der vorgestellten Penalisierungstechnik gegebene Möglichkeit, numerische Instabilitäten weitestgehend abfangen zu können, erlaubt eine Anwendung von Test-Statistiken, die die Verfügbarkeit von Schätzungen in der Modellklasse PPOM($\mathcal{G}$), $\mathcal{G} \neq \mathcal{P}$, explizit voraussetzen. Der Likelihood–Quotienten–Test auf das Vorliegen von globalen Kovariableneffekten erfordert mit der Schätzung im NPOM die Betrachtung mindestens eines Modells mit kategorienspezifischen Gewichten und ist in seiner klassischen Version daher häufig nicht anwendbar. Mit dem Konzept der penalisierten Schätzung kann jedoch ein alternativer LQ–Test formuliert werden, dessen Anwendbarkeit in nahezu allen Fällen gewährleistet ist.

Bezeichnen $pl_{\emptyset}\big(\hat{\gamma}_0, \{\hat{\gamma}_{\mathcal{P},r}\}_{r=1,\ldots,q}\big)$ bzw. $pl_{\mathcal{G}}\big(\tilde{\gamma}_0, \tilde{\gamma}_{\mathcal{G}}, \{\tilde{\gamma}_{\bar{\mathcal{G}},r}\}_{r=1,\ldots,q}\big)$ die maximalen penalisierten Log–Likelihoods des NPOM bzw. PPOM($\mathcal{G}$), so ist mit

$$LR_p \;=\; -2\big\{pl_{\mathcal{G}}\big(\tilde{\gamma}_0, \tilde{\gamma}_{\mathcal{G}}, \{\tilde{\gamma}_{\bar{\mathcal{G}},r}\}_{r=1,\ldots,q}\big) \;-\; pl_{\emptyset}\big(\hat{\gamma}_0, \{\hat{\gamma}_{\mathcal{P},r}\}_{r=1,\ldots,q}\big)\big\} \qquad (6.11)$$

die penalisierte Fassung des klassischen Likelihood–Quotienten zum Test auf proportionale Chancen für x_j, $j \in \mathcal{G} \subset \mathcal{P}$, definiert. Werden in $\hat{\boldsymbol{\beta}}$ und $\tilde{\boldsymbol{\beta}}$ die

Parameterschätzer des NPOM bzw. des PPOM($\mathcal{G}$) subsumiert, und bezeichnen $\widehat{\boldsymbol{\pi}}_i := (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{ik})'$ und $\widetilde{\boldsymbol{\pi}}_i := (\tilde{\pi}_{i1}, \ldots, \tilde{\pi}_{ik})'$ die zugehörigen geschätzten Responsewahrscheinlichkeiten, so kann (6.11) mit (6.9) und (5.25) als

$$LR_p \;=\; 2\sum_{i=1}^{N}\sum_{r=1}^{k} y_{ir}\log\left(\frac{\hat{\pi}_{ir}}{\tilde{\pi}_{ir}}\right) \;+\; P_{\tilde{\boldsymbol{\beta}}} \;-\; P_{\hat{\boldsymbol{\beta}}}$$

geschrieben werden. Für die kategorienspezifischen Gewichte kann die Stärke der erforderlichen Penalisierung in den beiden Modellen verschieden sein. Eine genauere Betrachtung der Differenz

$$\begin{aligned} P_{\tilde{\boldsymbol{\beta}}} \;-\; P_{\hat{\boldsymbol{\beta}}} \;&=\; \sum_{j\in\{0\}\cup\bar{\mathcal{G}}} \lambda_{j,\mathcal{G}} \sum_{r=1}^{q-1} (\Delta^1\tilde{\gamma}_{jr})^2 - \sum_{j=0}^{p} \lambda_{j,\emptyset} \sum_{r=1}^{q-1} (\Delta^1\hat{\gamma}_{jr})^2 \\ &=\; \sum_{j\in\{0\}\cup\bar{\mathcal{G}}} \sum_{r=1}^{q-1} \left(\lambda_{j,\mathcal{G}}(\Delta^1\tilde{\gamma}_{jr})^2 - \lambda_{j,\emptyset}(\Delta^1\hat{\gamma}_{jr})^2\right) \;-\; \sum_{j\in\mathcal{G}} \lambda_{j,\emptyset} \sum_{r=1}^{q-1} (\Delta^1\hat{\gamma}_{jr})^2 \end{aligned}$$

legt jedoch die Wahl identischer Glättungsparameter nahe. Falls $\lambda_{j,\mathcal{G}} = \lambda_{j,\emptyset}$, $j \in \{0\}\cup\bar{\mathcal{G}}$, wird der erste Ausdruck in obiger Darstellung vernachlässigbar klein, da $\tilde{\gamma}_{jr} \approx \hat{\gamma}_{jr}$, $r = 1, \ldots, q$. Der ausschlaggebende Anteil der Strafterme resultiert damit aus der Penalisierung der zu x_j, $j \in \mathcal{G}$, gehörigen Parameter, für die die Annahme der Gleichheit zur Diskussion steht.

Werden die kategorienspezifischen Parameter nicht penalisiert, gilt also

$$\lambda_{0,\mathcal{G}} \;=\; \lambda_{1,\mathcal{G}} \;=\; \ldots \;=\; \lambda_{|\bar{\mathcal{G}}|,\mathcal{G}} \;=\; \lambda_{0,\emptyset} \;=\; \lambda_{1,\emptyset} \;=\; \ldots \;=\; \lambda_{p,\emptyset} \;=\; 0,$$

reduziert sich LR_p auf den klassischen Likelihood–Quotienten, der asymptotisch χ^2–verteilt ist mit $|\mathcal{G}|\cdot(q-1)$ Freiheitsgraden (Fahrmeir & Tutz, 2001). Für penalisierte Schätzer existieren hingegen weder Aussagen hinsichtlich einer asymptotischen Grenzverteilung noch liegen gesicherte Erkenntnisse über deren Eigenschaften in endlichen Stichprobensituationen vor. Die Signifikanz eines beobachteten Wertes lr_p von LR_p wird daher via Monte Carlo Simulation bestimmt. Dazu werden Stichproben $\boldsymbol{y}_1^{(m)}, \ldots, \boldsymbol{y}_N^{(m)}$, $m = 1, \ldots, M$, unabhängiger, multinomialverteilter Zufallsvektoren $\boldsymbol{y}_i^{(m)} = \left(y_{i1}^{(m)}, \ldots, y_{iq}^{(m)}\right)'$ mit zugehörigen Auftretenswahrscheinlichkeiten $P(y_{ir}^{(m)} = 1) = \tilde{\pi}_{ir}$ generiert. Jede der so entstehenden M Datensituationen $\{\boldsymbol{y}_i^{(m)}, \boldsymbol{x}_i\}_{i=1,\ldots,N}$ wird als PPOM($\mathcal{G}$)

und als NPOM geschätzt. Derjenige Anteil der M korrespondierenden Werte von LR_p, die größer oder gleich lr_p sind, kann als Schätzung für den p-Wert herangezogen werden (Firth, Glosup & Hinkley, 1991). Diese Vorgehensweise motiviert sich direkt aus der Definition des p-Wertes als Wahrscheinlichkeit, unter der Nullhypothese (hier PPOM($\mathcal{G}$)) den beobachteten Prüfgrößenwert (hier lr_p), oder einen in Richtung der Alternative (hier NPOM) extremeren (hier größeren) Wert zu erhalten. Mit $\widetilde{\boldsymbol{\pi}}_i$, $i = 1, \ldots, N$, als geschätzten Responsewahrscheinlichkeiten im PPOM($\mathcal{G}$), gewährleistet die Ziehung von $\boldsymbol{y}_i^{(m)}$ aus der Multinomialverteilung $\mathcal{M}(1, \widetilde{\boldsymbol{\pi}}_i)$ dabei die erforderliche Kalkulation unter der Nullhypothese.

Eine weniger aufwendige Alternative zum LQ–Test stellt der Wald–Test dar. Für dessen Definition ist die zu testende Annahme proportionaler Chancen für x_j, $j \in \mathcal{G} := \{l_1, \ldots, l_{|\mathcal{G}|}\} \subseteq \mathcal{P}$, zunächst als lineare Hypothese zu formulieren. Setze dazu

$$C_s \;:=\; \left[\, 0_{(q-1)\times q} \;\mid\; -\mathbf{1}_{q-1} \otimes \boldsymbol{e}'_{l_s,p} \;\mid\; I_{q-1} \otimes \boldsymbol{e}'_{l_s,p} \,\right], \;\; s = 1, \ldots, |\mathcal{G}|,$$

Mit $\boldsymbol{\beta} := (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_{\mathcal{P},1}, \ldots, \boldsymbol{\gamma}_{\mathcal{P},q})'$ als Parametervektor im NPOM und der Deklaration $C := \left[\, C_1' \mid \ldots \mid C_{|\mathcal{G}|}' \,\right]'$ ist das PPOM($\mathcal{G}$) und damit die zu testende Annahme $\gamma_{j1} = \ldots = \gamma_{jq}$, $j \in \mathcal{G}$, äquivalent zu der linearen Hypothese $C\boldsymbol{\beta} = \mathbf{0}$.

Bezeichnet $\hat{\boldsymbol{\beta}}$ den Maximum–Likelihood–Schätzer von $\boldsymbol{\beta}$ im NPOM, so misst die Wald-Statistik die Distanz zwischen dem Schätzer $C\hat{\boldsymbol{\beta}}$ und dessen Erwartungswert $E\big(C\hat{\boldsymbol{\beta}} \,|\, \text{PPOM}(\mathcal{G})\big) = \mathbf{0}$ unter der Hypothese proportionaler Chancen für x_j, $j \in \mathcal{G}$,

$$W_p \;=\; \hat{\boldsymbol{\beta}}' C' \big(C\, \widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\, C'\big)^{-1} C\hat{\boldsymbol{\beta}}. \tag{6.12}$$

Die Berechnung der Wald–Statistik (6.12) erfordert im Gegensatz zum Likelihood–Quotienten lediglich eine Betrachtung des Non-Proportional Odds Modells. Der zugehörige Schätzer $\hat{\boldsymbol{\beta}}$ läßt sich bei numerischen Problemen wiederum nur aus der Maximierung der penalisierten Log–Likelihood gewinnen. In diesem Sinne stellt (6.12) ebenfalls eine penalisierte Statistik dar, wobei der Strafterm hier nur indirekt über die penalisierte Schätzung der unbekannten Kovariablengewichte einfliesst. Die Inverse der geschätzten Kovarianzmatrix $C\, \widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\, C'$ von $C\hat{\boldsymbol{\beta}}$ fungiert bei der angesprochenen Distanzmessung als zu-

sätzliches Gewicht. Zur Approximation von $\text{cov}(\hat{\boldsymbol{\beta}})$ wird der mit (5.23) gegebene Sandwich–Schätzer herangezogen.

Im Unterschied zum Wald–Test basiert die penalisierte Form des Score–Tests auf der Schätzung im PPOM($\mathcal{G}$). Die penalisierte Score–Funktion $ps_\emptyset(\boldsymbol{\beta})$ des NPOM ausgewertet am zugehörigen Maximum-Likelihood-Schätzer $\hat{\boldsymbol{\beta}}$ liefert einen Nullvektor der Länge $q\cdot(p+1)$: $ps_\emptyset(\hat{\boldsymbol{\beta}}) = \mathbf{0}_{q(p+1)}$. Ersetzt man $\hat{\boldsymbol{\beta}}$ durch den Parameterschätzer $\tilde{\boldsymbol{\beta}}$ des PPOM($\mathcal{G}$), so ist $ps_\emptyset(\tilde{\boldsymbol{\beta}})$ signifikant verschieden von null, sofern die Annahme proportionaler Chancen für x_j, $j \in \mathcal{G}$, nicht gerechtfertigt ist. Die penalisierte Score–Statistik

$$s_p \;=\; ps_\emptyset(\tilde{\boldsymbol{\beta}})' F_\emptyset(\tilde{\boldsymbol{\beta}})^{-1} ps_\emptyset(\tilde{\boldsymbol{\beta}}) \tag{6.13}$$

misst den gewichteten Abstand zwischen $ps_\emptyset(\tilde{\boldsymbol{\beta}})$ und dem Nullvektor $ps_\emptyset(\hat{\boldsymbol{\beta}})$. Als Gewichtungsmatrix fungiert dabei die Inverse der Kovarianzmatrix

$$\text{cov}\big(ps_\emptyset(\boldsymbol{\beta})\big) \;=\; \text{cov}\big(s_\emptyset(\boldsymbol{\beta})\big) \;=\; F_\emptyset(\boldsymbol{\beta})$$

ausgewertet an der Stelle $\tilde{\boldsymbol{\beta}}$. Da $\hat{\boldsymbol{\beta}}$ und $\tilde{\boldsymbol{\beta}}$ unterschiedlich dimensioniert sind, ist für die Berechnung von $ps_\emptyset(\tilde{\boldsymbol{\beta}})$ bzw. $F_\emptyset(\tilde{\boldsymbol{\beta}})$ die Länge von $\tilde{\boldsymbol{\beta}}$ entsprechend anzupassen. Dies geschieht via Einfügung von $\tilde{\boldsymbol{\gamma}}_\mathcal{G}$ in den Vektor $\tilde{\boldsymbol{\beta}}$ vor jedem $\tilde{\boldsymbol{\gamma}}_{\bar{\mathcal{G}},r}$, $r = 2,\ldots,q$,

$$\tilde{\boldsymbol{\beta}} \;=\; (\tilde{\boldsymbol{\gamma}}_0', \tilde{\boldsymbol{\gamma}}_\mathcal{G}', \tilde{\boldsymbol{\gamma}}_{\bar{\mathcal{G}},1}', \tilde{\boldsymbol{\gamma}}_{\bar{\mathcal{G}},2}', \ldots, \tilde{\boldsymbol{\gamma}}_{\bar{\mathcal{G}},q}')' \;\;\rightsquigarrow\;\; (\tilde{\boldsymbol{\gamma}}_0', \tilde{\boldsymbol{\gamma}}_\mathcal{G}', \tilde{\boldsymbol{\gamma}}_{\bar{\mathcal{G}},1}', \tilde{\boldsymbol{\gamma}}_\mathcal{G}', \tilde{\boldsymbol{\gamma}}_{\bar{\mathcal{G}},2}', \ldots, \tilde{\boldsymbol{\gamma}}_\mathcal{G}', \tilde{\boldsymbol{\gamma}}_{\bar{\mathcal{G}},q}')'.$$

Aufgrund der speziellen Gestalt des Differenzenpenalties ist $K_\emptyset\tilde{\boldsymbol{\beta}}$ – und damit die Score–Statistik s_p – für den so modifizierten Vektor unabhängig von den in $K_\emptyset$ spezifizierten Glättungsparametern $\lambda_{j,\emptyset}$, $j \in \mathcal{G}$. Eine Klassifizierung als penalisierte Test–Statistik ist somit nur dann zutreffend, wenn die Schätzung der zu x_j, $j \in \bar{\mathcal{G}}$, gehörigen Parameter im PPOM($\mathcal{G}$) einer Penalisierung bedarf. Ist dies nicht der Fall, so reduziert sich s_p auf die klassische Form

$$s_\emptyset(\tilde{\boldsymbol{\beta}})' F_\emptyset(\tilde{\boldsymbol{\beta}})^{-1} s_\emptyset(\tilde{\boldsymbol{\beta}})$$

der Score–Statistik. Gleiches gilt in der Situation $\mathcal{G} = \mathcal{P}$, d.h. wenn die Annahme ausschließlich globaler Effekte (POM) getestet werden soll.

Die nur partiell existente Hierarchie in der Menge $\{\text{PPOM}(\mathcal{G}) : \mathcal{G} \subseteq \mathcal{P}\}$ gestaltet die Suche nach einem adäquaten Modell zu einem recht aufwendigen

Testproblem. Statt alle Hypothesen $H_{\mathcal{G}} : \gamma_{j1} = \ldots = \gamma_{jq}$, $j \in \mathcal{G}$, innerhalb eines multiplen Testproblems zu behandeln, wird im folgenden eine stark vereinfachte Strategie betrachtet: Getestet werden alle Hypothesen $H_{\{j\}}$, $j \in \mathcal{P}$, mit dem rein kategorienspezifischen Modell als Alternative, d.h. die Modelle PPOM($\{j\}$), $j \in \mathcal{P}$, werden auf Basis der eingeführten Test–Statistiken mit dem NPOM = PPOM($\emptyset$) verglichen.

6.3.1 Simulation: Gütefunktionen

Eine qualitative Bewertung der vorgestellten Test–Statistiken kann in effektiver Weise anhand der korrespondierenden Gütefunktionen erfolgen. Für eine Simulationsstudie wird die folgende Prädiktorspezifikation zugrunde gelegt

$$\begin{aligned} \eta_{i1} &= -0.8 + (0.4 + \Delta) \cdot x_i \\ \eta_{i2} &= -0.4 + 0.4 \cdot x_i \\ \eta_{i3} &= 0.2 + (0.4 - \Delta) \cdot x_i \,, \end{aligned}$$

wobei die Daten x_i, $i = 1, \ldots, N$, Realisationen einer auf dem Intervall $[-1, 1]$ gleichverteilten Zufallsgröße sind. In Abhängigkeit vom Parameter $\Delta \geq 0$ ist das so spezifizierte Modell ein POM ($\Delta = 0$) oder ein NPOM ($\Delta > 0$). Wachsendes Δ erhöht die Nicht–Proportionalität der Chancen. In diesem Sinne definiert Δ ein „Abstandsmaß“ zwischen dem POM und dem NPOM.

Die Gütefunktionen der drei Test-Statistiken resultieren aus den geschätzten Wahrscheinlichkeiten, die Nullhypothese proportionaler Chancen (POM) für wachsendes Δ zu verwerfen. Beginnend bei $\Delta = 0$ (POM) wurden 160 Stichproben von je $N = 250$ unabhängigen Zufallsgrößen $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ entsprechend obiger Prädiktorspezifikation gezogen, d.h. $\boldsymbol{y}_i \sim \mathcal{M}\big(1, \boldsymbol{\pi}_i = h(\boldsymbol{\eta}_i)\big)$. Für jede der Stichproben wurden die Modelle POM und NPOM gefittet und auf deren Basis die Statistiken LR_p, W_p und s_p berechnet. Die Bestimmung der Signifikanz dieser drei Größen erfolgte via Monte Carlo Simulation mit je $M = 200$ Ziehungen. Anhand der so erhaltenen 160 p-Werte wurde die Wahrscheinlichkeit, die Nullhypothese proportionaler Chancen zu verwerfen, über die relative Häufigkeit in den Stichproben abgeschätzt. Diese Prozedur wurde dann für schrittweise wachsendes Δ bis zum maximalen Wert $\Delta = 0.4$ wiederholt.

Abbildung 6.3 zeigt, daß das spezifizierte Modell für jede Kombination von $x_i \in [-1, 1]$ und $\Delta \in [0, 0.4]$ der Ordnungsrestriktion $\eta_{i1} \leq \eta_{i2} \leq \eta_{i3}$ genügt.

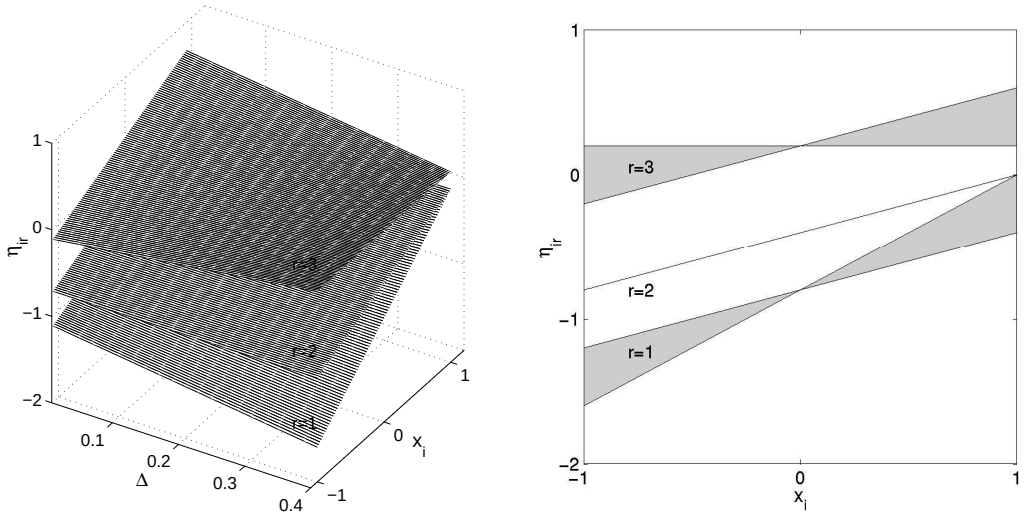

ABBILDUNG 6.3: *Links: Prädiktorwerte in Abhängigkeit von Einflußgröße x_i und Parameter Δ. Rechts: Projektionen in die x_i–η_{ir}–Ebene.*

In Abbildung 6.4 sind die geschätzten Gütefunktionen von LR_p, W_p und s_p für eine Penalisierung mit Glättungsparameter $\lambda_{1,\emptyset} = \exp(4)$ dargestellt. Die für wachsendes $\Delta \in [0, 0.4]$ generierten punktweisen Schätzungen wurden lokal durch ein Polynom zweiten Grades approximiert. Der jeweils resultierende lokal quadratische Fit ist ebenfalls geplottet. Die vertikalen Linien in den oberen beiden Darstellungen repräsentieren für jede Wahl von Δ den Anteil der fehlgeschlagenen Versuche, das korrespondierende Modell mit der angegebenen Penalisierung zu schätzen. Dieser nimmt erwartungsgemäß mit wachsender Nicht–Proportionalität der Chancen zu. Im Gegensatz zum Anteil der Fehlversuche beim Schätzen ohne Penalisierung – dargestellt im linken unteren Plot – ist dieser Zuwachs jedoch deutlich reduziert.

In der Abbildung rechts unten werden die geschätzten Gütefunktionen überlagert. Die Graphen der lokal quadratischen Fits sind nahezu identisch, ein signifikanter Unterschied zwischen LR_p, W_p und s_p ist nicht ersichtlich. Wie bereits erwähnt, stimmen für die hier vorliegende Testsituation (POM versus NPOM) penalisierte und klassische Version der Score–Statistik überein, die

Ergebnisse erlauben daher den Schluß, daß die penalisierten Testformen und der gewöhnliche, unpenalisierte Score–Test von vergleichbarer Qualität sind. Weitere, hier nicht präsentierte Simulationen zeigten, daß diese Aussage für jede Wahl des Glättungsparameters $\lambda_{1,\emptyset}$ verallgemeinerbar ist.

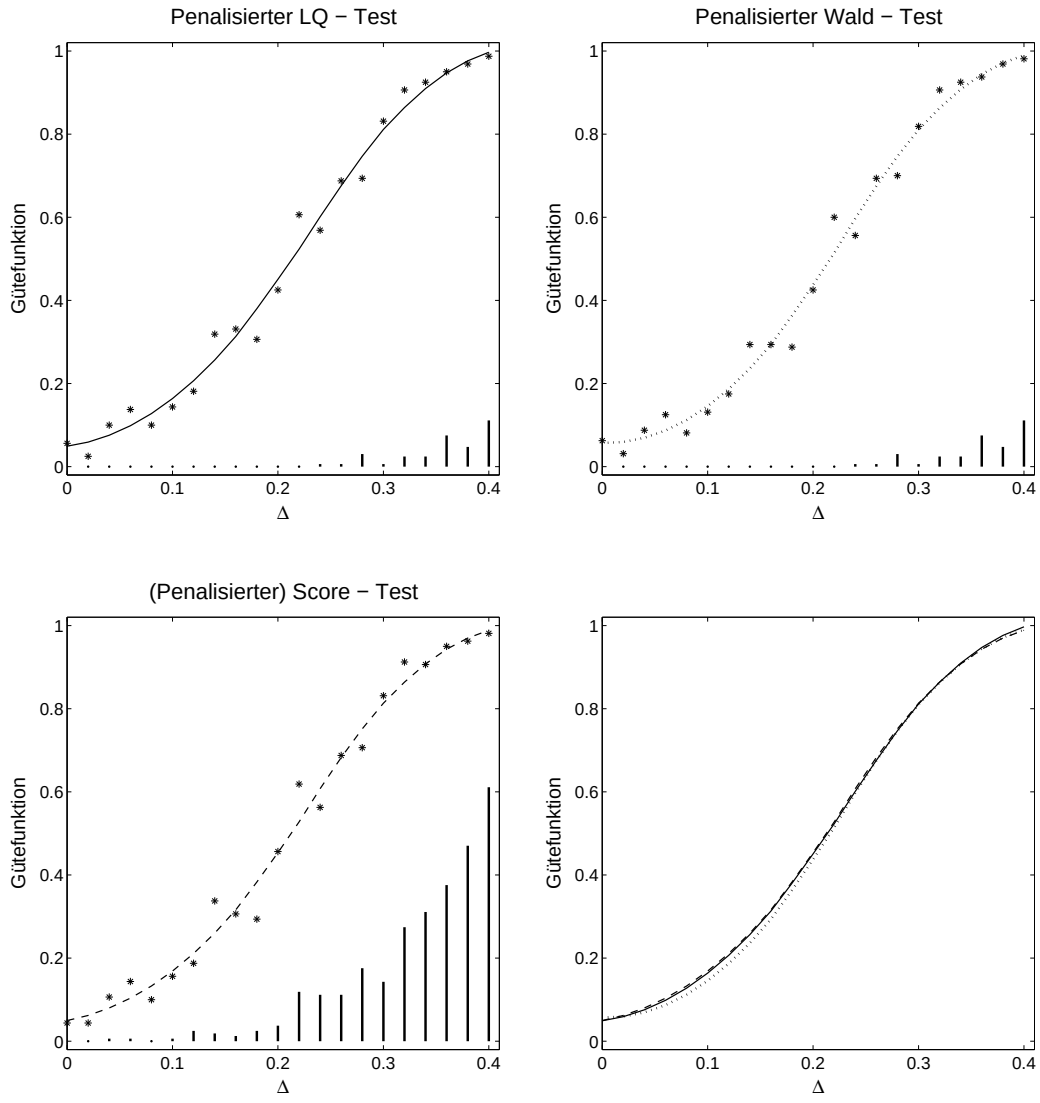

ABBILDUNG 6.4: *Geschätzte Gütefunktionen: LR_p (links oben), W_p (rechts oben) und s_p (links unten) für Penalisierung mit $\lambda_{1,\emptyset} = \exp(4)$. Vertikale Linien indizieren den Anteil der Fehlversuche (links bzw. rechts oben: mit Penalisierung, links unten: ohne Penalisierung ($\lambda_{1,\emptyset} = 0$)). Rechts unten: Überlagerte Gütefunktionen.*

6.4 Beispiel: Diabetische Retinopathie

Durch Diabetes mellitus verursachte, krankhafte Veränderungen der menschlichen Netzhaut bezeichnet man in der Medizin als diabetische Retinopathie. Gegenstand der folgenden Untersuchungen sind Daten aus einer sechsjährigen Verlaufsstudie zum Effekt von Risikofaktoren auf die Entstehung diabetischer Retinopathie (Jörgens et al., 1993, Mühlhauser et al., 1996). Die betrachteten Risikofaktoren sind das dichotome Merkmal Raucherstatus ($SM = 0$: Nichtraucher, $SM = 1$: Raucher) und die metrischen Größen Diabetesdauer (DD) in Jahren, glykosyliertes Hämoglobin (GH) in Prozent sowie der diastolische Blutdruck (BP) in Millimeter Quecksilbersäule (mmHg). Der zum Ende der Studie vorliegende Schweregrad bzw. Status der Retinopathie wird auf einer ordinalen Skala gemessen und definiert den zu modellierenden Response. Als dessen Ausprägungen werden die Kategorien keine Retinopathie (1), nonproliferative Retinopathie (2) und fortgeschrittene Retinopathie bzw. Blindheit (3) unterschieden.

Bender & Grouven (1998) analysieren dieses Datenbeispiel ebenfalls und weisen darauf hin, daß ein adäquates Modell die Diabetesdauer sowohl linear als auch quadriert (DDQ) berücksichtigen sollte. Aus diesem Grund betrachten wir zunächst das NPOM

$$\eta_{ir} = \gamma_{0r} + SM_i \cdot \gamma_{SM,r} + DD_i \cdot \gamma_{DD,r} + DDQ_i \cdot \gamma_{DDQ,r} + GH_i \cdot \gamma_{GH,r} + BP_i \cdot \gamma_{BP,r} \,,$$

für $i = 1, \ldots, N = 613$ Diabetespatienten und Kategorien $r = 1, 2$. Der Versuch, obiges Modell zu fitten, scheitert jedoch an numerischen Schwierigkeiten. Ohne Penalisierung bricht das Fisher–Scoring bereits nach wenigen Iterationen fehlerbedingt ab - Abbruchursache ist die schlechte Konditionierung der Fisher–Matrix. Selbst eine Skalierung sämtlicher Kovariablen auf das Intervall $[0, 1]$ kann dies nicht verhindern. Eine Konvergenz des Schätzalgorithmus kann aber durch einen kleinen Penalty $\lambda^{\min}_{DD,\emptyset} = \lambda^{\min}_{DDQ,\emptyset} = 1$ auf den Parametern $\gamma_{DD,r}$ und $\gamma_{DDQ,r}$, $r = 1, 2$, erzwungen werden. Für die übrigen Parameter ist keine zusätzliche Penalisierung erforderlich.

Für jedes Submodell PPOM($\{j\}$), $j \in \{SM, GH, BP\}$, treten die genannten numerischen Probleme ebenfalls auf. Einzige Ausnahme ist das Partial Proportional Odds Modell mit globalen Effekten für DD und DDQ, für das die

unpenalisierte Schätzung reibungslos abläuft. Linearer und quadratischer Effekt der Diabetesdauer werden dabei nicht separat, sondern ausschließlich simultan betrachtet. Ansätze, in denen einer der Effekte global, der andere hingegen kategorienspezifisch modelliert wird, erscheinen wenig sinnvoll. Für die problematischen Modelle PPOM($\{j\}$), $j \in \{SM, GH, BP\}$, kann die Verfügbarkeit von Schätzungen ebenfalls über Penalties $\lambda^{\min}_{DD,\{j\}} = \lambda^{\min}_{DDQ,\{j\}} = 1$ gesichert werden.

Da unpenalisierte Schätzungen im NPOM nicht zur Verfügung stehen, scheiden die klassischen Versionen des LQ– und des Wald–Tests für einen Test auf global vorliegende proportionale Chancen (POM vs. NPOM) aus. Der Score-Test als einzig verbleibende Alternative liefert für diese Situation einen Wert von 16.69, der – verglichen mit dem 95%–Quantil der $\chi^2(5)$–Verteilung – die Ablehnung der Nullhypothese proportionaler Chancen indiziert. Da das klassische Fisher–Scoring im angemesseneren NPOM jedoch fehlschlägt, ist man bei der Interpretation von Effekten auf die Ergebnisse des inadäquaten POM beschränkt. Die Untersuchung zur Signifikanz von SM im Proportional Odds Modell liefert die Werte 1.69 (Score–Test), 1.72 (Wald–Test) und 1.71 (LQ–Test). Der Vergleich mit dem 95%–Quantil der χ^2–Verteilung mit einem Freiheitsgrad impliziert einen nicht-signifikanten Einfluss des Raucherstatus auf die Entwicklung von Retinopathie. Bender & Grouven (1998), die den dichotomisierten Response analysieren, stellen jedoch fest, daß diese Nicht-Signifikanz ein aus der Unterstellung falscher Modellannahmen resultierendes Artefakt ist. Diese Aussage soll im folgenden unter Verwendung von penalisierten Test–Statistiken manifestiert werden.

Die Betrachtung der vier Hypothesenpaare

$$\text{PPOM}(\{j\}) \quad \text{vs.} \quad \text{NPOM}\,, \quad j \in \mathcal{P} = \big\{SM, \{DD, DDQ\}, GH, BP\big\}$$

soll zunächst Aufschluß geben, für welche Kovariablen die Annahme globaler Parameter ungerechtfertigt ist. Dazu werden die entsprechenden Modelle penalisiert gefittet und auf Basis der Fits die Test–Statistiken LR_p, W_p und s_p berechnet. Die Stärke der Penalisierung orientiert sich für $l \in \bar{\mathcal{G}} = \mathcal{P}\backslash\{j\}$ dabei am minimal erforderlichen Betrag, d.h. $\lambda_{l,\{j\}} = \lambda^{\min}_{l,\{j\}}$ und $\lambda_{l,\emptyset} = \lambda^{\min}_{l,\emptyset}$, mit $\lambda^{\min}_{l,\{j\}} = \lambda^{\min}_{l,\emptyset} = 1$, $l \in \{DD, DDQ\}$, und $\lambda^{\min}_{l,\{j\}} = \lambda^{\min}_{l,\emptyset} = 0$, $l \notin \{DD, DDQ\}$.

Für die untersuchte Variable j selbst wird deren Glättungsparameter $\lambda_{j,\emptyset}$ im Intervall $[\lambda_{j,\emptyset}^{\min}, \exp(14)]$ variiert, um eventuelle Abhängigkeiten der Test-Statistiken von der Penalisierungsstärke aufdecken zu können. Linearer und quadratischer Effekt der Diabetesdauer werden dabei simultan betrachtet.

Die Abbildungen 6.5 und 6.6 zeigen für jedes der möglichen Hypothesenpaare die via Monte Carlo Simulation bestimmten p–Werte der Test–Statistiken s_p, W_p und LR_p in Abhängigkeit vom jeweiligen Glättungsparameter $\lambda_{j,\emptyset}$. Diese auch als *significance traces* bekannten Darstellungen werden schon von Azzalini & Bowman (1993) (vgl. auch Bowman & Azzalini (1997)) als nützliches Analysewerkzeug beschrieben.

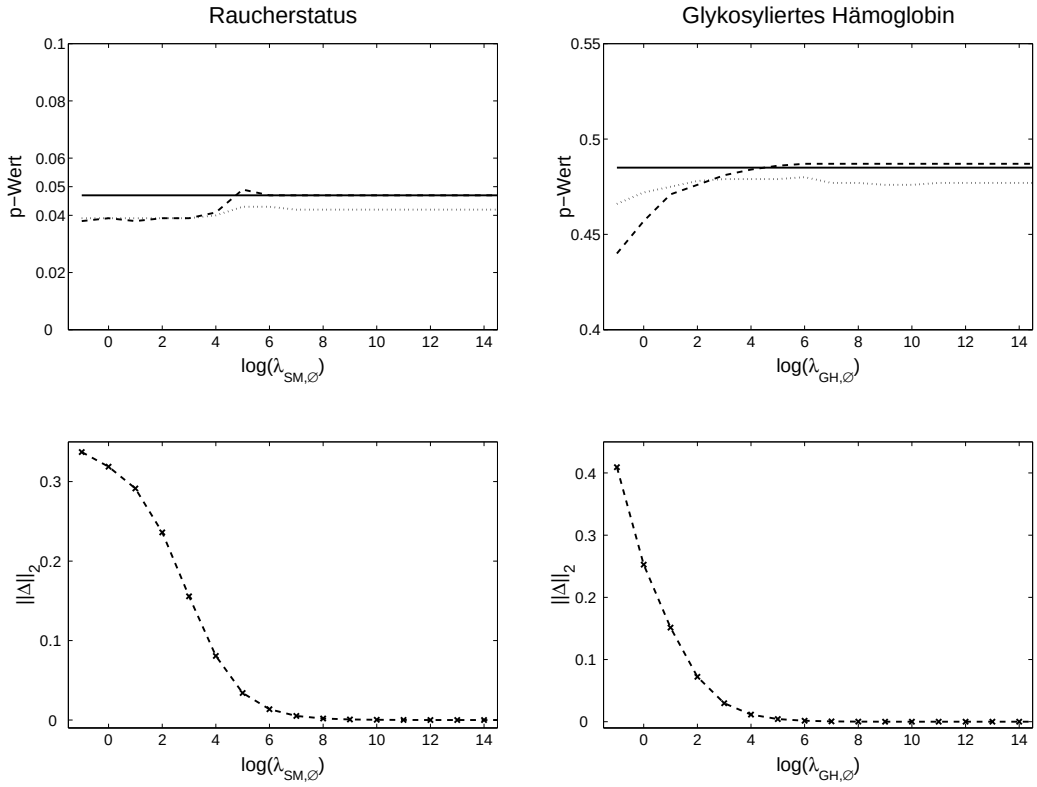

ABBILDUNG 6.5: *Oben: Geschätzte p-Werte für Tests PPOM($\{SM\}$) (links) und PPOM($\{GH\}$) (rechts) vs. NPOM der Test–Statistiken s_p (durchgezogene Linie), W_p (gestrichelte Linie) und LR_p (gepunktete Linie) bei verschiedenen Penalisierungsstärken. Unten: Abweichungen der penalisierten kategorienspezifischen Parameter von korrespondierender Annahme globaler Effekte für SM (rechts) und GH (links).*

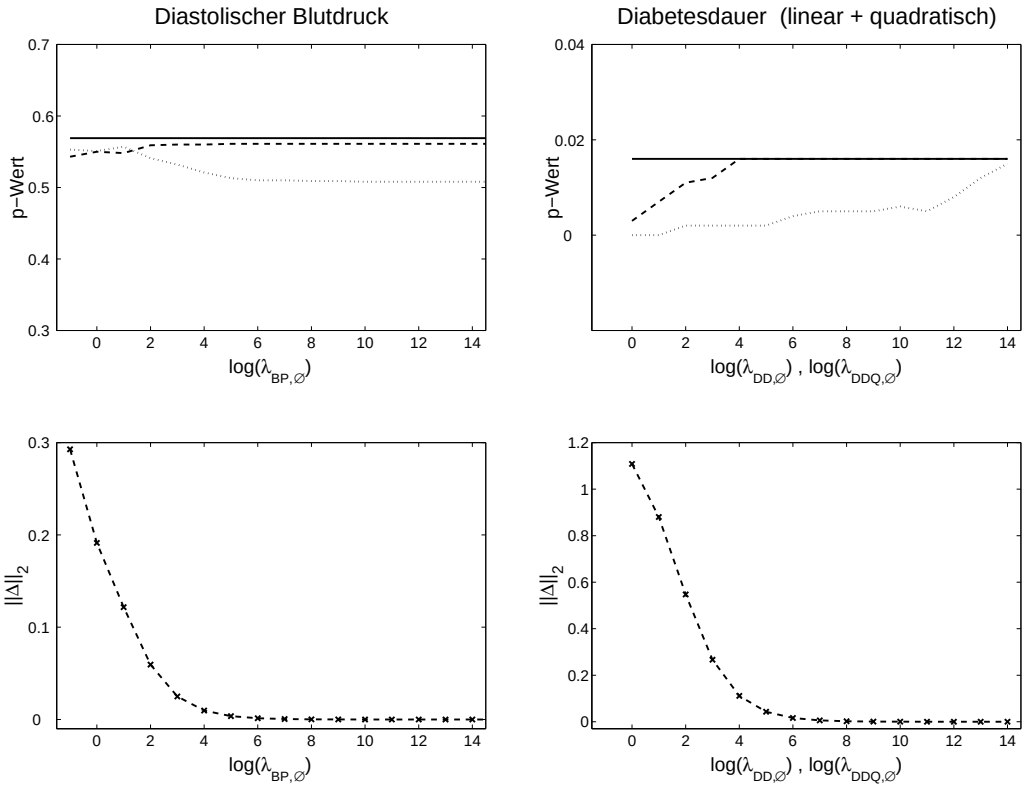

ABBILDUNG 6.6: *Oben: Geschätzte p-Werte für Tests PPOM($\{BP\}$) (links) und PPOM($\{DD, DDQ\}$) (rechts) vs. NPOM der Test-Statistiken s_p (durchgezogene Linie), W_p (gestrichelte Linie) und LR_p (gepunktete Linie) bei verschiedenen Stärken des Penalties. Unten: Abweichungen der penalisierten kategorienspezifischen Gewichte von korrespondierender Annahme globaler Effekte für BP (rechts) und DD, DDQ (links).*

Die für alle Hypothesenpaare horizontal verlaufenden significance traces der Score–Statistik s_p spiegeln deren bereits erwähnte Unabhängigkeit vom Glättungsparameter $\lambda_{j,\emptyset}$ wieder. Auch für die penalisierten Versionen von LQ- und Wald–Test deuten die relativ stabilen Verläufe der significance traces auf eine weitestgehende Unabhängigkeit des zugehörigen p-Wertes und der damit verbundenen Testentscheidung von der gewählten Penalisierungsstärke hin. Mit p-Werten um 0.5 implizieren obige Darstellungen proportionale Chancen für glykosyliertes Hämoglobin und den diastolischen Blutdruck. Die um 0.05 und 0.01 schwankenden p-Wertschätzungen für den Raucherstatus bzw. die Diabetesdauer (linear und quadratisch) liefern ein klares Indiz, daß diese Effekte kategorienspezifisch sind.

In den Abbildungen 6.5 und 6.6 ist neben den geschätzten significance traces die Abweichung der Schätzungen für $\gamma_{j,1}$ und $\gamma_{j,2}$ im NPOM von der korrespondierenden Annahme globaler Effekte in Abhängigkeit von $\lambda_{j,\emptyset}$ dargestellt. Die Projektion des Vektors $(\hat{\gamma}_{j,1}, \hat{\gamma}_{j,2})'$ auf die durch $\gamma_{j,1} = \gamma_{j,2}$ definierte Hyperebene liefert dabei

$$||\Delta||_2 := \sqrt{q^{-1} \cdot \sum_{s=1}^{q-1} \sum_{t=s+1}^{q} (\hat{\gamma}_{j,s} - \hat{\gamma}_{j,t})^2} = \sqrt{0.5} \cdot |\hat{\gamma}_{j,1} - \hat{\gamma}_{j,2}|$$

als geeignetes Abweichungsmaß. Aus den unteren Darstellungen geht hervor, daß mit dem für jedes $\lambda_{j,\emptyset}$ betrachteten Intervall $[\lambda_{j,\emptyset}^{\min}, \exp(14)]$ der relevante Bereich abgedeckt ist. In allen Testsituationen wird für $\log(\lambda_{j,\emptyset}) > 6$ das Modell mit globalen Gewichten für die Kovariable(n) j geschätzt, so daß der Bereich, in dem die penalisierte Schätzung im NPOM noch zu kategorienspezifischen Effekten für j führt, jeweils durch das Intervall $[\lambda_{j,\emptyset}^{\min}, \exp(6)]$ bestimmt ist. Auch wenn die zu LR_p und W_p gehörigen significance traces in diesen Intervallen eine gewisse Variation aufweisen, so bleiben die Testentscheidungen bezüglich globaler oder kategorienspezifischer Modellierung von j davon letztlich unberührt.

Basierend auf den significance traces kann die Prädiktorspezifikation

$$\eta_{ir} = \gamma_{0r} + GH_i \cdot \gamma_{GH} + BP_i \cdot \gamma_{BP} + SM_i \cdot \gamma_{SM,r} + DD_i \cdot \gamma_{DD,r} + DDQ_i \cdot \gamma_{DDQ,r},$$

mit globalen Gewichten für glykosyliertes Hämoglobin und den diastolischen Blutdruck als hinreichend komplex angesehen werden. In Tabelle 6.3 sind die geschätzten Kovariableneffekte und Standardabweichungen für obiges Modell aufgelistet. Die kategorienspezifischen Parameterschätzungen der Einflußgröße Diabetesdauer (linear und quadratisch) wurden dabei penalisiert mit Glättungsparametern

$$\lambda_{DD,\{GH,BP\}} = \lambda_{DDQ,\{GH,BP\}} = \lambda_{DD,\{GH,BP\}}^{\min} = \lambda_{DDQ,\{GH,BP\}}^{\min} = 1.$$

Ohne diese Penalisierungen scheitert das Schätzen des PPOM($\{GH, BP\}$) an den bekannten numerischen Schwierigkeiten. Zur Schätzung der Standardabweichungen wurde die Approximationsformel (5.23) benutzt.

Tests zur Signifikanz der Kovariableneffekte können unter Verwendung geeignet adaptierter Versionen der penalisierten Test–Statistiken aus Abschnitt 6.3

	Effekt	Standard-abweichung	p-Werte LR_p	W_p	s_p
γ_{01}	6.031	0.523			
γ_{02}	7.642	0.587			
γ_{GH}	-3.685	0.605	0.000	0.000	0.000
γ_{BP}	-2.932	0.627	0.000	0.000	0.000
$\gamma_{SM,1}$	-0.409	0.207	0.048	0.049	0.049
$\gamma_{SM,2}$	0.059	0.244	0.810	0.809	0.817
$\gamma_{DD,1}$	-11.263	1.656	0.000	0.000	0.000
$\gamma_{DD,2}$	-11.880	1.701	0.000	0.000	0.000
$\gamma_{DDQ,1}$	8.265	1.837	0.000	0.000	0.000
$\gamma_{DDQ,2}$	7.319	1.822	0.000	0.000	0.001

TABELLE 6.3: *Geschätzte Kovariableneffekte und Standardabweichungen im Modell PPOM($\{GH, BP\}$). Spalten rechts: p–Werte der penalisierten Test–Statistiken LR_p, W_p und s_p für Tests auf Signifikanz der Effekte.*

durchgeführt werden. Als wesentlichste Modifikation sind Null– und Alternativhypothese neu zu formulieren. Im vollen Modell werden alle zur Verfügung stehenden Kovariablen berücksichtigt. Ob deren Parameter global oder kategorienspezifisch sind, ist in einem vorgeschalteten Schritt zu testen. Als Submodelle kommen all jene Konstellationen in Frage, die aus dem Nichtberücksichtigen einzelner Kovariableneffekte resultieren. Die Präsenz kategorienspezifischer Gewichte im Null- und/oder Alternativ–Modell erfordert unter Umständen eine penalisierte Schätzung. Geeignete Adaptionen der definierenden Größen erlauben dann die Anwendung der Test–Statistiken LR_p, W_p und s_p.

Tabelle 6.3 zeigt die Ergebnisse für die Retinopathie–Daten. Ausgehend vom PPOM($\{GH, BP\}$) als vollem Modell wurden alle möglichen Submodelle gefittet und adaptierte Versionen von LR_p, W_p sowie s_p berechnet. Die Kalkulation der zugehörigen p–Werte erfolgte auf der Basis von Monte Carlo Simulationen mit $M = 5000$ Ziehungen. Der Raucherstatus zeigt für alle Tests einen

signifikant negativen Effekt in der ersten Kategorie, woraus sich schlußfolgern läßt, daß Raucher ein höheres Risiko tragen, mindestens nonproliferative Retinopathie auszubilden. Auf eine Entwicklung fortgeschrittener Retinopathie hat der Raucherstatus hingegen keinen signifikanten Einfluß. Für die übrigen Variablen weisen die p–Werte auf hoch signifikante (globale bzw. kategorienspezifische) Effekte hin.

Obige Ergebnisse bestätigen damit die von Bender & Grouven (1998) getroffenen Aussagen: Der nicht–signifikante Effekt des Raucherstatus im POM ist ein aus der fälschlichen Annahme proportionaler Chancen resultierendes Artefakt. Wird der Raucherstatus (in korrekter Weise) kategorienspezifisch modelliert, sind dessen Effekte in zumindest einer Kategorie signifikant.

6.5 Restringierte Modelle

Die Identifikation globaler Effekte ist insbesondere im Hinblick auf eine parameterökonomische Modellierung erstrebenswert. Proportionale Chancen stellen jedoch in vielen Situationen eine zu rigide Einschränkung dar. In der Literatur finden sich zahlreiche alternative Wege, Kovariableneffekte im Sinne einer Modellvereinfachung zu restringieren, ohne deren kategorienspezifischen Charakter dabei völlig aufgeben zu müssen. Peterson & Harrell (1990) analysieren kumulative Logit–Modelle mit der Prädiktorspezifikation

$$\eta_{ir} \;=\; \gamma_{0r} \;+\; \boldsymbol{x}_i'\boldsymbol{\gamma} \;+\; \sum_{j\in\bar{\mathcal{G}}} x_{ij}\tilde{\gamma}_j\delta_{jr}\,, \tag{6.14}$$

für vordefinierte Skalare δ_{jr}, $j \in \bar{\mathcal{G}}$, $r = 1,\ldots,q$. Unter Berücksichtigung der zusätzlichen Nebenbedingungen $\gamma_{jr} = \gamma_j + \tilde{\gamma}_j \cdot \delta_{jr}$, $j \in \bar{\mathcal{G}}$, läßt sich (6.14) in bekannter Form auch als Partial Proportional Odds Modell

$$\eta_{ir} \;=\; \gamma_{0r} \;+\; \boldsymbol{x}'_{i,\mathcal{G}}\boldsymbol{\gamma}_{\mathcal{G}} \;+\; \boldsymbol{x}'_{i,\bar{\mathcal{G}}}\boldsymbol{\gamma}_{\bar{\mathcal{G}},r}$$

formulieren. Mit der Restringierung der kategorienspezifischen Parameter γ_{jr} reduzieren sich die für jede Kovariable x_j, $j \in \bar{\mathcal{G}}$, zu schätzenden Effekte auf die beiden unbekannten Gewichte γ_j und $\tilde{\gamma}_j$. Die daraus resultierende Modellvereinfachung wird jedoch auf Kosten einer häufig eher willkürlichen Annahme der Konstanten δ_{jr}, $r = 1,\ldots,q$, erzielt.

Im Kontext penalisierter Schätzungen läßt sich jedwede Willkür bei der Wahl der Skalare δ_{jr} durch eine rein datenadaptive Form der Festlegung umgehen. Vereinfachend werde dazu das NPOM mit der Prädiktorspezifikation

$$\eta_{ir} \;=\; \gamma_{0r} + \sum_{j=1}^{p} x_{ij}\gamma_{jr} \;=\; (\boldsymbol{e}'_{r,q}, \boldsymbol{e}'_{r,q} \otimes \boldsymbol{x}'_i)\,\boldsymbol{\beta},\;\; \boldsymbol{\beta} = (\boldsymbol{\gamma}'_0, \boldsymbol{\gamma}'_{\mathcal{P},1}, \ldots, \boldsymbol{\gamma}'_{\mathcal{P},q})'$$

betrachtet. Statt den Zusammenhang $\gamma_{jr} = \gamma_j + \tilde{\gamma}_j \cdot \delta_{jr}$ mit fixen Konstanten δ_{jr} anzunehmen, wird eine Reduktion der Parameterzahl durch stark penalisiertes Schätzen mit verschiedenen Differenzenordnungen erreicht. Der schon für P–Splines abgeleitete Zusammenhang zwischen starker Penalisierung und polynomialen Schätzungen bildet hierbei das zugrunde liegende Konzept. Erfolgt die Schätzung unter Verwendung eines Differenzenpenalties d–ter Ordnung

$$P_{\boldsymbol{\beta}} := \sum_{j=1}^{p} \lambda_{j,\emptyset} \sum_{r=1}^{q-d} (\Delta^d \gamma_{jr})^2,\;\; 1 \le d \le q-1\,, \tag{6.15}$$

so sind die geschätzten Parameter $\hat{\gamma}_{jr}$, $r = 1, \ldots, q$, für $\lambda_{j,\emptyset} \longrightarrow \infty$ und festes $j \in \mathcal{P} = \{1, \ldots, p\}$ durch ein Polynom vom Grad $d-1$ in r bestimmt. Wählt man beispielsweise $d = 2$, liegen die zu x_j gehörigen Gewichte auf einer Geraden, d.h. $\hat{\gamma}_{jr} = \alpha_{j0} + \alpha_j r$. Die effektive Anzahl gefitteter Parameter reduziert sich damit auf zwei: Intercept α_{j0} und Anstieg α_j. Im Ansatz von Peterson & Harrell (1990) entspricht dies der Wahl von $\delta_{jr} = r$, $r = 1, \ldots, q$.

Zum Nachweis der Beziehung zwischen starker Penalisierung und polynomialer Struktur in den entsprechenden Parameterschätzungen wird das eindeutig bestimmte Polynom $\tilde{p} \in \mathbb{P}_{q-1}$ betrachtet, das an den Stützstellen $1, \ldots, q$ die Stützwerte $\hat{\gamma}_{j1}, \ldots, \hat{\gamma}_{jq}$ annimmt: $\tilde{p}(r) = \hat{\gamma}_{jr}$, $r = 1, \ldots, q$. Laut dem Ansatz von Newton hat das Interpolationspolynom $\tilde{p}$ die Gestalt

$$\tilde{p}(x) \;=\; \tilde{\alpha}_{j0} \;+\; \tilde{\alpha}_{j1} \cdot (x-1) \;+\; \ldots \;+\; \tilde{\alpha}_{j,q-1} \cdot (x-1) \cdot \ldots \cdot (x-q+1)\,,$$

mit Koeffizienten $\tilde{\alpha}_{js}$, die unter Verwendung von (2.4) darstellbar sind als

$$\tilde{\alpha}_{js} \;=\; \big[(s+1)\ldots 1\big]\tilde{p} \;=\; (s!)^{-1} \cdot \Delta^s \tilde{p}(1) \;=\; (s!)^{-1} \cdot \Delta^s \hat{\gamma}_{j1}\,,\;\; s = 0, \ldots, q-1.$$

Für penalisiertes Schätzen mit (6.15) und $\lambda_{j,\emptyset} \longrightarrow \infty$, $j \in \mathcal{P}$, gilt $\Delta^d \hat{\gamma}_{jr} \longrightarrow 0$, $r = 1, \ldots, q-d$, und damit $\Delta^s \hat{\gamma}_{j1} \longrightarrow 0$ für $s \ge d$, d.h. $\tilde{p}$ reduziert sich auf

$$\tilde{p}(x) \;=\; \tilde{\alpha}_{j0} \;+\; \tilde{\alpha}_{j1} \cdot (x-1) \;+\; \ldots \;+\; \tilde{\alpha}_{j,d-1} \cdot (x-1) \cdot \ldots \cdot (x-d+1)\,,$$

bzw. nach Ausmultiplizieren und Rearrangieren der Koeffizienten

$$\tilde{p}(x) = \alpha_{j0} + \alpha_{j1} \cdot x + \alpha_{j2} \cdot x^2 + \ldots + \alpha_{j,d-1} \cdot x^{d-1}.$$

Mit der geforderten Interpolationseigenschaft folgt daraus

$$\hat{\gamma}_{jr} = \tilde{p}(r) = \alpha_{j0} + \alpha_{j1} \cdot r + \alpha_{j2} \cdot r^2 + \ldots + \alpha_{j,d-1} \cdot r^{d-1}$$

für $r = 1, \ldots, q$ und damit die Behauptung.

Die mit starker Penalisierung gegebene Reduktion der effektiven Parameteranzahl je Kovariable von q auf d ist unabhängig von der konkreten Datenlage für jede Differenzenordnung $d \in \{1, \ldots, q-1\}$ möglich. Die Wahl von d sollte jedoch nicht willkürlich, sondern datenadaptiv über ein geeignetes Kriterium (z.B. AIC) erfolgen. Liefert dieses Kriterium für die Ordnung d einen unendlichen Betrag als optimale Penalisierungsstärke der Gewichte von x_j, können die Parameter γ_{jr}, $r = 1, \ldots, q$, durch ein Polynom vom Grad $d-1$ in r beschrieben werden, und es ist legitim, das NPOM unter Beachtung einer entsprechenden Nebenbedingung zu fitten.

6.5.1 Beispiel: Übelkeit bei Chemotherapie

Ein einfaches Beispiel, welches auch von Peterson & Harrell (1990) analysiert wird, basiert auf einem Datensatz aus Farewell (1982). Farewell (1982) untersucht den Einfluß des Medikamentes Cisplatin auf den Grad des Empfindens von Übelkeit bei Chemotherapie–Patienten. Tabelle 6.4 zeigt die Daten.

	Übelkeitsempfinden						*Total*
	Kein					*Stark*	
	1	*2*	*3*	*4*	*5*	*6*	
Ohne Cisplatin	43	39	13	22	15	29	161
Mit Cisplatin	7	7	3	12	15	14	58

TABELLE 6.4: *Daten zum Empfinden von Übelkeit bei Chemotherapie.*

Farewell (1982) verwirft die Annahme globaler Effekte für die dichotome Kovariable. Peterson & Harrell (1990) fitten ein Non-Proportional Odds Modell

und vermuten, daß lediglich für die letzte Dichotomisierung der Responsevariablen eine signifikante Abweichung von der Annahme proportionaler Chancen vorliegt. Ein anschließender Test der Hypothese $\delta_1 = \ldots = \delta_4 = 0$, $\delta_5 = 1$ im restringierten Non-Proportional Odds Modell

$$\eta_{ir} = \gamma_{0r} + x_i\gamma + x_i\tilde{\gamma}\delta_r, \quad r = 1, \ldots, q = 5, \tag{6.16}$$

bestätigt diese Vermutung. Statt einer eher willkürlichen Festlegung von Parametern $\delta_1, \ldots, \delta_5$ wird das NPOM

$$\eta_{ir} = \gamma_{0r} + x_i\gamma_r, \quad r = 1, \ldots, q,$$

für den hier propagierten Ansatz penalisiert geschätzt mit verschiedenen Differenzenordnungen und wachsender Penalisierungsstärke. Die Abbildung 6.7 zeigt für die möglichen Differenzenordnungen $d \in \{1, 2, 3, 4\}$ den Verlauf des AIC in Abhängigkeit vom Glättungsparameter $\lambda_\emptyset$.

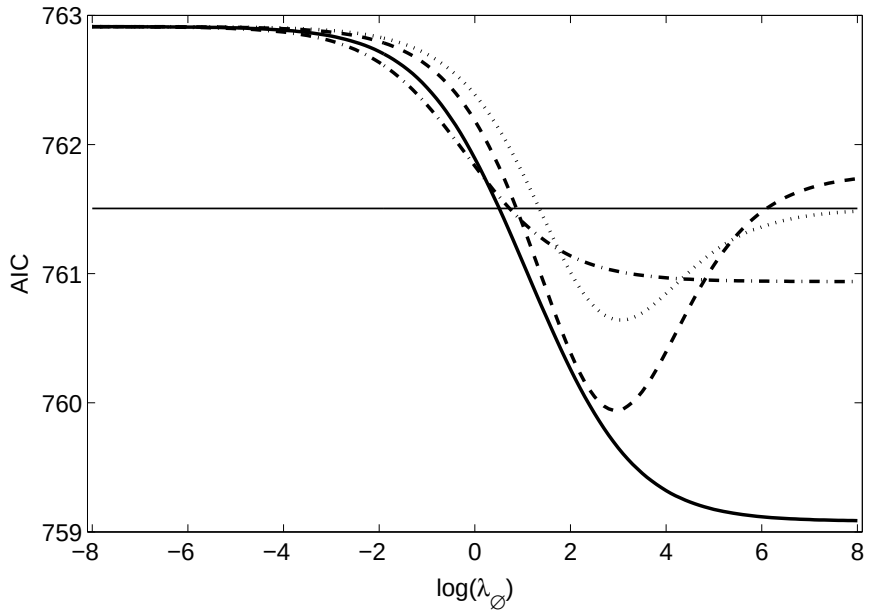

ABBILDUNG 6.7: *AIC des NPOM für variierenden Glättungsparameter und verschiedene Differenzenordnungen (1: gepunktet, 2: gestrichelt, 3: durchgezogen, 4: strich–punkt). Horizontale Linie kennzeichnet AIC des korrespondierenden Proportional Odds Modells.*

Für die Differenzenordnungen 1 und 2 weist das Akaike-Informations-Kriteri-

um ausgeprägte Minima im gewählten Bereich des Glättungsparameters auf, während es für $d \in \{ 3, 4 \}$ mit wachsender Stärke der Penalisierung beständig abnimmt. Im Extremfall $\lambda_\emptyset = \exp(8)$ sind die geschätzten Kovariableneffekte für jedes d durch ein Polynom $(d-1)$–ten Grades bestimmt, doch nur für $d \in \{3, 4\}$ ist diese Penalisierungsstärke äquivalent zur AIC–optimalen Konstellation. Der für $d = 3$ und $\lambda_\emptyset = \exp(8)$ resultierende Wert des AIC ist darüber hinaus minimal unter allen Paaren $(d, \lambda_\emptyset)$ aus Differenzenordnung und Penalisierungsstärke. Daher scheint es sinnvoll, das NPOM unter der Nebenbedingung $\gamma_r = \alpha_0 + \alpha_1 r + \alpha_2 r^2$ zu fitten. Die korrespondierende Log-Likelihood ist mit –371.54 geringfügig größer als der in Peterson & Harrell (1990) angegebene Wert –372.19 für die Log-Likelihood des Modells (6.16) mit den Konstanten $\delta_1 = \ldots = \delta_4 = 0$ und $\delta_5 = 1$.

Vergleicht man die AIC–Werte für die Differenzenordnungen 3 und 4 im Extremfall $\lambda_\emptyset = \exp(8)$, ergibt sich eine Differenz von etwa zwei Einheiten. Für $d = 3$ sind die geschätzten Kovariableneffekte in diesem Fall durch ein quadratisches, für $d = 4$ durch ein kubisches Polynom in r bestimmt – die effektiven Parameterzahlen unterscheiden sich also um eins. Die korrespondierenden Devianzen müssen daher nach (4.2) identisch sein. Für $d = 3$ und $d = 4$ resultiert bei starker Penalisierung somit derselbe Fit. Das Schätzen im Non-Proportinal Odds Modell unter der Nebenbedingung

$$\gamma_r = \alpha_0 + \alpha_1 r + \alpha_2 r^2 + \alpha_3 r^3$$

führt dementsprechend auf $\alpha_3 = 0$.

7 Semiparametrische ordinale Regression

Als Erweiterung des rein parametrischen Partial Proportional Odds Modells aus Abschnitt 6.2 wird in diesem Kapitel die Einbeziehung nonparametrischer Komponenten in die Modellierung von Kovariableneffekten bei ordinaler Responsestruktur betrachtet. Jede der im folgenden untersuchten Verallgemeinerungen des PPOM (6.8) betrifft ausschließlich den Prädiktor, wobei der nonparametrischen Schätzung wiederum das Konzept penalisierter Basisfunktionen zugrunde liegt. Damit behalten die Aussagen aus Abschnitt 6.2 zur Einbettung ordinaler Regressionsmodelle in den Rahmen multivariater GLM's auch hier ihre Gültigkeit.

Die Ausführungen zu diesem Kapitel erfolgen im wesentlichen dreigeteilt. Zunächst werden semiparametrische Formen des POM untersucht. Die Betrachtungen konzentrieren sich dabei auf eine Anwendung mit Interaktionen, deren theoretische Behandlung Inhalt der Sektionen 3.4 bzw. 3.5 war. Ein weiterer Abschnitt widmet sich der Analyse nonparametrischer Effekte, die spezifisch für die einzelnen Kategorien sein können. In diesem Zusammenhang wird das Penalisierungskonzept für kategorienspezifische Parameter aus Kapitel 6 vom rein diagnostischen Tool beim Testen der Hypothese proportionaler Chancen zum festen Bestandteil bei der Modellierung bzw. Schätzung in semiparametrischen PPOM's weiterentwickelt. Das Kapitel schließt mit einem Abschnitt zu multiplikativ verknüpften Effekten, die eine parameterökonomische Alternative bei der Modellierung von kategorienspezifischen glatten Komponenten darstellen.

7.1 Semiparametrische Erweiterungen des POM

Die strukturelle Erweiterung der restriktiven parametrischen Prädiktorform

$$\eta_{ir} = \gamma_{0r} + \boldsymbol{x}_i'\boldsymbol{\gamma}, \;\; i = 1, \ldots, N, \; r = 1, \ldots, q$$

des Proportional Odds Modells (6.3) um flexiblere nonparametrische Effekte kann ohne theoretischen Mehraufwand realisiert werden. Aufgrund des globalen Charakters der Kovariableneffekte im POM ist zu diesem Zweck ledig-

lich der lineare Einflußterm durch die Komponenten des universellen Prädiktors (3.1) zu ersetzen

$$\eta_{ir} = \gamma_{0r} + \eta_{i,L} + \eta_{i,A} + \eta_{i,V} + \eta_{i,O}.$$

Die B-Spline basierte Umsetzung der einzelnen Komponenten sowie die (penalisierte) Maximum–Likelihood–Schätzung der zugehörigen Basiskoeffizienten waren Gegenstand von Kapitel 3 und bedürfen hier keiner erneuten Betrachtung. Den Schwerpunkt dieses Abschnitts bilden demzufolge auch nicht die jeweils zugrunde liegenden Modellierungsaspekte. Der Fokus richtet sich vielmehr auf die bis dato nur unter theoretischen Gesichtspunkten behandelten Interaktionsformen (vgl. Abschnitte 3.4 und 3.5). Anhand eines konkreten Anwendungsbeispiels soll deren Berücksichtigung im ordinalen Regressionsmodell demonstriert werden.

7.1.1 Beispiel: Untersuchung von Waldschäden

Im Bereich des Forstamtes Rothenbuch (Spessart, Bayern) werden seit 1983 jährliche Waldschadensinventuren durchgeführt. Die hier analysierten Daten umfassen einen Zeitraum von 19 Jahren, 1983 - 2001, in dem der prozentuale Blattverlust (Entlaubungsgrad) $\tilde{y}_{it}$, $i = 1, \ldots, N$, $t = 1, \ldots, 19$, als jährliche ordinale Kenngröße mit den möglichen Ausprägungen

$$\tilde{y}_{it} = 1: \ 0\%, \qquad \tilde{y}_{it} = 2: \leq 25\%, \qquad \tilde{y}_{it} = 3: > 25\% \quad \text{Blattverlust}$$

an $N = 75$ verschiedenen Standorten visuell erfasst wurde. Einzig betrachtete (weil vorherrschende) Baumart ist die Buche. Untersucht werden soll der Effekt der Kovariablen

AGE_{it}:	Alter (in Jahren) von Baum i $(1, \ldots, 75)$ im Jahr t $(1, \ldots, 19)$
BsG_{it}:	Beschirmungsgrad (Ausnutzung des zur Verfügung stehenden Lichtraumes) (in %) am Standort i im Jahr t

auf den beobachteten Entlaubungsgrad. Dem metrischen Charakter der Einflußgrößen entsprechend wird ein nonparametrisches Proportional Odds Modell mit der Prädiktorspezifikation

$$\eta_{itr} = \gamma_{0r} + \gamma_{(B)}(\text{BsG}_{it}) + \gamma_{(A)}(\text{AGE}_{it}), \ r \in \{1, 2\},$$

betrachtet. Da es sich bei den beobachteten Responsevariablen um Longitudinaldaten handelt, ist die Unabhängigkeitsannahme für die $\tilde{y}_{it}$ fragwürdig. Der zeitlichen Korrelation kann jedoch durch die Berücksichtigung der Kalenderzeit t auf Kovariablenseite Rechnung getragen werden

$$\eta_{itr} = \gamma_{0r} + \gamma_{(T)}(t) + \gamma_{(B)}(\mathrm{BsG}_{it}) + \gamma_{(A)}(\mathrm{AGE}_{it}), \quad r \in \{1,2\}. \tag{7.1}$$

Ebenfalls vorhandene räumliche Abhängigkeitsbeziehungen, die sich aus natürlichen Nachbarschaften von Standorten ergeben, werden in den folgenden Analysen nicht explizit modelliert. Untersuchungen, die diese räumliche Korrelation der Daten berücksichtigen, finden sich in Pruscha & Göttlein (2002) bzw. Fahrmeir & Lang (2001) und Fahrmeir, Kneib & Lang (2003) im Kontext bayesianischer Modellierung.

Für die Generierung von Interaktionen werden die metrische Kovariable

$\mathrm{AGE}_{i,83}$: Alter (in Jahren) von Baum i $(1,\ldots,75)$ im Jahr 1983

und deren Dummy–Kodierung

$$\mathrm{AGE}^{(1)}_{i,83} = \begin{cases} 1, & \mathrm{AGE}_{i,83} < 50 \\ 0, & \text{sonst} \end{cases} \qquad \mathrm{AGE}^{(2)}_{i,83} = \begin{cases} 1, & 50 \leq \mathrm{AGE}_{i,83} \leq 120 \\ 0, & \text{sonst} \end{cases}$$

eingeführt. Durch alternatives Modellieren des Alterseffekts in (7.1) ergeben sich daraus ein Modell mit variierenden Koeffizienten

$$\eta_{itr} = \gamma_{0r} + \gamma_{(T)}(t) + \gamma_{(B)}(\mathrm{BsG}_{it}) + \alpha_{(1)}(t)\cdot\mathrm{AGE}^{(1)}_{i,83} + \alpha_{(2)}(t)\cdot\mathrm{AGE}^{(2)}_{i,83} \tag{7.2}$$

und ein Modell mit einem Oberflächeneffekt

$$\eta_{itr} = \gamma_{0r} + \gamma_{(T)}(t) + \gamma_{(B)}(\mathrm{BsG}_{it}) + \gamma_{(A,T)}(\mathrm{AGE}_{i,83}, t), \tag{7.3}$$

deren nonparametrische Komponenten über die P–Spline Ansätze aus Kapitel 3 im Rahmen multivariater GLM's geschätzt wurden.

Die Approximation der Haupteffekte in (7.2) erfolgte unter Verwendung von B–Splines dritten Grades. Für die variierenden Koeffizienten der dichotomen Altersvariablen wurden B–Splines vom Grad zwei herangezogen. Differenzenpenalties der Ordnung $d = 2$ beschränkten die Variation der zugehörigen Ba-

siskoeffizienten. Die Optimierung der korrespondierenden Glättungsparameter erfolgte auf Basis des BIC und unter Zuhilfenahme des genetischen Algorithmus aus Abschnitt 4.2. Abbildung 7.1 zeigt die resultierenden Schätzungen zusammen mit den approximativen Konfidenzbändern (5.24) ($\alpha = 0.05$).

Der geschätzte Haupteffekt der Kalenderzeit t – interpretierbar als zeitlicher Effekt für Bäume aus der Altersreferenzkategorie (in 1983 älter als 120 Jahre) – weist einen ausgeprägt nicht-linearen Verlauf auf. Einer deutlichen Zustandsverschlechterung der alten Bäume bis ins Jahr 1987 folgte eine knapp vierjährige Erholungsphase, nach der dann eine wieder zunehmende Entlaubung zu verzeichnen war. Seit etwa 1996 deutet sich für Bäume der Altersreferenzkategorie eine leichte Regeneration des Blattbewuchses an. Deutlich nicht-linear verläuft auch der geschätzte Effekt des Beschirmungsgrades. Bis zu einem Wert von ca. 55% wächst die Wahrscheinlichkeit, mehr als ein Viertel der Blätter zu verlieren, mit zunehmendem Beschirmungsgrad. Über diesen Wert hinausgehende Blattdichten können hingegen als Indikator für eine geringere Entlaubung betrachtet werden. Die beiden mittleren Darstellungen in Abbildung 7.1 zeigen die Schätzungen der (zeit-)variierenden Koeffizienten. Informativer sind jedoch die zusätzlich angegebenen zeitlichen Effekte der zwei modellierten Alterskategorien, die aus der Summation von zugehörigem variierenden Koeffizienten und Haupteffekt der Kalenderzeit hervorgehen. Die Interpretation der unteren Darstellungen führt zu gleichlautenden Aussagen wie für alte Bäume. Die sich ab 1996 abzeichnende Erholung ist für Bäume mittleren Alters jedoch deutlicher ausgeprägt, wohingegen junge Bäume seit 1992 eine kontinuierliche Zustandsverschlechterung aufweisen.

Die Haupteffekte von Modell (7.3) wurden mit kubischen B–Splines approximiert. Differenzenpenalties zweiter Ordnung beschränkten die Variation der zugehörigen Basiskoeffizienten. Für die Modellierung des Oberflächeneffekts kamen 140 (14×10) zweidimensionale B–Splines vierten Grades zur Anwendung. Ein Zeilen- und ein Spaltenpenalty erster Ordnung dienten zur Variationskontrolle der entsprechenden Basiskoeffizienten. BIC–optimierte Glättungsparameter wurden unter Verwendung des genetischen Algorithmus aus Kapitel 4 gewonnen. Abbildung 7.2 zeigt die korrespondierenden Schätzungen. Für den Beschirmungsgrad ergibt sich ein ähnliches Bild wie in Modell

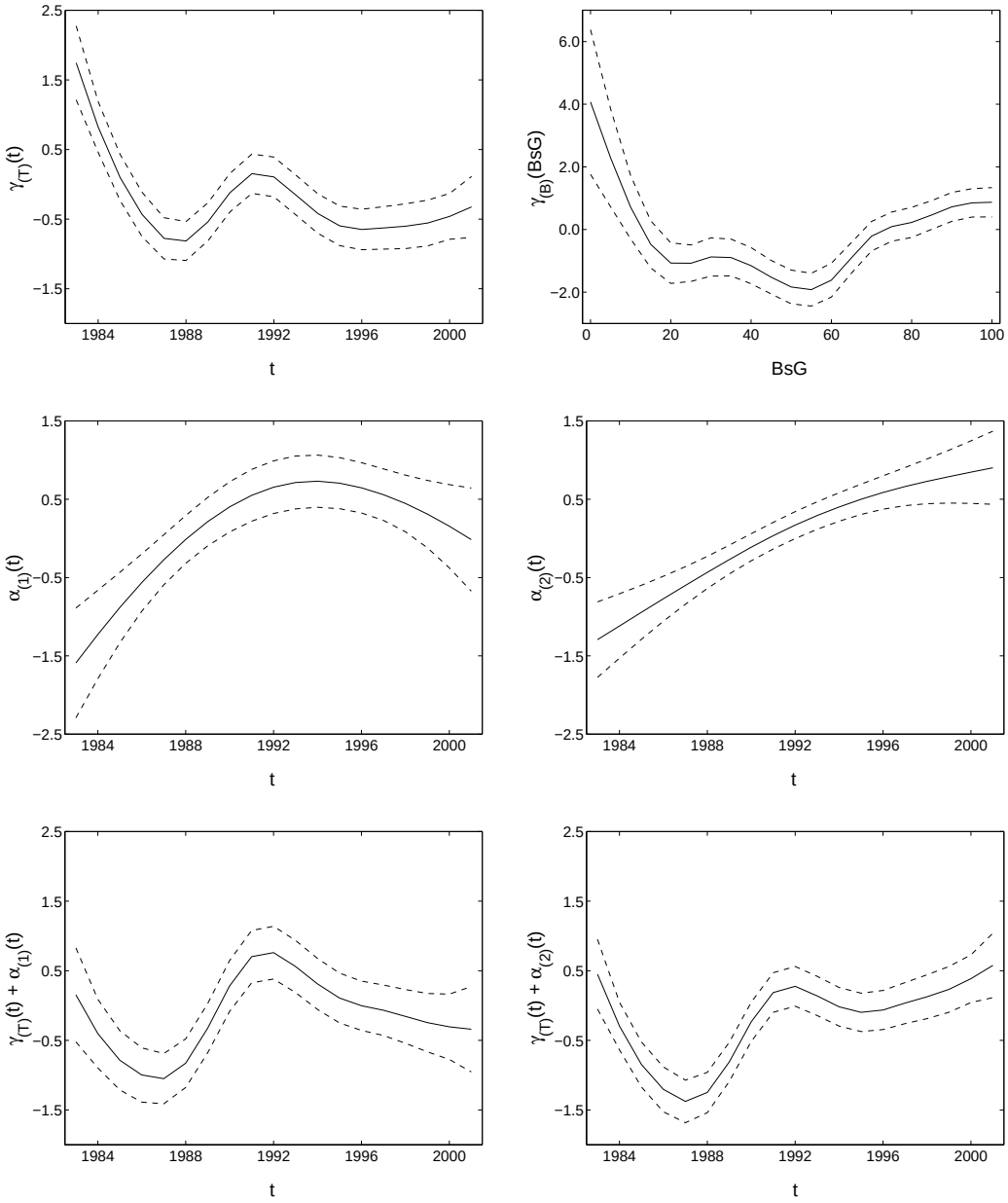

ABBILDUNG 7.1: *Oben: Geschätzte Haupteffekte der Kalenderzeit und des Beschirmungsgrades im Proportional Odds Modell (7.2). Mitte: Variierende Koeffizienten für junge Bäume (links) und Bäume mittleren Alters (rechts). Unten: Kalenderzeiteffekte für junge Bäume (links) sowie Bäume mittleren Alters (rechts). Approximative Konfidenzbänder nach (5.24) sind als gestrichelte Linien dargestellt.*

(7.2). Die interpretatorischen Aussagen können daher übernommen werden. Zeitliche Effekte für Bäume eines bestimmten Alters resultieren als Summe aus zugehöriger Schnittfunktion des Oberflächeneffekts parallel zur Zeitachse und Haupteffekt der Kalenderzeit. Die Schnittfunktionen parallel zur Altersachse lassen sich als jahresspezifische Alterseffekte interpretieren.

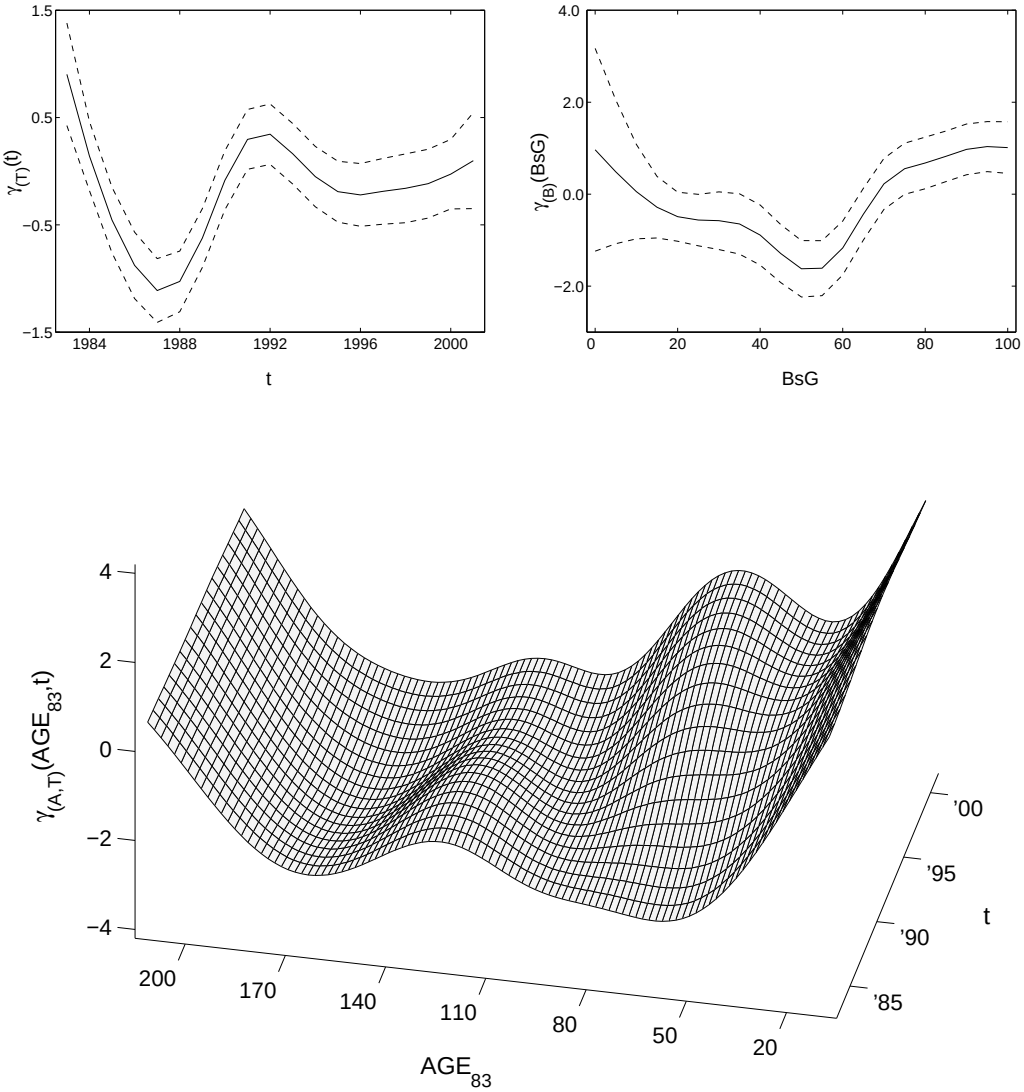

ABBILDUNG 7.2: *Oben: Geschätzte Haupteffekte der Kalenderzeit und des Beschirmungsgrades im Modell (7.3). Approximative Konfidenzbänder sind als gestrichelte Linien dargestellt. Unten: Geschätzter Interaktionseffekt des Alters zu Beginn des Beobachtungszeitraumes (1983) und der Kalenderzeit.*

7.2 Semiparametrische Erweiterungen des PPOM

In Verallgemeinerung des PPOM($\mathcal{G}$), $\mathcal{G} \subset \mathcal{P} = \{1, \ldots, p\}$, werden in diesem Abschnitt die Effekte stetiger Kovariablen x_j, $j \in \mathcal{S} \subset \mathcal{P}$, in unspezifizierter funktionaler Form modelliert. Dementsprechend modifizieren sich die Prädiktoren (6.8) nach Partitionierung der parametrischen Terme zu

$$\begin{aligned}\eta_{ir} &= \gamma_{0r} + \eta_{i,L} + \eta_{i,A} + \eta_{ir,L} + \eta_{ir,A} \\ &= \gamma_{0r} + \sum_{j \in \mathcal{G} \cap \mathcal{D}} x_{ij}\gamma_j + \sum_{j \in \mathcal{G} \cap \mathcal{S}} \alpha_{(j)}(x_{ij}) + \sum_{j \in \bar{\mathcal{G}} \cap \mathcal{D}} x_{ij}\gamma_{jr} + \sum_{j \in \bar{\mathcal{G}} \cap \mathcal{S}} \alpha_{(j),r}(x_{ij}), \quad (7.4)\end{aligned}$$

wobei in $\mathcal{D} := \mathcal{P} \backslash \mathcal{S}$ die Indizes aller diskreten Kovariablen x_j, $j \in \mathcal{P}$, zusammengefasst sind.

Die Modellierung der nonparametrischen Komponenten in (7.4) erfolgt unter Verwendung entsprechender B–Spline–Basen. In Anlehnung an Abschnitt 3.3 setzt man

$$\alpha_{(j)}(x_{ij}) = \sum_{s=1}^{P} \alpha_{js} G_{js}(x_{ij}) \quad \text{bzw.} \quad \alpha_{(j),r}(x_{ij}) = \sum_{s=1}^{P} \alpha_{jrs} G_{js}(x_{ij}) \qquad (7.5)$$

und ermöglicht damit die Rückführung des semiparametrischen PPOM in die Klasse multivariater GLM's. Da dies für Prädiktoren mit parametrischen und nonparametrischen Komponenten globaler bzw. kategorienspezifischer Natur wiederholt demonstriert wurde, kann hier auf Einzelheiten verzichtet werden.

Inhaltlich neu ist hingegen die für semiparametrische PPOM's mögliche Kombination der verschiedenen Penalisierungskonzepte. Die Verwendung der Darstellungen (7.5) zur Modellierung glatter Komponenten erfordert eine Bestrafung der jeweiligen Basiskoeffizienten. Darüber hinaus ist mit der Präsenz kategorienspezifischer Parameter deren Penalisierung über die Responsekategorien hinweg möglich. Bei Berücksichtigung aller denkbaren Strafterme ist die penalisierte Log–Likelihood für Modell (7.4) von der Form

$$\begin{aligned}pl(\boldsymbol{\eta}) &= l(\boldsymbol{\eta}) - \frac{1}{2} \cdot \sum_{j \in \{0\} \cup \bar{\mathcal{G}} \cap \mathcal{D}} \delta_j \sum_{r=1}^{q-d} (\Delta_v^d \gamma_{jr})^2 - \frac{1}{2} \cdot \sum_{j \in \mathcal{G} \cap \mathcal{S}} \lambda_j \sum_{s=1}^{P-d} (\Delta_h^d \alpha_{js})^2 \\ &\quad - \frac{1}{2} \cdot \sum_{j \in \bar{\mathcal{G}} \cap \mathcal{S}} \left\{ \sum_{r=1}^{q} \lambda_{jr} \sum_{s=1}^{P-d} (\Delta_h^d \alpha_{jrs})^2 + \sum_{s=1}^{P} \delta_{js} \sum_{r=1}^{q-d} (\Delta_v^d \alpha_{jrs})^2 \right\}. \quad (7.6)\end{aligned}$$

Aufgrund ihres kategorienspezifischen Charakters unterliegen die Basiskoeffizienten der Funktionen $\alpha_{(j),r}$ einer doppelten Penalisierung. Zur besseren Unterscheidung sind die Differenzenoperatoren Δ^d daher mit einem zusätzlichen Index versehen. Dabei bedeuten

$$\Delta_h^d \alpha_{jrs} = \Delta_h^{d-1} \alpha_{jr,s+1} - \Delta_h^{d-1} \alpha_{jrs} \text{ und } \Delta_v^d \alpha_{jrs} = \Delta_v^{d-1} \alpha_{j,r+1,s} - \Delta_v^{d-1} \alpha_{jrs}.$$

Der Index h kennzeichnet die sogenannte *horizontale* Penalisierung der Basiskoeffizienten innerhalb einer Kategorie. Für $r \in \{1, \ldots, q\}$ fest, bestimmt der horizontale Penalty die Variation des Effekts $\alpha_{(j),r}$ über benachbarte Knoten hinweg. Umgekehrt kennzeichnet der Index v die sogenannte *vertikale* Penalisierung der zu einem Knoten gehörigen Basiskoeffizienten in den Kategorien. Der vertikale Penalty entspricht der in Kapitel 6 vorgestellten Penalisierungsform und bestimmt die Variation von $\alpha_{(j),r}$ über benachbarte Kategorien hinweg. Abbildung 7.3 verdeutlicht nochmals das Konzept der doppelten Penalisierung kategorienspezifischer Basiskoeffizienten.

Die Stärke der Penalisierung kategorienspezifischer Basiskoeffizienten wird in horizontaler Richtung durch Glättungsparameter λ_{jr}, $r = 1, \ldots, q$, in vertikaler Richtung durch Glättungsparameter δ_{js}, $s = 1, \ldots, P$, bestimmt. Eine kategorienspezifische horizontale Penalisierung ist aufgrund kleiner Kategorienzahlen oft vertretbar, die für jeden Knoten einer Basis variable Penalisierung der zugehörigen Basiskoeffizienten erscheint jedoch als zu komplex. Daher sei vereinbart, daß die vertikale Penalisierungsstärke für alle Knoten einer Basis identisch ist, d.h. $\delta_{js} \equiv \delta_j$ für $j \in \bar{\mathcal{G}} \cap \mathcal{S}$ und $s = 1, \ldots, P$.

Die verbliebenen Penalties in (7.6) betreffen die Schwellenwerte, die kategorienspezifischen Gewichte diskreter Kovariablen und die Basiskoeffizienten global und unspezifiziert modellierter metrischer Einflußgrößen. Obwohl für diese Parameter nur eine Form der Penalisierung in Frage kommt, wurden auch deren Strafterme aus Konsistenzgründen als horizontal bzw. vertikal wirkend gekennzeichnet. Auf Einzelheiten der Maximierung von (7.6) wird an dieser Stelle verzichtet.

Kategorienübergreifende Penalties dienten im Kapitel 6 ausschließlich einer Robustifizierung des Fisher–Scoring. Die Wahl der Glättungsparameter war

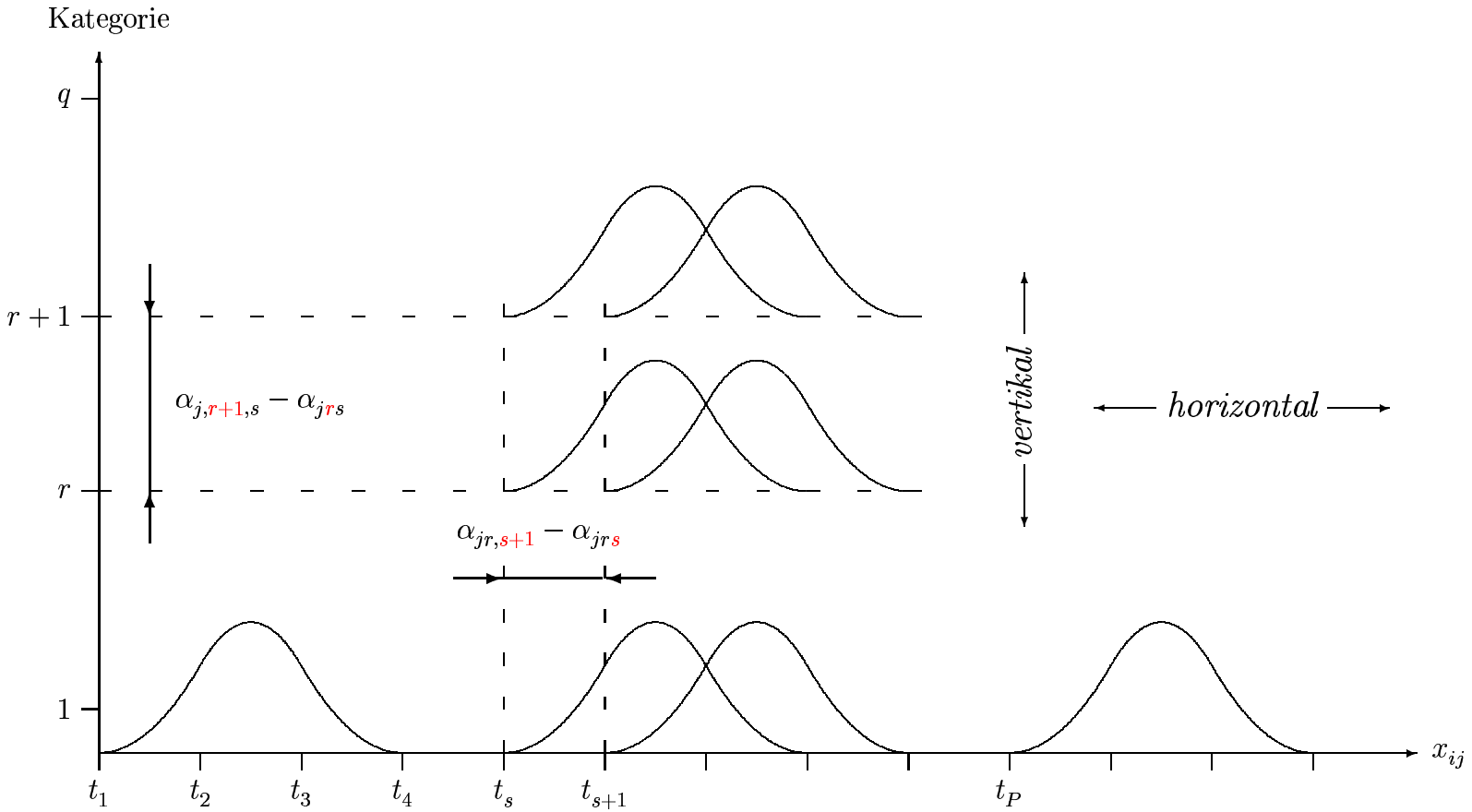

ABBILDUNG 7.3: *Veranschaulichung horizontaler und vertikaler Penalties erster Ordnung für kategorienspezifische Basiskoeffizienten und B–Splines zweiten Grades.*

zweckgebunden, da sie sich an rein numerischen Notwendigkeiten orientierte. Hier hingegen ist die vertikale Penalisierung kategorienspezifischer Parameter als konzeptioneller Bestandteil der Maximum–Likelihood–Schätzung im Partial Proportional Odds Modell aufzufassen. Neben den horizontalen sind entsprechend auch die vertikalen Glättungsparameter in (7.6) bezüglich eines geeigneten Kriteriums (z.B. AIC) zu optimieren.

7.2.1 Beispiel: Diabetische Retinopathie

In Abschnitt 6.4 wurde der Einfluß verschiedener Risikofaktoren auf die Entwicklung diabetischer Retinopathie untersucht. Die Analyse im rein parametrischen Modell bestätigte die Hypothese proportionaler Chancen für die Kovariablen glykosyliertes Hämoglobin und diastolischer Blutdruck. Für die Risikofaktoren Raucherstatus und Diabetesdauer indizierten die Resultate hingegen eine kategorienspezifische Modellierung. Darauf aufbauend werden die Effekte der metrischen Kovariablen in diesem Abschnitt in funktionaler Form modelliert und das semiparametrische PPOM($\{GH, BP\}$) mit der Prädiktorspezifikation

$$\eta_{ir} = \gamma_{0r} + SM_i \cdot \gamma_{SM,r} + \alpha_{(GH)}(GH_i) + \alpha_{(BP)}(BP_i) + \alpha_{(DD),r}(DD_i),$$

$r \in \{1, 2\}$, gefittet. Ein möglicher Effekt der quadrierten Diabetesdauer DDQ wie im parametrischen Modell wird durch die Funktionen $\alpha_{(DD),r}$, $r \in \{1, 2\}$, ebenfalls abgedeckt. Für die nonparametrischen Komponenten sind vier horizontale sowie ein vertikaler Glättungsparameter zu bestimmen. Hinzu kommt ein weiterer vertikaler Glättungsparameter für die kategorienspezifischen Gewichte des Raucherstatus, so daß bei Nicht-Penalisierung der Schwellenwerte insgesamt sechs Glättungsparameter zu optimieren sind. Da $q = 2$, kommen für die vertikalen Penalties nur Differenzen erster Ordnung in Frage, die vier horizontalen Strafterme werden mit $d = 2$ angesetzt. Die Basisrepräsentation der nonparametrischen Komponenten erfolgt in je 20 äquidistanten B-Splines dritten Grades.

Zur Bestimmung der Glättungsparameter wird das Akaike-Informations-Kriterium herangezogen, dessen Minimierung unter Verwendung des genetischen Algorithmus aus Kapitel 4 erfolgt. Abbildung 7.4 zeigt die maximale Fitness

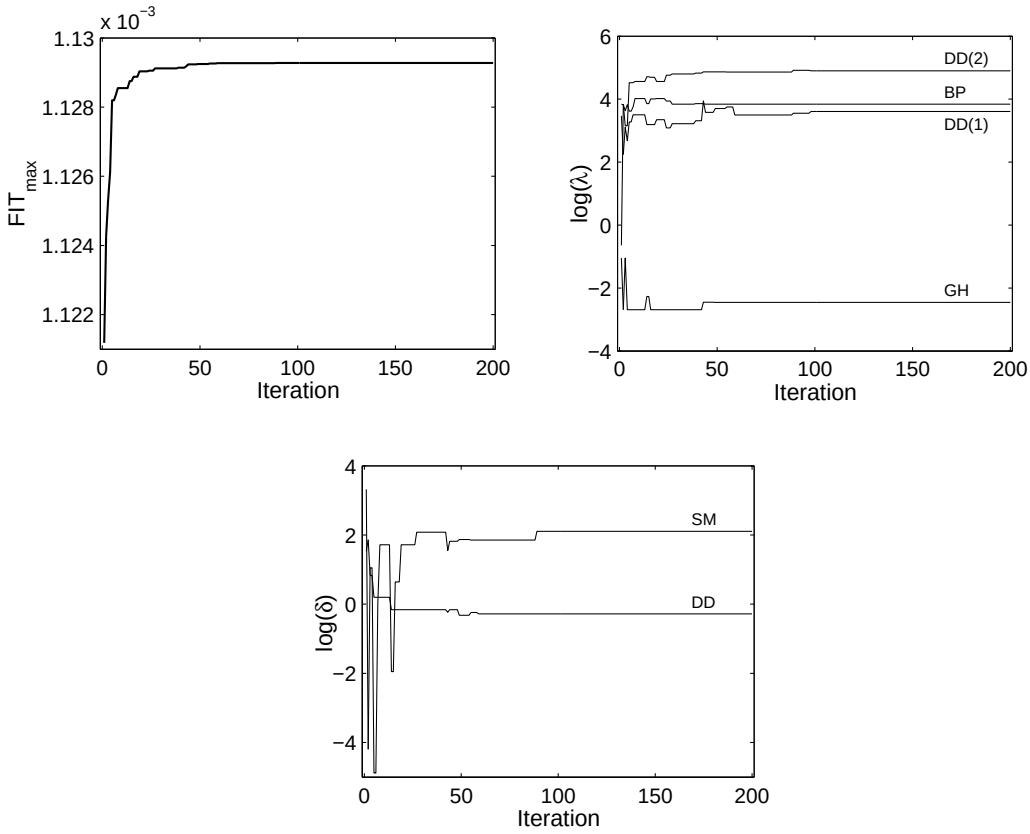

ABBILDUNG 7.4: *Verlauf der maximalen Fitnessfunktion (links) und der zugehörigen, logarithmierten Glättungsparameter (horizontal: rechts, vertikal: unten) im Datensatz zur diabetischen Retinopathie.*

j	λ_j	λ_{j1}	λ_{j2}	δ_j
SM	–.–	–.–	–.–	8.20
GH	0.09	–.–	–.–	–.–
BP	46.45	–.–	–.–	–.–
DD	–.–	36.97	134.54	0.76

TABELLE 7.1: *AIC–optimale, horizontale und vertikale Glättungsparameter im Datensatz zur diabetischen Retinopathie.*

der aktuellen Population in Abhängigkeit vom Iterationszähler sowie die Entwicklung der korrespondierenden horizontalen und vertikalen Glättungsparameter. Nach etwa 90 Iterationen stellt sich ein stabiler Zustand ein, der über die folgenden mehr als 100 Schritte nicht wieder verlassen wird. In Tabelle 7.1 sind die diesen Zustand charakterisierenden Glättungsparameter aufgelistet.

Tabelle 7.2 zeigt die Parameterschätzungen und geschätzten Standardabweichungen für die Schwellenwerte sowie die kategorienspezifischen Gewichte des Raucherstatus. Geschätzte Standardabweichungen liefert auch hier der Sandwich–Schätzer aus Abschnitt 5.3. Wie schon im rein parametrischen Modell weist der Raucherstatus lediglich in der ersten Kategorie einen signifikanten Effekt auf.

	$\hat{\gamma}_{0r}$	$\hat{s}(\hat{\gamma}_{0r})$	$\hat{\gamma}_{SM,r}$	$\hat{s}(\hat{\gamma}_{SM,r})$
Kategorie 1	1.771	0.757	-0.399	0.206
Kategorie 2	2.748	0.760	-0.049	0.227

TABELLE 7.2: *Parameterschätzungen und geschätzte Standardabweichungen der parametrischen Effekte im Datensatz zur diabetischen Retinopathie.*

In Abbildung 7.5 sind die mit den AIC–optimalen Glättungsparametern aus Tabelle 7.1 korrespondierenden glatten Schätzungen für die metrischen Kovariablen dargestellt. Der monoton fallende globale Effekt $\alpha_{(BP)}$ deutet auf eine mit steigendem Blutdruck abnehmende Wahrscheinlichkeit hin, nicht an Retinopathie zu erkranken, wohingegen die Wahrscheinlichkeit für fortgeschrittene Retinopathie wächst. Gleiches gilt für den geschätzten Effekt der Kovariablen GH, lediglich für Patienten mit mehr als 12% glykosyliertem Hämoglobin kehren sich die Aussagen um. Der Verlauf des Effekts $\alpha_{(DD),2}$ läßt auf ein wachsendes Risiko schließen, fortgeschrittene Retinopathie zu entwickeln, je länger der Diabetes dauert. Umgekehrt sinkt die Wahrscheinlichkeit, nicht an Retinopathie zu erkranken, mit fortdauerndem Diabetes zunächst ab, um etwa 25 Jahre nach Diagnose des Diabetes wieder anzusteigen.

Abbildung 7.5 zeigt darüber hinaus für jeden geschätzten glatten Effekt die gemäß (5.24) kalkulierten approximativen Konfidenzbänder.

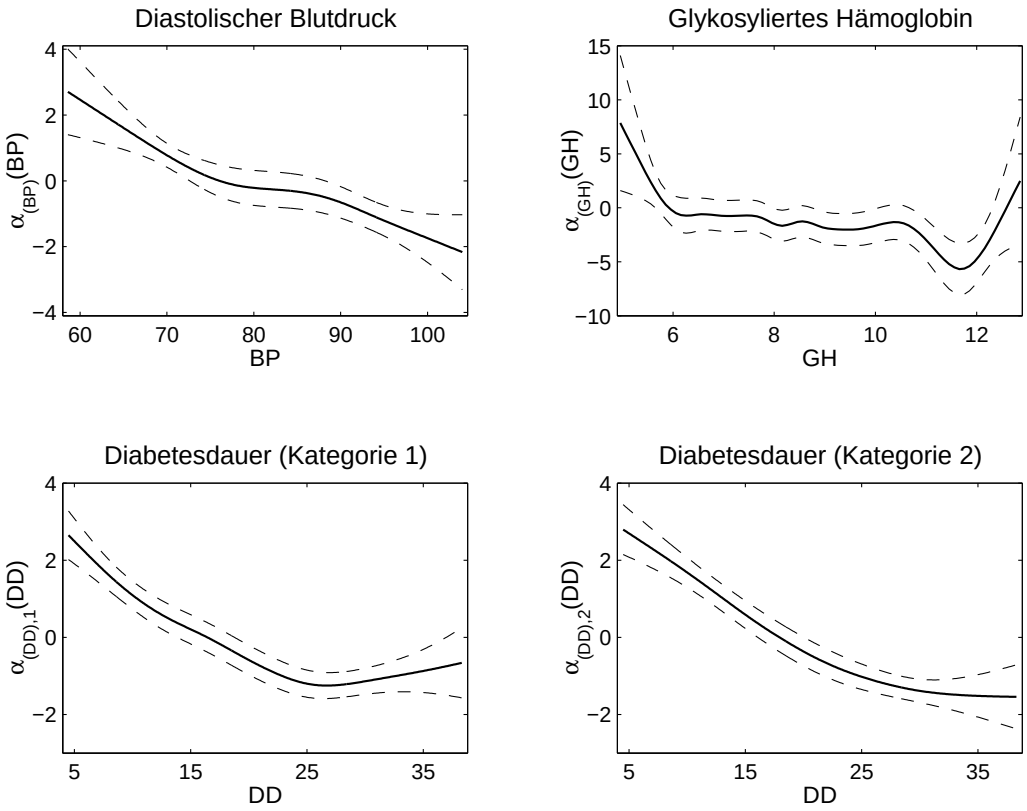

ABBILDUNG 7.5: *Geschätzte nonparametrische Effekte im Datensatz zur diabetischen Retinopathie. Oben: globale Effekte des diastolischen Blutdrucks und des glykosylierten Hämoglobin. Unten: kategorienspezifische Effekte der Diabetesdauer. Approximative Konfidenzbänder sind gestrichelt dargestellt.*

7.3 Modelle mit multiplikativen Effekten

Modelle mit kategorienspezifischen Komponenten stellen ein äußerst flexibles Instrument bei der Nachbildung von Einflußstrukturen dar, für die ein globales Wirken nicht adäquat erscheint. Insbesondere für glatte Effekte geht diese Flexibilität jedoch deutlich zu Lasten einer überschaubaren Modellkomplexität. Daraus resultierende numerische Probleme beim Fitten des Modells können zwar durch entsprechend starke Penalties vermieden werden, einfache Interpretationen der geschätzten Effekte sind hingegen kaum noch möglich, da eine Effektbewertung unter Berücksichtigung der Kurven aller Responsekategorien zu erfolgen hat.

Die folgenden Ausführungen behandeln einen Ansatz, der die Dimensionalität im semiparametrischen PPOM wesentlich reduziert, ohne auf die Flexibilität kategorienspezifischer Modellierung ganz verzichten zu müssen. Für Erklärungszwecke werde vereinfachend ein Non–Proportional Odds Modell mit ausschließlich glatten Effekten betrachtet, für die die übliche Repräsentation in Basisdarstellung angesetzt wird

$$\eta_{ir} \;=\; \gamma_{0r} + \sum_{j=1}^{p} \alpha_{(j),r}(x_{ij}) \;=\; \gamma_{0r} + \sum_{j=1}^{p} \sum_{s=1}^{P} \alpha_{jrs} G_{js}(x_{ij})\,, \;\; r = 1, \ldots, q. \quad (7.7)$$

Statt einer direkten Maximum–Likelihood–Schätzung zerlegt man die kategorienspezifischen Basiskoeffizienten zunächst faktoriell in

$$\alpha_{jrs} \;=\; \alpha_{jr} \tilde{\alpha}_{js}\,, \;\; r = 1, \ldots, q, \; s = 1, \ldots, P.$$

Für die nonparametrischen Effekte resultiert damit die Darstellung

$$\alpha_{(j),r}(x) \;=\; \sum_{s=1}^{P} \alpha_{jrs} G_{js}(x) \;=\; \alpha_{jr} \sum_{s=1}^{P} \tilde{\alpha}_{js} G_{js}(x) \;:=\; \alpha_{jr} \alpha_{(j)}(x)\,,$$

als Produkt aus einer globalen Funktion $\alpha_{(j)}(x)$, die den Basiseffekt der Kovariablen x_j repräsentiert und kategorienspezifischen Faktoren α_{jr}, $r = 1, \ldots, q$, die den Basiseffekt in den Responsekategorien multiplikativ modifizieren. Die Anzahl der zur Modellierung eines Kovariableneffekts erforderlichen Parameter reduziert sich damit von $P \cdot q$ in (7.7) auf $P + q$ im nachstehenden *Modell mit multiplikativen Effekten*

$$\eta_{ir} \;=\; \gamma_{0r} \;+\; \sum_{j=1}^{p} \alpha_{jr} \alpha_{(j)}(x_{ij}) \;=\; \gamma_{0r} \;+\; \sum_{j=1}^{p} \alpha_{jr} \sum_{s=1}^{P} \tilde{\alpha}_{js} G_{js}(x_{ij})\,, \qquad (7.8)$$

$r = 1, \ldots, q$. Die penalisierte Log–Likelihood für Modell (7.8) hat die Gestalt

$$pl(\boldsymbol{\beta}) \;=\; l(\boldsymbol{\beta}) \;-\; \frac{1}{2} \cdot P_{\boldsymbol{\beta}}\,,$$

wobei

$$\begin{aligned} P_{\boldsymbol{\beta}} \;&=\; \delta_0 \sum_{r=1}^{q-d} (\Delta_v^d \gamma_{0r})^2 \;+\; \sum_{j=1}^{p} \lambda_j \sum_{s=1}^{P-d} (\Delta_h^d \tilde{\alpha}_{js})^2 \;+\; \sum_{j=1}^{p} \delta_j \sum_{r=1}^{q-d} (\Delta_v^d \alpha_{jr})^2 \\ &=\; \boldsymbol{\gamma}_0' K_q^{(0)} \boldsymbol{\gamma}_0 \;+\; \sum_{j=1}^{p} \left\{ \tilde{\boldsymbol{\alpha}}_j' K_P^{(j)} \tilde{\boldsymbol{\alpha}}_j \;+\; \boldsymbol{\alpha}_j' K_q^{(j)} \boldsymbol{\alpha}_j \right\}, \end{aligned}$$

mit $\boldsymbol{\beta} = (\boldsymbol{\gamma}_0', \tilde{\boldsymbol{\alpha}}_1', \ldots, \tilde{\boldsymbol{\alpha}}_p', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_p')'$, $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0q})'$ und

$$\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jq})', \quad \tilde{\boldsymbol{\alpha}}_j = (\tilde{\alpha}_{j1}, \ldots, \tilde{\alpha}_{jP})', \quad K_P^{(j)} = \lambda_j (D_P^d)' D_P^d,$$

für $j = 1, \ldots, p$, sowie $K_q^{(j)} = \delta_j (D_q^d)' D_q^d$, für $j = 0, \ldots, p$.

Eine direkte Maximierung der penalisierten Log–Likelihood ist zwar möglich, diese kann aber aufgrund der multiplikativen Verknüpfung einzelner Parameter nicht im Kontext generalisierter linearer Modelle erfolgen. Statt einer direkten Schätzprozedur wird daher ein in Tutz (2003b) vorgeschlagenes zweistufiges Verfahren betrachtet:

Stufe 1: Schwellenwerte und Basiskoeffizienten

Mit $\boldsymbol{c}_{ij} := \big(G_{j1}(x_{ij}), \ldots, G_{jP}(x_{ij})\big)'$, $j = 1, \ldots, p$, ist (7.8) äquivalent zu

$$\eta_{ir} = \gamma_{0r} + \alpha_{1r} \boldsymbol{c}_{i1}' \tilde{\boldsymbol{\alpha}}_1 + \ldots + \alpha_{pr} \boldsymbol{c}_{ip}' \tilde{\boldsymbol{\alpha}}_p, \quad r = 1, \ldots, q. \tag{7.9}$$

Der volle Prädiktor $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iq})'$ ist damit gegeben als

$$\boldsymbol{\eta}_i = \boldsymbol{\gamma}_0 + D_{(1)} C_{i1} \tilde{\boldsymbol{\alpha}}_1 + \ldots + D_{(p)} C_{ip} \tilde{\boldsymbol{\alpha}}_p, \tag{7.10}$$

wobei $D_{(j)} = \mathrm{diag}(\boldsymbol{\alpha}_j)$ und $C_{ij} = \mathbf{1}_q \otimes \boldsymbol{c}_{ij}'$, $j = 1, \ldots, p$. Für fixierte Parameter $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ erhält man daraus ein multivariates GLM $\boldsymbol{\eta}_i = Z_{1i} \boldsymbol{\beta}_1$ mit der Designmatrix

$$Z_{1i} = \big[I_q \mid D_{(1)} C_{i1} \mid \ldots \mid D_{(p)} C_{ip} \big]$$

und unbekanntem Parametervektor $\boldsymbol{\beta}_1 = \big(\boldsymbol{\gamma}_0', \tilde{\boldsymbol{\alpha}}_1', \ldots, \tilde{\boldsymbol{\alpha}}_p'\big)'$. Die korrespondierende penalisierte Log–Likelihood

$$pl(\boldsymbol{\beta}_1) = l(\boldsymbol{\beta}_1) - \frac{1}{2} \cdot \big\{ \boldsymbol{\gamma}_0' K_q^{(0)} \boldsymbol{\gamma}_0 + \tilde{\boldsymbol{\alpha}}_1' K_P^{(1)} \tilde{\boldsymbol{\alpha}}_1 + \ldots + \tilde{\boldsymbol{\alpha}}_p' K_P^{(p)} \tilde{\boldsymbol{\alpha}}_p \big\}$$

kann demnach im Rahmen generalisierter linearer Modelle maximiert werden. Als penalisierte Score–Funktion erhält man

$$ps(\boldsymbol{\beta}_1) = \partial pl(\boldsymbol{\beta}_1) / \partial \boldsymbol{\beta}_1 = \sum_{i=1}^{N} Z_{1i}' D_i \Sigma_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\pi}_i) - K_1 \boldsymbol{\beta}_1,$$

mit $D_i = \partial h(\boldsymbol{\eta}_i) / \partial \boldsymbol{\eta}$, $\Sigma_i = \mathrm{cov}(\boldsymbol{y}_i) = \mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i'$ und der Penaltymatrix $K_1 = \mathrm{Diag}\big(K_q^{(0)}, K_P^{(1)}, \ldots, K_P^{(p)}\big)$. Die Schätzgleichungen $ps(\boldsymbol{\beta}_1) = \mathbf{0}$ werden iterativ via Fisher–Scoring gelöst

$$\hat{\boldsymbol{\beta}}_1^{(k+1)} = \hat{\boldsymbol{\beta}}_1^{(k)} + \tilde{F}\left(\hat{\boldsymbol{\beta}}_1^{(k)}\right)^{-1} ps\left(\hat{\boldsymbol{\beta}}_1^{(k)}\right),\ k = 0, 1, 2, \ldots \tag{7.11}$$

mit der Pseudo–Fisher–Matrix

$$\tilde{F}(\boldsymbol{\beta}_1) = \sum_{i=1}^{N} Z_{1i}'\, D_i \Sigma_i^{-1} D_i'\, Z_{1i} + K_1.$$

Für $j \in \{1, \ldots, p\}$ und eine Konstante $c \neq 0$ gilt gemäß (7.10)

$$\boldsymbol{\eta}_i = \boldsymbol{\gamma}_0 + \ldots + D_{(j)} C_{ij} \tilde{\boldsymbol{\alpha}}_j + \ldots = \boldsymbol{\gamma}_0^* + \ldots + D_{(j)} C_{ij} \tilde{\boldsymbol{\alpha}}_j^* + \ldots,$$

mit $\tilde{\boldsymbol{\alpha}}_j^* = \tilde{\boldsymbol{\alpha}}_j + c \cdot \mathbf{1}_P$ und $\boldsymbol{\gamma}_0^* = \boldsymbol{\gamma}_0 - c \cdot \boldsymbol{\alpha}_j$, d.h. auch hier treten die bereits in Kapitel 3 erwähnten Identifikationsprobleme beim Modellieren mit B–Splines auf. Einfache Abhilfe schaffen wiederum die Restriktionen

$$\tilde{\boldsymbol{\alpha}}_j' \mathbf{1}_P = 0 \quad \text{bzw.} \quad \tilde{\alpha}_{jP} = -\tilde{\alpha}_{j1} - \ldots - \tilde{\alpha}_{j,P-1},\ j = 1, \ldots, p, \tag{7.12}$$

mit einhergehender Reduktion des Parametervektors $\boldsymbol{\beta}_1$.

Eine entsprechende Berücksichtigung im Design resultiert aus der Ersetzung von $\boldsymbol{c}_{ij}$ durch $\tilde{\boldsymbol{c}}_{ij} = \left[I_{P-1} \,|\, -\mathbf{1}_{P-1}\right] \boldsymbol{c}_{ij}$, $j = 1, \ldots, p$. Die ebenfalls notwendige Modifikation der Differenzenmatrix D_P^d erfolgt analog zu Abschnitt 3.3.1.

Stufe 2: Kategorienspezifische Faktoren

Ausgehend von (7.9) läßt sich der volle Prädiktor auch schreiben als

$$\boldsymbol{\eta}_i = \boldsymbol{\gamma}_0 + (\boldsymbol{c}_{i1}' \tilde{\boldsymbol{\alpha}}_1)\, \boldsymbol{\alpha}_1 + \ldots + (\boldsymbol{c}_{ip}' \tilde{\boldsymbol{\alpha}}_p)\, \boldsymbol{\alpha}_p\,. \tag{7.13}$$

Für festgehaltene Parameter $\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_p$ erhält man daraus ein multivariates GLM $\boldsymbol{\eta}_i = \boldsymbol{\gamma}_0 + Z_{2i}\boldsymbol{\beta}_2$ mit bekanntem Offset $\boldsymbol{\gamma}_0$, Designmatrix

$$Z_{2i} = \left[\, (\boldsymbol{c}_{i1}' \tilde{\boldsymbol{\alpha}}_1) I_q \mid \ldots \mid (\boldsymbol{c}_{ip}' \tilde{\boldsymbol{\alpha}}_p) I_q \,\right]$$

und unbekanntem Parametervektor $\boldsymbol{\beta}_2 = \left(\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_p'\right)'$. Die korrespondierende penalisierte Log–Likelihood

$$pl(\boldsymbol{\beta}_2) = l(\boldsymbol{\beta}_2) - \frac{1}{2} \cdot \left\{ \boldsymbol{\alpha}_1' K_q^{(1)} \boldsymbol{\alpha}_1 + \ldots + \boldsymbol{\alpha}_p' K_q^{(p)} \boldsymbol{\alpha}_p \right\}$$

kann demnach im Rahmen generalisierter linearer Modelle maximiert werden. Als penalisierte Score–Funktion erhält man

$$ps(\boldsymbol{\beta}_2) \;=\; \partial pl(\boldsymbol{\beta}_2)/\partial \boldsymbol{\beta}_2 \;=\; \sum_{i=1}^{N} Z_{2i}' D_i \Sigma_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\pi}_i) - K_2 \boldsymbol{\beta}_2,$$

mit $D_i = \partial h(\boldsymbol{\eta}_i)/\partial \boldsymbol{\eta}$, $\Sigma_i = \mathrm{cov}(\boldsymbol{y}_i) = \mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i'$ und der Penaltymatrix $K_2 = \mathrm{Diag}\big(K_q^{(1)}, \ldots, K_q^{(p)}\big)$. Das Lösen der Schätzgleichungen $ps(\boldsymbol{\beta}_2) = \mathbf{0}$ erfolgt wiederum iterativ via Fisher–Scoring

$$\hat{\boldsymbol{\beta}}_2^{(k+1)} = \; \hat{\boldsymbol{\beta}}_2^{(k)} \;+\; \tilde{F}\Big(\hat{\boldsymbol{\beta}}_2^{(k)}\Big)^{-1} ps\Big(\hat{\boldsymbol{\beta}}_2^{(k)}\Big), \;\; k = 0, 1, 2, \ldots \qquad (7.14)$$

mit der Pseudo–Fisher–Matrix

$$\tilde{F}(\boldsymbol{\beta}_2) \;=\; \sum_{i=1}^{N} Z_{2i}' \, D_i \Sigma_i^{-1} D_i' \, Z_{2i} \;+\; K_2.$$

Die unbekannten Parameter $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ sind dabei in jedem Iterationsschritt eindeutig identifizierbar. Zusätzliche Restriktionen zur Ausführung von Stufe 2 können daher entfallen. Iteratives Hintereinanderschalten der Stufen 1 und 2 bis zur Konvergenz erlaubt eine Schätzung der unbekannten Parameter im Rahmen generalisierter linearer Modelle. Nachstehendes Schema verdeutlicht nochmals die einzelnen Schritte dieses iterativen Algorithmus:

Step 1: Initialisiere $\hat{\boldsymbol{\beta}}_{(0)} := \big(\hat{\boldsymbol{\beta}}_{(0),1}', \hat{\boldsymbol{\beta}}_{(0),2}'\big)'$ und $t := 0$.

Step 2: Berechne $\hat{\boldsymbol{\beta}}_{(t+1)}$:

Step 2.1: Berechne $\hat{\boldsymbol{\beta}}_{(t+1),1}$ (entspricht Stufe 1):
Fixiere $(\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_p') := \hat{\boldsymbol{\beta}}_{(t),2}'$. Initialisiere $\hat{\boldsymbol{\beta}}_1^{(0)} := \hat{\boldsymbol{\beta}}_{(t),1}$, berechne Folgeiterierte $\hat{\boldsymbol{\beta}}_1^{(k+1)}$, $k = 0, 1, \ldots$, via (7.11) bis Konvergenz. Setze $\hat{\boldsymbol{\beta}}_{(t+1),1} := \hat{\boldsymbol{\beta}}_1^{(k+1)}$.

Step 2.2: Berechne $\hat{\boldsymbol{\beta}}_{(t+1),2}$ (entspricht Stufe 2):
Fixiere $(\boldsymbol{\gamma}_0', \tilde{\boldsymbol{\alpha}}_1', \ldots, \tilde{\boldsymbol{\alpha}}_p') := \hat{\boldsymbol{\beta}}_{(t+1),1}'$. Initialisiere $\hat{\boldsymbol{\beta}}_2^{(0)} := \hat{\boldsymbol{\beta}}_{(t),2}$ und berechne Folgeiterierte $\hat{\boldsymbol{\beta}}_2^{(k+1)}$, $k = 0, 1, \ldots$, via (7.14) bis Konvergenz. Setze $\hat{\boldsymbol{\beta}}_{(t+1),2} := \hat{\boldsymbol{\beta}}_2^{(k+1)}$.

Step 2.3: Setze $\hat{\boldsymbol{\beta}}_{(t+1)} := \big(\hat{\boldsymbol{\beta}}_{(t+1),1}', \hat{\boldsymbol{\beta}}_{(t+1),2}'\big)'$ und berechne die relative Änderung $\Delta := \big|\big|\hat{\boldsymbol{\beta}}_{(t+1)} - \hat{\boldsymbol{\beta}}_{(t)}\big|\big| \,/\, \big|\big|\hat{\boldsymbol{\beta}}_{(t)}\big|\big|$. Falls $\Delta \leq \varepsilon$ für ein $\varepsilon > 0$, beende Algorithmus. Andernfalls setze $t := t+1$ und wiederhole *Step 2.*

Konvergenz des beschriebenen Algorithmus, d.h. das Erreichen des Abbruchkriteriums in *Step 2.3*, setzt insbesondere die Identifizierbarkeit der Parameter im Gesamtmodell voraus. Nun ist aber

$$\alpha_{jrs} \;=\; \alpha_{jr}\tilde{\alpha}_{js} \;=\; \alpha^*_{jr}\tilde{\alpha}^*_{js}\,,$$

für $\alpha^*_{jr} = c \cdot \alpha_{jr}$, $\tilde{\alpha}^*_{js} = \tilde{\alpha}_{js}/c$ und ein $c \neq 0$, so daß zwischen den Komponenten der faktoriellen Zerlegung Konstanten multiplikativ ausgetauscht werden können. Das damit erwachsende Identifikationsproblem wird auch durch die Restriktionen (7.12) nicht verhindert, da ebenso $\tilde{\boldsymbol{\alpha}}_j^{*\prime}\mathbf{1}_P = 0$, $j = 1,\ldots,p$, gilt. Für die kategorienspezifischen Faktoren sind somit zwar keine Restriktionen zur Schätzung in Stufe 2 erforderlich, sehr wohl aber zur Gewährleistung der Identifizierbarkeit im Gesamtmodell. Da Restriktionen des Typs (7.12) nicht den gewünschten Effekt erzielen, betrachtet man alternativ

$$\alpha_{j1} \;=\; 1\,,\quad j = 1,\ldots,p. \tag{7.15}$$

Für die damit notwendigen Modifikationen in Stufe 2 definiere die reduzierten Vektoren $\boldsymbol{\alpha}_{j-} := (\alpha_{j2},\ldots,\alpha_{jq})'$, $j = 1,\ldots,p$, und $\boldsymbol{\beta}_{2-} := (\boldsymbol{\alpha}'_{1-},\ldots,\boldsymbol{\alpha}'_{p-})'$. Mit diesen Deklarationen läßt sich (7.13) auch schreiben als

$$\boldsymbol{\eta}_i \;=\; \boldsymbol{\gamma}_0 \;+\; \sum_{j=1}^{p} \boldsymbol{c}'_{ij}\tilde{\boldsymbol{\alpha}}_j \cdot \boldsymbol{e}_{1,q} \;+\; \tilde{Z}_{2i}\boldsymbol{\beta}_{2-},$$

mit bekanntem Offset $\boldsymbol{\gamma}_0 + \sum_j \boldsymbol{c}'_{ij}\tilde{\boldsymbol{\alpha}}_j \cdot \boldsymbol{e}_{1,q}$ und modifizierter Designmatrix

$$\tilde{Z}_{2i} \;=\; \big[\,(\boldsymbol{c}'_{i1}\tilde{\boldsymbol{\alpha}}_1)\tilde{I}_{q-1} \mid \ldots \mid (\boldsymbol{c}'_{ip}\tilde{\boldsymbol{\alpha}}_p)\tilde{I}_{q-1}\,\big],$$

wobei $\tilde{I}_{q-1} := \big[\,\mathbf{0}_{q-1} \mid I_{q-1}\,\big]'$. Zur Berücksichtigung der Restriktionen im Penalty betrachte hier nur den Fall erster Differenzen

$$D^1_q\boldsymbol{\alpha}_j \;=\; D^1_q\begin{bmatrix}1\\ \boldsymbol{\alpha}_{j-}\end{bmatrix} \;=\; \begin{bmatrix}\alpha_{j2}-1\\ D^1_{q-1}\boldsymbol{\alpha}_{j-}\end{bmatrix} \;=\; \begin{bmatrix}\boldsymbol{e}'_{1,q-1}\boldsymbol{\alpha}_{j-}-1\\ D^1_{q-1}\boldsymbol{\alpha}_{j-}\end{bmatrix} \;=\; \tilde{D}^1_q\boldsymbol{\alpha}_{j-} - \boldsymbol{e}_{1,q-1},$$

für $\tilde{D}^1_q := \big[\,\boldsymbol{e}_{1,q-1} \mid (D^1_{q-1})'\,\big]'$. Damit ist

$$\begin{aligned}
\boldsymbol{\alpha}'_j K^{(j)}_q \boldsymbol{\alpha}_j \;&=\; \delta_j\boldsymbol{\alpha}'_j(D^1_q)'D^1_q\boldsymbol{\alpha}_j \;=\; \delta_j\big(\boldsymbol{\alpha}'_{j-}(\tilde{D}^1_q)' - \boldsymbol{e}'_{1,q-1}\big)\big(\tilde{D}^1_q\boldsymbol{\alpha}_{j-} - \boldsymbol{e}_{1,q-1}\big)\\
&=\; \delta_j\boldsymbol{\alpha}'_{j-}(\tilde{D}^1_q)'\tilde{D}^1_q\boldsymbol{\alpha}_{j-} \;-\; 2\,\delta_j\boldsymbol{e}'_{1,q-1}\tilde{D}^1_q\boldsymbol{\alpha}_{j-} \;+\; \delta_j\boldsymbol{e}'_{1,q-1}\boldsymbol{e}_{1,q-1}\\
&=\; \boldsymbol{\alpha}'_{j-}\tilde{K}^{(j)}_q\boldsymbol{\alpha}_{j-} \;-\; 2\,\delta_j\boldsymbol{e}'_{1,q-1}\boldsymbol{\alpha}_{j-} \;+\; \delta_j\,,
\end{aligned}$$

wobei $\tilde{K}_q^{(j)} = \delta_j(\tilde{D}_q^1)'\tilde{D}_q^1$, $j = 1, \ldots, p$. Als penalisierte Log–Likelihood erhält man

$$pl(\boldsymbol{\beta}_{2\_}) \;=\; l(\boldsymbol{\beta}_{2\_}) \;-\; \frac{1}{2}\cdot\left\{\boldsymbol{\beta}_{2\_}'\,\tilde{K}_2\,\boldsymbol{\beta}_{2\_} \;-\; 2\cdot(\boldsymbol{\delta}\otimes\boldsymbol{e}_{1,q-1})'\boldsymbol{\beta}_{2\_} \;+\; \boldsymbol{\delta}'\mathbf{1}_p\right\},$$

mit $\tilde{K}_2 = \mathrm{Diag}\big(\tilde{K}_q^{(1)}, \ldots, \tilde{K}_q^{(p)}\big)$ sowie $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)'$. Differenzieren liefert die penalisierte Score–Funktion

$$ps(\boldsymbol{\beta}_{2\_}) \;=\; \partial pl(\boldsymbol{\beta}_{2\_})/\partial\boldsymbol{\beta}_{2\_} \;=\; \sum_{i=1}^{N}\tilde{Z}_{2i}'D_i\Sigma_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) \;-\; \tilde{K}_2\boldsymbol{\beta}_{2\_} \;+\; \boldsymbol{\delta}\otimes\boldsymbol{e}_{1,q-1}.$$

Das Fisher–Scoring zum Lösen von $ps(\boldsymbol{\beta}_{2\_}) = \mathbf{0}$ hat die Form

$$\hat{\boldsymbol{\beta}}_{2\_}^{(k+1)} = \;\hat{\boldsymbol{\beta}}_{2\_}^{(k)} \;+\; \tilde{F}\Big(\hat{\boldsymbol{\beta}}_{2\_}^{(k)}\Big)^{-1} ps\Big(\hat{\boldsymbol{\beta}}_{2\_}^{(k)}\Big),\;\; k = 0, 1, 2, \ldots$$

mit der Pseudo–Fisher–Matrix

$$\tilde{F}(\boldsymbol{\beta}_{2\_}) \;=\; \sum_{i=1}^{N}\tilde{Z}_{2i}'\,D_i\Sigma_i^{-1}D_i'\,\tilde{Z}_{2i} \;+\; \tilde{K}_2.$$

Aus den Restriktionen (7.15) resultiert als Modifikation für Stufe 1 lediglich die Ersetzung von $\boldsymbol{\alpha}_j$ durch $(1, \boldsymbol{\alpha}_{j\_}')'$, $j = 1, \ldots, p$.

7.3.1 Simulation

Zur Demonstration des vorgestellten zweistufigen Schätzverfahrens wird ein kumulatives Logit–Modell mit den Prädiktoren

$$\begin{aligned}
\eta_{i1} &= -1.5 \;+\; 1.0\cdot\sin(x_i)\\
\eta_{i2} &= -0.8 \;+\; 0.7\cdot\sin(x_i)\\
\eta_{i3} &= -0.4 \;+\; 0.4\cdot\sin(x_i)\\
\eta_{i4} &= \;\;0.2 \;+\; 0.1\cdot\sin(x_i)
\end{aligned}$$

und Beobachtungen x_i, $i = 1, \ldots, 200$, aus dem Intervall $[0, 2\pi]$ betrachtet. Die gewählte Spezifikation repräsentiert exemplarisch den Fall nur einer Kovariablen in der Modellklasse (7.8). Als Basiseffekt der Einflußgröße wird die Sinusfunktion zugrunde gelegt, deren Amplitude durch abnehmende kategorienspezifische Faktoren mehr und mehr gedämpft wird. Wie der linken Darstellung in Abbildung 7.6 zu entnehmen ist, genügen die Prädiktoren im Intervall $[0, 2\pi]$ der erforderlichen Ordnungsrelation $\eta_{i1} \leq \ldots \leq \eta_{i4}$.

Für die Simulation wurden 100 Stichproben, bestehend aus jeweils 200 unabhängigen, multinomialverteilten Zufallsgrößen gezogen, deren Auftretenswahrscheinlichkeiten via (6.5) aus der angesetzten Prädiktorspezifikation resultieren. Zu jeder der so generierten 100 Responsesituationen wurde Modell (7.8) für $j = 1$ sowie $x_{i1} := x_i$, $i = 1, \ldots, 200$, unter Verwendung des vorgestellten zweistufigen Verfahrens geschätzt. Die Optimierung der beiden involvierten Glättungsparameter λ_1 und δ_1 erfolgte jeweils über den genetischen Algorithmus aus Kapitel 4. Auf die Penalisierung der Schwellenwerte wurde verzichtet ($\delta_0 = 0$).

In Tabelle 7.3 sind die Mittelwerte und empirischen Standardabweichungen der 100 AIC–optimalen Schätzungen für die Schwellenwerte und kategorienspezifischen Faktoren aufgelistet. Alle wahren Modellparameter liegen innerhalb der einfachen Schwankungsintervalle $\bar{x} \pm s$. Die größere Verzerrung der geschätzten kategorienspezifischen Faktoren ist auf deren vertikale Penalisierung zurückzuführen.

	$\overline{\hat{\gamma}_{0r}}$	$s(\hat{\gamma}_{0r})$	$\overline{\hat{\alpha}_{1r}}$	$s(\hat{\alpha}_{1r})$
Kategorie 1	-1.516	0.211	1.000	0.000
Kategorie 2	-0.800	0.167	0.763	0.151
Kategorie 3	-0.388	0.148	0.492	0.172
Kategorie 4	0.223	0.141	0.255	0.213

TABELLE 7.3: *Gemittelte Schätzungen der Schwellenwerte und der kategorienspezifischen Faktoren zusammen mit den zugehörigen empirischen Standardabweichungen.*

In Abbildung 7.6 sind der Mittelwert der 100 AIC–optimalen Schätzungen für den Basiseffekt der Kovariablen sowie die zugrunde liegende Sinusfunktion dargestellt. Darüber hinaus zeigt die Abbildung die aus den optimalen Effektschätzungen bestimmten, punktweise definierten, empirischen 5% und 95% Quantilbänder.

Aus der rechten Darstellung ist ersichtlich, daß der mittlere, geschätzte Basiseffekt den wahren Zusammenhang weitgehend formgetreu nachbildet. Ab-

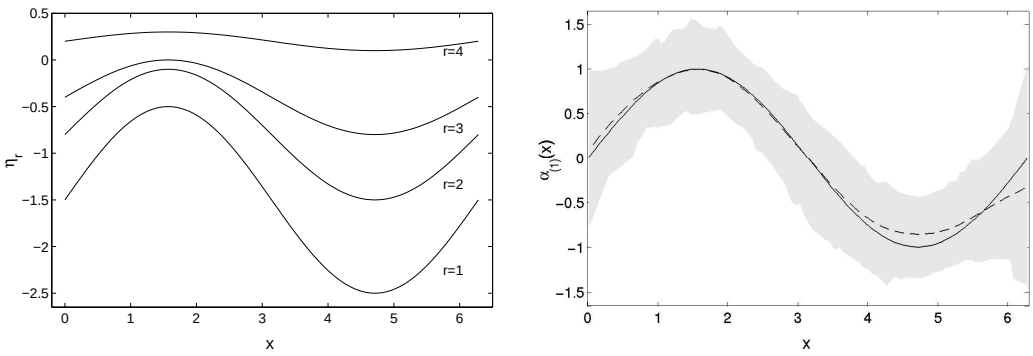

ABBILDUNG 7.6: *Links: Zugrunde liegende Prädiktorspezifikation in den Kategorien. Rechts: Wahrer (——) und mittlerer geschätzter (– – –) Basiseffekt der Kovariablen. Grau unterlegte Fläche kennzeichnet Bereich zwischen punktweisen empirischen* 5% *und* 95% *Quantilen.*

weichungen sind nur für große Kovariablenwerte zu beobachten. Ungeachtet dieser Verzerrungen ist die zugrunde liegende Sinusfunktion in dem von den empirischen Quantilbändern begrenzten Bereich enthalten.

Zusammenfassung und Ausblick

Semiparametrische Regressionsmodelle erweisen sich in Situationen, in denen die Kovariablenmenge neben diskreten auch stetige Merkmale umfasst als flexibles Instrumentarium bei der Analyse komplexer Abhängigkeitsstrukturen. Während diese Modelle im univariaten Fall einer – durch zahlreiche Publikationen belegten – erschöpfenden Betrachtung unterzogen wurden, existieren bis dato nur wenige Arbeiten, die kategoriale Responsevariablen fokussieren. In der vorliegenden Dissertationsschrift wurden bewährte parametrische Regressionsmodelle für kategorial–nominale und –ordinale abhängige Variablen um verschiedene nonparametrische Komponenten ergänzt.

Konzeptionelle Grundlage für die Modellierung dieser Komponenten bildete die Approximierbarkeit hinreichend glatter, funktionaler Kovariableneffekte durch sogenannte Polynom–Splines. Die Existenz von Basisrepräsentationen dieser stückweise polynomialen Funktionen gewährleistete eine Rückführbarkeit des Prädiktors auf rein parametrische Strukturen und ermöglichte damit die Einbettung der betrachteten semiparametrischen Ansätze in den bekannten Rahmen (multivariater) generalisierter linearer Modelle. Die Schätzung der nonparametrischen Modellkomponenten erfolgt simultan und unter Vermeidung iterativer Backfitting-Schritte. Regulierende Mechanismen in Form diskreter Strafterme dienten der Variationskontrolle in den korrespondierenden Schätzungen und führten zu penalisierten Maximum–Likelihood–Prozeduren bei der Bestimmung der unbekannten Modellparameter.

Theoretische Betrachtungen im Kapitel 2 qualifizierten B–Splines und Differenzenpenalties als die ideale Paarung aus Basisdarstellung eines Polynom-Splines und diskretem Strafterm. Eine Diskussion von Einsatzmöglichkeiten dieser als P–Splines bezeichneten Kombination (Eilers & Marx, 1996) bei der Berücksichtigung nonparametrischer Komponenten erbrachte neben grundlegenden Modellierungsaspekten einen Lösungsansatz für auftretende Identifikations– und Singularitätsprobleme. Über die Behandlung von Haupteffekten hinaus, wurden Kovariableninteraktionen einer detaillierten Betrachtung unterzogen.

Unter Verwendung des propagierten P–Spline Ansatzes gelang eine Erweite-

rung des multinomialen Logit–Modells für nominalen Response auf semiparametrische Strukturen. Dabei wurde explizit zwischen globalen Einflußgrößen und kategorienspezifischen Charakteristiken, deren Werte individuenspezifisch und charakteristisch für die einzelnen Responsekategorien sind, unterschieden. Die zugrunde liegende Prädiktorstruktur ist insoweit als allgemeingültig zu bezeichnen, als das Kovariablen beider Typen linear als auch unspezifiziert funktional ins Modell eingehen können. In einer Analyse zu den Auftretensformen von Sichelzellenanämie und in einer Simulationsstudie konnten die Stärken des P–Spline Ansatzes bei der flexiblen Modellierung von Kovariableneffekten im multinomialen Logit–Modell demonstriert werden.

Das zweifelsohne populärste Modell im Kontext ordinaler Regression ist das Proportional Odds Modell. Häufig stellt die Annahme ausschließlich globaler Kovariablengewichte jedoch eine ungerechtfertigte Vereinfachung dar, die zu falschen Schlüssen bei der Interpretation von Ergebnissen führen kann. Gängige iterative Schätzverfahren, wie das Fisher–Scoring, schlagen für das angemessenere Partial Proportional Odds Modell hingegen oft fehl. Durch den Übergang zur penalisierten Log–Likelihood, die eine kategorienübergreifende Parameterbestrafung durch Differenzenpenalties erster Ordnung einschließt, konnten numerische Schwierigkeiten bei der Schätzung nicht-globaler Effekte weitgehend umgangen werden. In numerisch kritischen Fällen wiesen die penalisierten Schätzungen geringere Verlustfunktionen als die korrespondierenden Schätzungen im Proportional Odds Modell auf. Neben einer verbesserten Schätzgüte ermöglichte das Penalisierungsprinzip die Betrachtung von Teststatistiken, die eine Verfügbarkeit kategorienspezifischer Parameterschätzungen explizit voraussetzen. Die qualitative Bewertung von Tests auf das Vorliegen identischer Chancenverhältnisse erbrachte im Non-Proportional Odds Modell annähernd deckungsgleiche Gütefunktionen für den klassischen (unpenalisierten) Score-Test und die penalisierten Formen des Likelihood–Quotienten- und des Wald-Tests. Anhand von Daten zur diabetischen Retinopathie wurde das Verhalten penalisierter Teststatistiken bei der Identifizierung proportionaler Chancen untersucht. Die analysierten significance traces liessen nur geringe Abhängigkeiten der geschätzten p–Werte von der gewählten Penalisierungsstärke erkennen. Damit hatte die konkrete Belegung der Glättungsparameter keinen Einfluß auf die Testentscheidung und konnte sich an

rein numerischen Notwendigkeiten orientieren. Inwieweit dieser Aussage Allgemeingültigkeit zukommt, wurde im Rahmen der vorliegenden Arbeit nicht geklärt.

Die mit der Penalisierung kategorienspezifischer Gewichte auferlegte Variationskontrolle läßt sich als Umsetzung kategorienübergreifender Glattheitsanforderungen interpretieren. Starke Differenzenpenalties erzwingen polynomiale Abhängigkeiten, deren Grad von der zugrunde liegenden Differenzenordnung bestimmt wird. Darauf basierend wurde gezeigt, daß vereinfachte kategorienspezifische Ansätze aus Situationen resultieren, in denen eine Optimierung des Akaike–Informations–Kriteriums (AIC) zu maximalen Glättungsparametern führt.

In einem weiteren Schritt wurde die Penalisierung über Responsekategorien hinweg als fester Bestandteil der Schätzung in ordinalen Modellen mit kategorienspezifischen Effekten integriert. Der in diesem Zusammenhang geprägte Begriff der vertikalen Penalisierung diente der Abgrenzung zur (horizontalen) Bestrafung von Basiskoeffizienten im Kontext nonparametrischer Modellierung. Glättungsparameter vertikaler Penalties wurden nicht mehr unter ausschließlich stabilitätsfördernden Gesichtspunkten festgelegt, sondern auf Basis eines geeigneten Kriteriums optimiert. Die Kombination von horizontaler und vertikaler Penalisierung im Datenbeispiel zur Retinopathie lieferte ein semiparametrisches Modell mit kategorienspezifischen Effekten, die über die Responsekategorien glatt variieren.

Globale funktionale Effekte, die in den Responsekategorien multiplikativ modifiziert werden, stellen eine parameterökonomische Alternative zur kategorienspezifischen Spline–Approximation nonparametrischer Komponenten dar. Für kumulative Modelle mit multiplikativen Effektverknüpfungen wurde ein zweistufiges Verfahren zur separaten Schätzung von globalen Basiskoeffizienten und modifizierenden, kategorienspezifischen Faktoren entwickelt. In einer Simulationsstudie konnte das Potential dieses Ansatzes verdeutlicht werden.

Die Bestimmung der Glättungsparameter in den Anwendungen erfolgte, sofern nicht von numerischen Erfordernissen dominiert, stets auf der Basis geeigneter Zielkriterien. Deren Optimierung wurde unter Zuhilfenahme eines

eigens dafür adaptierten genetischen Algorithmus vollzogen, der ein praktischen Ansprüchen genügendes Auswahl- und Konvergenzverhalten zeigte.

Im Rahmen der numerischen Auswertungen sind einige Softwarekomponenten entstanden, die interessierten Anwendern im Internet als Download zur Verfügung stehen. Ein spezieller Abschnitt im Anhang dieser Arbeit widmet sich detaillierten Erläuterungen zur Nutzung dieser Komponenten.

Fragen der Existenz und Eindeutigkeit von Maximum–Likelihood–Schätzern in generalisierten linearen Modellen waren Gegenstand zahlreicher Publikationen der jüngeren Vergangenheit (vgl. Wedderburn, 1976 bzw. Kaufmann, 1988 für kategorialen Response). Ebenso ausführlich untersucht wurden verschiedene asymptotische Eigenschaften dieser Schätzer (s. Fahrmeir & Kaufmann, 1985) und korrespondierende Hypothesentests (Fahrmeir, 1987). Für die penalisierte Maximum–Likelihood–Schätzung existieren bis dato nur erste Ansätze zur Behandlung der genannten Problemstellungen (Wand, 1999; Aerts, Claeskens & Wand, 2002). Die entsprechend hohe Anzahl ungeklärter Fragen bietet ein weites Betätigungsfeld für zukünftige Forschungen. Insbesondere zur Signifikanzbeurteilung der vorgestellten, penalisierten Test–Statistiken wären Erkenntnisse über deren (asymptotische) Verteilung von großem Wert.

Als potentielle Erweiterung der universellen Prädiktorstruktur in Kapitel 3 käme die Modellierung korrelierter Daten in Betracht. In Tutz (2003a) wird Korrelationen in ordinalen Regressionsmodellen über die Einbeziehung von Random Effects Rechnung getragen. Der Modellierung stetiger Einflußgrößen liegt der auch hier verwendete P–Spline Ansatz zugrunde. Die betrachteten Kovariableneffekte sind jedoch ausschließlich globaler Natur. Eine denkbare Erweiterung des Modells um kategorienspezifische Komponenten würde den Rahmen dieser Arbeit sprengen, da die Maximierung der zugehörigen penalisierten marginalen Likelihood mit erheblichem Aufwand verbunden ist.

Die Glättungsparameterbestimmung anhand geeigneter Zielkriterien bildete in allen durchgeführten Auswertungen den zeitintensivsten Teil. Auch wenn sich die Optimierungszeit mit effizienteren Algorithmen verkürzen ließe, bliebe der Zeitaufwand im Vergleich zur eigentlichen Schätzung unverhältnismä-

ßig hoch. Die Reformulierung penalisierter Basisfunktionsansätze als Mixed Models (Wand, 2003; Ruppert, Wand & Carroll, 2003) sowie bayesianische P-Splines (Lang & Brezger, 2003; Brezger & Lang, 2003) stellen alternative Ansätze dar, die aufwendiges Optimieren durch simultanes Mitschätzen von Glättungsparametern vermeiden. Da diese Alternativen auf der auch in der vorliegenden Arbeit propagierten Spline–Approximation nonparametrischer Komponenten beruhen, läge ein Vergleich der Methoden nahe. Die Darstellungen penalisierter Basisfunktionsansätze als Mixed Model wurden jedoch nur für Modelle mit univariatem Response untersucht. Entsprechende Aussagen für den hier relevanten mehrkategorialen Fall fehlen bis dato. Auf einen Vergleich mit bayesianischen P–Splines wurde aus Zeitgründen verzichtet.

Der stetige Charakter einer Einflußgröße bildete in allen angestellten Überlegungen das ausschlaggebende Kriterium für deren nonparametrische Modellierung. Parametrische Effekte stetiger Kovariablen sind zwar über Polynom–Splines ebenso approximierbar, wurden für die explizite Modellierung jedoch als zu restriktive Annahme ausgeschlossen. Eine Rechtfertigung dieser Annahme ließe sich nur mit entsprechenden Tests ableiten. Crainiceanu, Ruppert, Claeskens & Wand (2003) testen polynomiale Regressionsmodelle gegen nonparametrische Alternativen, letztere modelliert über die Ridge–penalisierte Truncated–Power-Basis. Die Nullhypothese polynomialer Strukturen wird dabei als Bedingung an die Koeffizienten der Basisdarstellung formuliert. Mit den Beziehungen aus Abschnitt 2.3 ließe sich dieser Test – unter Ausnutzung der damit verbundenen (numerischen) Vorteile – äquivalent in den Koeffizienten der zugehörigen B–Spline-Basis ausdrücken. Die praktische Umsetzung dieser Idee konnte jedoch noch nicht realisiert werden.

A Diverses

Im Anhang dieser Arbeit sind einige, im Text ausgelassene oder nur angedeutete Beweise und Erläuterungen zusammengefaßt.

Beweis von Proposition 2.2. (vollständige Induktion über k)

Induktionsanfang: Sei $j \in \mathbb{Z}$ beliebig. Für $k = 0$ ist

$$\Delta^0 f(z_j) \;=\; f(z_j) \;=\; (-1)^0 \binom{0}{0} f(z_j) \;=\; \sum_{l=0}^{0} (-1)^{0+l} \binom{0}{l} f(z_{j+l})$$

Induktionsvoraussetzung: Für $j \in \mathbb{Z}$ beliebig und $k \geq 0$ gelte

$$\Delta^k f(z_j) \;=\; \sum_{l=0}^{k} (-1)^{k+l} \binom{k}{l} f(z_{j+l})$$

Induktionsbehauptung: Dann gilt auch

$$\Delta^{k+1} f(z_j) \;=\; \sum_{l=0}^{k+1} (-1)^{k+l+1} \binom{k+1}{l} f(z_{j+l})$$

Induktionsschritt: $(k \to k+1)$

$$\Delta^{k+1} f(z_j) \;=\; \Delta^k f(z_{j+1}) \;-\; \Delta^k f(z_j)$$

$$\overset{I.V.}{=} \; \sum_{l=0}^{k} (-1)^{k+l} \binom{k}{l} \cdot f(z_{j+l+1}) \;-\; \sum_{l=0}^{k} (-1)^{k+l} \binom{k}{l} \cdot f(z_{j+l})$$

$$= \; \sum_{l=1}^{k+1} (-1)^{k+l+1} \binom{k}{l-1} \cdot f(z_{j+l}) \;+\; \sum_{l=0}^{k} (-1)^{k+l+1} \binom{k}{l} \cdot f(z_{j+l})$$

$$= \; f(z_{j+k+1}) \;+\; (-1)^{k+1} f(z_j) \;+\; \sum_{l=1}^{k} (-1)^{k+l+1} \left\{ \binom{k}{l-1} + \binom{k}{l} \right\} f(z_{j+l})$$

$$= \; (-1)^{2(k+1)} \binom{k+1}{k+1} \cdot f(z_{j+k+1}) \;+\; (-1)^{k+1} \binom{k+1}{0} \cdot f(z_j)$$

$$+ \sum_{l=1}^{k} (-1)^{k+l+1} \binom{k+1}{l} f(z_{j+l}) \;=\; \sum_{l=0}^{k+1} (-1)^{k+l+1} \binom{k+1}{l} f(z_{j+l})$$

□

Beweis von Gleichung (2.11).

Eine Anwendung der Leibnizschen Regel auf die Funktion

$$q_l(x,t) = (x-t)_+^l = (x-t)\cdot(x-t)_+^{l-1} = (x-t)\cdot q_{l-1}(x,t)$$

liefert für $z_i := t_i$, $i = j,\dots,j+k$ und $k := l+1$

$$\begin{aligned}[t_{j+l+1}\dots t_j]q_l(x,.) &= \sum_{i=j}^{j+l+1}\big([t_i\dots t_j](x-.)\big)\cdot\big([t_{j+l+1}\dots t_i]q_{l-1}(x,.)\big)\\ &= (x-t_j)\cdot[t_{j+l+1}\dots t_j]q_{l-1}(x,.) -\\ &\qquad -[t_{j+l+1}\dots t_{j+1}]q_{l-1}(x,.), \quad \text{(A.1)}\end{aligned}$$

da $[t_{i+1}\,t_i](x-.) = -1$ für $i = j,\dots,j+l$ und damit $[t_i\dots t_j](x-.) = 0$ für $i \geq j+2$. Ferner ist mit (2.2)

$$\begin{aligned}&[t_{j+l+1}\dots t_j]q_{l-1}(x,.) \;=\\ &= \frac{1}{t_{j+l+1}-t_j}\big([t_{j+l+1}\dots t_{j+1}]q_{l-1}(x,.) - [t_{j+l}\dots t_j]q_{l-1}(x,.)\big).\end{aligned}$$

Einsetzen in (A.1) liefert die Behauptung. □

Beweis von Proposition 2.4. (vollständige Induktion über l)

Induktionsanfang: Sei $j \in \mathbb{Z}$ beliebig. Für $l = 0$ ist

$$(t_{j+1}-t_j)[t_{j+1}\,t_j]\tilde{q}_0(.,x) \overset{(2.5)}{=} B_{0j}(x) = I_{[t_j,t_{j+1})}(x) = -(t_{j+1}-t_j)[t_{j+1}\,t_j]q_0(x,.)$$

Induktionsvoraussetzung: Für $j \in \mathbb{Z}$ beliebig und $l \geq 0$ gelte

$$[t_{j+l+1}\dots t_j]\tilde{q}_l(.,x) = (-1)^{l+1}\cdot[t_{j+l+1}\dots t_j]q_l(x,.)$$

Induktionsbehauptung: Dann gilt auch

$$[t_{j+l+2}\dots t_j]\tilde{q}_{l+1}(.,x) = (-1)^{l+2}\cdot[t_{j+l+2}\dots t_j]q_{l+1}(x,.)$$

Induktionsschritt: $(l \to l+1)$

$$\begin{aligned}&[t_{j+l+2}\dots t_j]\tilde{q}_{l+1}(.,x) \;=\\ &\overset{(2.10)}{=} \frac{t_{j+l+2}-x}{t_{j+l+2}-t_j}\cdot[t_{j+l+2}\dots t_{j+1}]\tilde{q}_l(.,x) - \frac{t_j-x}{t_{j+l+2}-t_j}\cdot[t_{j+l+1}\dots t_j]\tilde{q}_l(.,x)\\ &\underset{(2.11)}{\overset{I.V.}{=}} (-1)^{l+1}\cdot\big(-[t_{j+l+2}\dots t_j]q_{l+1}(x,.)\big) = (-1)^{l+2}\cdot[t_{j+l+2}\dots t_j]q_{l+1}(x,.)\end{aligned}$$

□

Beweis von Proposition 2.7.

Es bezeichne $\hat{\boldsymbol{\beta}}$ den für $d = n+1$ und ein $\lambda_{\boldsymbol{\beta}} \geq 0$ aus (2.22) hervorgehenden Maximum–Likelihood–Schätzer. Der korrespondierende Fit

$$\hat{f}_B(x) = \sum_{m=-n}^{M-1} \hat{\beta}_m B_{nm}(x)\,, \quad x \in [a,b]\,,$$

ist ein Element des Raumes $S_n(\Omega_M)$ und als solches darstellbar in der zugehörigen Truncated–Power-Basis. Die entsprechenden Basiskoeffizienten $\tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}})$ und $\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}})$ lassen sich via (2.13) und (2.14) aus $\hat{\boldsymbol{\beta}}$ berechnen. Da $\hat{\boldsymbol{\beta}}$ und $\tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}})$, $\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}})$ im selben Fit resultieren, folgt die Gleichheit der Log–Likelihoods

$$l(\hat{\boldsymbol{\beta}}) \;=\; l\big(\tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}), \boldsymbol{\alpha}(\hat{\boldsymbol{\beta}})\big).$$

Ferner gilt mit (2.14), (2.16), (2.19) und $\lambda_{\boldsymbol{\alpha}} := (n!h^n)^2\,\lambda_{\boldsymbol{\beta}}$

$$\begin{aligned}\lambda_{\boldsymbol{\beta}} P_\Delta^{n+1}(\hat{\boldsymbol{\beta}}) \;&=\; \lambda_{\boldsymbol{\alpha}} (n!h^n)^{-2} P_\Delta^{n+1}(\hat{\boldsymbol{\beta}}) \overset{(2.19)}{=} \lambda_{\boldsymbol{\alpha}} \sum_{m=1}^{M-1} \Big((n!h^n)^{-1} \Delta^{n+1} \hat{\beta}_{m-n-1}\Big)^2 \\ &\underset{(2.16)}{\overset{(2.14)}{=}} \lambda_{\boldsymbol{\alpha}} P_{ridge}\big(\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}})\big),\end{aligned}$$

also $pl(\hat{\boldsymbol{\beta}}) = pl\big(\tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}), \boldsymbol{\alpha}(\hat{\boldsymbol{\beta}})\big)$.

Man betrachte jetzt den für oben definiertes $\lambda_{\boldsymbol{\alpha}}$ aus (2.17) resultierenden Maximum–Likelihood–Schätzer $\big(\widehat{\tilde{\boldsymbol{\alpha}}}', \hat{\boldsymbol{\alpha}}'\big)'$ mit zugehörigem Fit

$$\hat{f}_T(x) = \sum_{r=0}^{n} \widehat{\tilde{\alpha}}_r p_r(x) \;+\; \sum_{m=1}^{M-1} \hat{\alpha}_m q_{nm}(x)\,, \quad x \in [a,b]\,. \tag{A.2}$$

Unter den angegebenen Voraussetzungen ist die Äquivalenz der Schätzansätze (2.17) und (2.22) gleichbedeutend mit $\hat{f}_B(x) = \hat{f}_T(x)$ für alle $x \in [a,b]$, wobei sich diese Identität aus

$$\widehat{\tilde{\boldsymbol{\alpha}}} = \tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}) \;\;\wedge\;\; \hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}) \tag{A.3}$$

folgern ließe. Wir beweisen (A.3), indem wir die Annahme

$$\widehat{\tilde{\boldsymbol{\alpha}}} \neq \tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}) \;\;\vee\;\; \hat{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}) \tag{A.4}$$

zu einem Widerspruch führen. Setzt man die Existenz einer eindeutigen Lösung von (2.17) voraus, so folgt aus (A.4)

$$pl(\hat{\boldsymbol{\beta}}) \;=\; pl\big(\tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}), \boldsymbol{\alpha}(\hat{\boldsymbol{\beta}})\big) \;<\; pl\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big), \tag{A.5}$$

weil $\left(\widehat{\tilde{\boldsymbol{\alpha}}}', \hat{\boldsymbol{\alpha}}'\right)'$ (2.17) maximiert.

Da die Schätzung $\hat{f}_T$ in (A.2) ein Element aus $S_n(\Omega_M)$ ist, muß eine Darstellung in der zugehörigen B–Spline-Basis mit Koeffizienten $\boldsymbol{\beta}\left(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\right)$ existieren. Für diese Darstellung resultiert derselbe Fit, und damit ist

$$l(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}) \;=\; l\Big(\boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big)\Big). \tag{A.6}$$

Obgleich die Berechnungsvorschrift für den Vektor $\boldsymbol{\beta}\left(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\right)$ nicht explizit angegeben wird, und damit die Gestalt seiner Komponenten unbekannt ist, seine Existenz ist nichtsdestotrotz durch die Existenz von $\hat{f}_T$ gesichert.

Ausgehend von dieser Existenzaussage müssen zu $\boldsymbol{\beta}\left(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\right)$ wiederum Koeffizienten

$$\tilde{\boldsymbol{\alpha}}\Big(\boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big)\Big) \quad \text{und} \quad \boldsymbol{\alpha}\Big(\boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big)\Big)$$

existieren, die denselben Fit ergeben und sich via (2.13) und (2.14) berechnen lassen. Da Basisdarstellungen eindeutig sind, impliziert dies die Entsprechungen

$$\tilde{\boldsymbol{\alpha}}\Big(\boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big)\Big) = \widehat{\tilde{\boldsymbol{\alpha}}} \quad \text{und} \quad \boldsymbol{\alpha}\Big(\boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big)\Big) = \hat{\boldsymbol{\alpha}},$$

so daß die Aussagen (2.13) und (2.14) auch in den Komponenten der Vektoren $\left(\widehat{\tilde{\boldsymbol{\alpha}}}', \hat{\boldsymbol{\alpha}}'\right)'$ und $\boldsymbol{\beta}\left(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\right)$ gelten, und damit ist

$$\begin{aligned}\lambda_{\boldsymbol{\alpha}} P_{ridge}\big(\hat{\boldsymbol{\alpha}}\big) &\overset{(2.16)}{=} \lambda_{\boldsymbol{\beta}} \sum_{m=1}^{M-1} \big(n! h^n \, \hat{\alpha}_m\big)^2 \overset{(2.14)}{=} \lambda_{\boldsymbol{\beta}} \sum_{m=1}^{M-1} \big(\Delta^{n+1} \beta_{m-n-1}(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}})\big)^2 \\ &\overset{(2.19)}{=} \lambda_{\boldsymbol{\beta}} P_{\Delta}^{n+1}\big(\boldsymbol{\beta}(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}})\big).\end{aligned}$$

Mit (A.6) folgt $pl(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}) = pl\Big(\boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big)\Big)$. Berücksichtigt man dies in (A.5) erhält man schließlich

$$pl(\hat{\boldsymbol{\beta}}) \;<\; pl\Big(\boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big)\Big),$$

im Widerspruch dazu, daß $\hat{\boldsymbol{\beta}}$ der Maximum–Likelihood-Schätzer für das Optimierungsproblem (2.22) ist. Demnach muß (A.4) falsch sein bzw. (A.3) gelten. Ein umgekehrter, vom Maximum–Likelihood–Schätzer von (2.17) ausgehender Argumentationsweg liefert die Identität

$$\hat{\boldsymbol{\beta}} \;=\; \boldsymbol{\beta}\big(\widehat{\tilde{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}\big). \tag{A.7}$$

(A.3) und (A.7) implizieren jeweils die Äquivalenz von (2.17) und (2.22). □

Beweis von Proposition 2.8. (vollständige Induktion über d)

Induktionsanfang: Sei $l \geq 1$ beliebig. Für $d = 0$ ist

$$B_{l\nu}^{(0)}(x) = B_{l\nu}(x) = B_l(x, t_\nu) = (-h)^{-0} \cdot \Delta_2^0 \, B_l(x, t_\nu).$$

Für $d = 1$ ist die Richtigkeit von Proposition 2.8 durch (2.25) gegeben.

Induktionsvoraussetzung: Für $l \geq 1$ beliebig und $0 \leq d < l - 1$ gelte

$$B_{l\nu}^{(d)}(x) \;=\; (-h)^{-d} \cdot \Delta_2^d B_{l-d}(x, t_\nu)$$

Induktionsbehauptung: Dann gilt auch

$$B_{l\nu}^{(d+1)}(x) \;=\; (-h)^{-d-1} \cdot \Delta_2^{d+1} B_{l-d-1}(x, t_\nu)$$

Induktionsschritt: $(d \to d+1)$

$$\begin{aligned}
B_{l\nu}^{(d+1)}(x) \;&=\; \frac{\partial}{\partial x} B_{l\nu}^{(d)}(x) \;\overset{I.V.}{=}\; (-h)^{-d} \cdot \frac{\partial}{\partial x} \, \Delta_2^d \, B_{l-d}(x, t_\nu) \\
&\overset{Prop.\ 2.2}{=} (-h)^{-d} \cdot \sum_{k=0}^{d} (-1)^{d+k} \binom{d}{k} \cdot \frac{\partial}{\partial x} \, B_{l-d}(x, t_{\nu+k}) \\
&\overset{(2.24)}{=} (-h)^{-d} \cdot h^{-1} \cdot \sum_{k=0}^{d} (-1)^{d+k} \binom{d}{k} \cdot \left\{ B_{l-d-1}(x, t_{\nu+k}) - B_{l-d-1}(x, t_{\nu+k+1}) \right\} \\
&\overset{Prop.\ 2.2}{=} (-h)^{-d} \cdot (-h)^{-1} \cdot \left\{ \Delta_2^d B_{l-d-1}(x, t_{\nu+1}) - \Delta_2^d B_{l-d-1}(x, t_\nu) \right\} \\
&= \; (-h)^{-d-1} \cdot \Delta_2^{d+1} B_{l-d-1}(x, t_\nu)
\end{aligned}$$

□

B Software

Die numerische Auswertung der Datenbeispiele sowie sämtliche Simulationsstudien erfolgten unter Verwendung eigens dafür erstellter S–Plus Routinen. Im Sinne einer ökonomischen Einsetzbarkeit wurde dabei die folgende Modularisierungsstruktur zugrunde gelegt:

- Aufbau des Designs mit B–Spline Approximation der glatten Effekte
- Kodierung der Responsebeobachtungen
- Aufbau der horizontalen/vertikalen Differenzenpenalties
- Fisher–Scoring zur Maximierung der penalisierten Log–Likelihood

Jedes dieser Module wurde als separate S–Plus Funktion implementiert, deren Aufruf, Ein– und Ausgabeparameter nachstehend erläutert werden. Die Darstellungen beschränken sich dabei auf die Schätzung von Haupteffekten. Interaktionsterme können jedoch – wie in Abschnitt 7.1 demonstriert – mit den gegebenen Routinen ebenfalls berücksichtigt werden.

Im Vorfeld einer Anwendung sind die Kovariablenbeobachtungen gemäß

$$Data = \left[X_L \,\middle|\, X_{L,c} \,\middle|\, X_A \,\middle|\, X_{A,c} \,\middle|\, X_L^{(c)} \,\middle|\, X_{L,c}^{(c)} \,\middle|\, X_A^{(c)} \,\middle|\, X_{A,c}^{(c)} \right]$$

spaltenweise in einer Datenmatrix zu gruppieren, wobei in den Teilmatrizen

X_L die global und linear,	$X_{L,c}$ die kategorienspezifisch und linear,
X_A die global und glatt,	$X_{A,c}$ die kategorienspezifisch und glatt

zu modellierenden globalen Kovariablen zusammengefaßt werden, während

$X_L^{(c)}$ die global und linear,	$X_{L,c}^{(c)}$ die kategorienspezifisch und linear,
$X_A^{(c)}$ die global und glatt,	$X_{A,c}^{(c)}$ die kategorienspezifisch und glatt

modellierten kategorienspezifischen Charakteristiken subsumieren. Für einen konkreten Anwendungsfall reduziert sich die Datenmatrix auf die vorgesehenen Kombinationen von Einflußgrößentyp und Modellierungsform.

Für die getätigten Auswertungen zu den Datensätzen der Sichelzellenanämie (vgl. Abschnitt 5.3.2) und der Retinopathie (vgl. Abschnitte 6.4 bzw. 7.2.1)

sind die untersuchten Kovariablen demnach wie folgt anzuordnen

$$(1) \quad Data = \left[X_{L,c} \,\middle|\, X_{A,c} \right] = \left[\text{GENDER} \,\middle|\, \text{AGE}, \text{ESR} \right]$$

$$(2) \quad Data = \left[X_{L} \,\middle|\, X_{L,c} \right] = \left[\text{GH}, \text{BP} \,\middle|\, \text{SM}, \text{DD}, \text{DDQ} \right]$$

$$(3) \quad Data = \left[X_{L,c} \,\middle|\, X_{A} \,\middle|\, X_{A,c} \right] = \left[\text{SM} \,\middle|\, \text{GH}, \text{BP} \,\middle|\, \text{DD} \right]$$

Zum Aufbau der globalen Designmatrix verwendet man die S–Funktion

Design(x,q,num,model,link,anzint,bdeg)

mit den Eingabeparametern

x	Rearrangierte Datenmatrix *Data*
q	Anzahl der modellierten Responsekategorien ($q = k - 1$)
num	Vektor der Länge 8, dessen Komponenten die Anzahl der pro Kombination von Einflußgrößentyp und Modellierungsform vorhandenen Kovariablen angeben (Vektor mit Teilspaltenzahlen von *Data*)
model	Modellspezifizierer mit möglichen Belegungen *'nominal'*, *'cumulative'* und *'sequential'*
link	Linkspezifizierer mit möglichen Belegungen *'identity'*, *'minextreme'*, *'probit'* und *'logit'* (Default). Für nominales Modell nur *'logit'* möglich
anzint	Vektor der Länge $num[3] + num[4] + \big(num[7] + num[8]\big)/(q+1)$, mit dem die äquidistante Zerlegung Ω_M (hier als Anzahl der Intervalle) für jede glatt zu modellierende Kovariable festgelegt wird
bdeg	Grad der B–Splines, global für alle Effekte gültig (2 = Default)

und den Ausgabeparametern

design	Gemäß Belegung der Eingabeparameter erstellte Designmatrix
NoB	Vektor der Länge $num[3] + num[4] + \big(num[7] + num[8]\big)/(q+1)$; beinhaltet die für jede glatt zu modellierende Kovariable verwendete Anzahl an B–Splines
ModelId	Interne numerische Kodierung des Modellspezifizierers
LinkId	Interne numerische Kodierung des Linkspezifizierers

Für die obigen Beispiele lauten die entsprechenden Funktionsaufrufe mit den

zugrunde gelegten Belegungen für die Eingabeparameter

```
     zero <- rep(0,4)
(1)  out.d <- Design(Data,2,c(0,1,0,2,zero),'nom',,rep(20,2))
(2)  out.d <- Design(Data,2,c(2,3,0,0,zero),'cum',,NULL)
(3)  out.d <- Design(Data,2,c(0,1,2,1,zero),'cum',,rep(20,3),3)
```

Für den Aufbau der horizontalen Penalties steht die S–Funktion

HorPen(q,num,ncd,anzbf,lambda,ddeg)

mit den Eingabeparametern

q	vgl. *Design*
num	vgl. *Design*
ncd	Anzahl Spalten der Designmatrix
anzbf	Ausgabeparameter *NoB* von *Design*
lambda	Vektor der Länge $num[3]+num[4]\cdot q+num[7]/(q+1)+num[8]$; beinhaltet die Glättungsparameter der glatt modellierten Kovariablen
ddeg	Grad der Differenzen, global für alle Penalties gültig (1 = Default)

zur Verfügung. Rückgabewert von *HorPen* ist die Matrix K der horizontalen Strafterme. Die zugehörigen Funktionsaufrufe für die Beispiele mit nonparametrischen Komponenten lauten

```
     cols <- ncol(out.d$design)
(1)  lambda <- c(5.32,10.49,11.46,0.82)
     out.ph <- HorPen(2,c(0,1,0,2,zero),cols,out.d$NoB,lambda)
(3)  lambda <- c(0.09,46.45,36.97,134.54)
     out.ph <- HorPen(2,c(0,1,2,1,zero),cols,out.d$NoB,lambda,2)
```

Dem Aufbau vertikaler Penalties dient die S–Funktion

VerPen(q,num,ncd,anzbf,lambda,ddeg)

mit den Eingabeparametern

q	vgl. *Design*
num	vgl. *Design*
ncd	vgl. *HorPen*
anzbf	vgl. *HorPen*
lambda	Vektor der Länge $1+num[2]+num[4]$; beinhaltet die Glättungsparameter der Schwellenwerte und der kategorienspezifisch modellierten Kovariablen
ddeg	vgl. *HorPen*

Rückgabewert von *VerPen* ist die Matrix K der vertikalen Penalties. Für die Beispiele mit vertikaler Bestrafung der kategorienspezifischen Parameter lauten die entsprechenden Funktionsaufrufe

(2)
```
lambda <- c(0,0,1,1)
out.pv <- VerPen(2,c(2,3,0,0,zero),cols,out.d$NoB,lambda)
```

(3)
```
lambda <- c(0,8.2,0.76)
out.pv <- VerPen(2,c(0,1,2,1,zero),cols,out.d$NoB,lambda)
```

Die eigentliche Maximum-Likelihood-Schätzung der unbekannten Parameter wird über die S–Funktion

CatReg(q,nrx,ncd,model.int,link.int,design,resp,pen,init,iter.max,eps,gm)

realisiert, deren Eingabeparameter als

q	vgl. *Design*
nrx	Anzahl der Beobachtungen (Anzahl der Zeilen von *Data*)
ncd	Anzahl Spalten der Designmatrix
model.int	Ausgabeparameter *ModelId* von *Design*
link.int	Ausgabeparameter *LinkId* von *Design*
design	Ausgabeparameter *design* von *Design*
resp	Dummy–kodierter Response
pen	Summe aus horizontaler und vertikaler Penaltymatrix
init	Vektor der Länge *ncd*; beinhaltet Initialisierungswerte des Parametervektors $\boldsymbol{\beta}$
iter.max	Maximale Anzahl von Fisher–Scoring Iterationen (30 = Default)

eps	Fisher–Scoring stoppt, wenn relative Änderung zwischen Folge-Iterierten kleiner ist als *eps* (0.001 = Default) oder nach *iter.max* Schritten
gm	Faktor γ vor Spur der Hatmatrix im Zielkriterium zur Optimierung der Glättungsparameter (2 = Default)

gegeben sind. Als Ausgabeparameter werden die Größen

konv	Logischer Indikator; falls `konv==T`, ist Fisher-Scoring nach höchstens *iter.max* Schritten konvergiert
inv	Logischer Indikator; falls `inv==T`, war (Pseudo–)Fisher–Matrix $\tilde{F}$ in jeder Iteration invertierbar. Algorithmus bricht ab, sobald `inv==F`
coef	Geschätzter Parametervektor im Konvergenzstadium
prob	Geschätzte Responsewahrscheinlichkeiten im Konvergenzstadium
fish	(Pseudo–)Fisher–Matrix im Konvergenzstadium
fish.inv	Inverse der (Pseudo–)Fisher–Matrix im Konvergenzstadium
zk	Zielkriterium im Konvergenzstadium

zurückgeliefert. Für die genannten drei Beispiele lauten die Funktionsaufrufe entsprechend

```
     rows <- nrow(Data)
(1)  start <- rep(0,cols)
     out.r <- CatReg(2,rows,cols,out.d$ModelId,out.d$LinkId,
                     out.d$design,response,out.ph,start,,,4)
(2)  start <- c(1:2,rep(0,cols-2))
     out.r <- CatReg(2,rows,cols,out.d$ModelId,out.d$LinkId,
                     out.d$design,response,out.pv,start)
(3)  start <- c(1:2,rep(0,cols-2))
     out.r <- CatReg(2,rows,cols,out.d$ModelId,out.d$LinkId,
                     out.d$design,response,out.ph+out.pv,start)
```

Die Schätzung komplexer multivariater GLM's in der Interpretersprache S–Plus beansprucht lange Rechenzeiten und hohe Speicherressourcen. Ein Teil der rechen- und speicherintensiven Prozesse wurde aus diesem Grund in der

Kompilersprache C implementiert und über ein vorhandenes Interface in die S–Plus Routinen eingebunden. Detaillierte Erläuterungen zur Schnittstellennutzung zwischen C und S–Plus finden sich in Venables & Ripley (1999) und Burns (1998).

Zu den ausgelagerten Programmabschnitten gehört die Dummy–Kodierung des kategorialen Response, hier realisiert über die C–Funktion

$$\textit{void resp(int } *x\textit{, int } *y\textit{, int } *n\textit{)}$$

mit den Parametern

- x Responsevektor
- y Nullvektor der Länge $N \cdot (q+1)$
- n Vektor der Länge 2 mit den Komponenten $(N, q+1)$

Bezeichnet $\boldsymbol{z}$ den Vektor der Responsebeobachtungen, so ist dem Aufruf von *CatReg* in den drei Beispielen stets die Zuweisung

```
response <- .C('resp',x=as.integer(z),y=as.integer(rep(0,3*rows
                )),n=as.integer(c(rows,3)))$y
```

vorauszuschicken. Alle sonstigen Auslagerungen nach C sind Interna der beschriebenen S–Funktionen und daher hier nicht von primärem Interesse. Für die Benutzung der einzelnen Funktionen ist es jedoch zwingend erforderlich, den integrierten C–Code zunächst in S–Plus zu laden. Dies geschieht durch den S–Befehl

```
dyn.load('objects.o')
```

mit einem Objektfile *objects.o*, in dem der zu ladende C–Code akkumuliert vorliegt.

Die vorgestellten S–Plus Funktionen zur Verwendung von P–Splines in kategorialen Regressionsmodellen sind unter *www.stat.uni-muenchen.de/∼scholz* frei im Internet erhältlich. Unter der genannten Adresse wird auch der zu ladende C–Code als Objektfile *objects.o* zur Verfügung gestellt.

Der hier zur Bestimmung der Glättungsparameter genutzte genetische Algorithmus wurde von R. Krause als allgemeines Instrumentarium zur Optimie-

rung einer Zielfunktion mit ausschließlich metrischen Argumenten implementiert. Der zugehörige MATLAB–Quellcode ist im Internet unter der Adresse *www.stat.uni-muenchen.de/∼krause* ebenfalls frei zugänglich.

C Notationen

Die nachstehende Auflistung beschreibt die in der Arbeit gewählten Schreibweisen für mathematische Größen bzw. Operatoren. Ausgenommen sind Bezeichnungen, die sich an bekannten Standards orientieren. Dies betrifft insbesondere die Symbolisierung von Zahlenmengen und mathematischen Konstanten. Notationen, denen eine wichtige inhaltliche Rolle zukommt, werden an der Stelle ihres ersten Auftretens im Text definiert.

Mathematische Größen

- Schreibweise von *Vektoren*: klein und fett. Beispiel: $\boldsymbol{a} \in \mathbb{R}^n$ mit Komponenten $a_1, \ldots, a_n$. *Spezielle Vektoren*:
  - *Einservektor* $\boldsymbol{a} = \mathbf{1}_n$: $a_i = 1,\ i = 1, \ldots, n$
  - *Nullvektor* $\boldsymbol{a} = \mathbf{0}_n$: $a_i = 0,\ i = 1, \ldots, n$
  - *Einheitsvektor* $\boldsymbol{a} = \boldsymbol{e}_{k,n}$: $a_k = 1,\ a_i = 0,\ i \neq k$
- Schreibweise von *Matrizen*: groß. Beispiel: $A \in M_{\mathbb{R}}(s,t)$ mit Einträgen $A(i,j),\ i = 1, \ldots, s,\ j = 1, \ldots, t$. *Spezielle Matrizen*:
  - *Einsermatrix* $A = 1_{s \times t}$: $A(i,j) = 1,\ i = 1, \ldots, s,\ j = 1, \ldots, t$
  - *Nullmatrix* $A = 0_{s \times t}$: $A(i,j) = 0,\ i = 1, \ldots, s,\ j = 1, \ldots, t$
  - *Einheitsmatrix* $A = I_s$: $A(i,i) = 1,\ i = 1, \ldots, s$ und $A(i,j) = 0$, $i, j = 1, \ldots, s,\ i \neq j$
- Eine notationelle Unterscheidung von *stochastischen* und *deterministischen* Größen wird nicht vorgenommen. Der Charakter einer Größe ist dem jeweiligen Kontext zu entnehmen.

Operatoren

- *Kronecker–Produkt* $\otimes$ von Matrizen $A \in M_{\mathbb{R}}(s,t)$ und $B \in M_{\mathbb{R}}(u,v)$: $A \otimes B \in M_{\mathbb{R}}(s \cdot u, t \cdot v)$ resultiert aus der Vorschrift: Ersetze jedes Element $A(i,j)$ in A durch $A(i,j) \cdot B$.

- *Zeilenweises Kronecker–Produkt* $\widetilde{\otimes}$:
  Für Matrizen $A \in M_{\mathbb{R}}(s,t)$ und $B \in M_{\mathbb{R}}(s,v)$ definiere

  $$A \mathbin{\widetilde{\otimes}} B \in M_{\mathbb{R}}(s, t \cdot v), \text{ mit } \big(A \mathbin{\widetilde{\otimes}} B\big)(i,\cdot) := A(i,\cdot) \otimes B(i,\cdot), \;\; i = 1, \ldots, s.$$

- Generieren von *Diagonalmatrizen* via diag
  Für einen Vektor $\boldsymbol{x} := (x_1, \ldots, x_n)' \in \mathbb{R}^n$ definiere

  $$\operatorname{diag}(\boldsymbol{x}) = \operatorname{diag}(x_1, \ldots, x_n) := \begin{pmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{pmatrix}$$

- Generieren von *Blockdiagonalmatrizen* via Diag
  Für Matrizen $A_1, \ldots, A_n$, $A_i \in M_{\mathbb{R}}(s_i, t_i)$, $i = 1, \ldots, n$, definiere

  $$\operatorname{Diag}(A_1, \ldots, A_n) := \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_n \end{pmatrix} \in M_{\mathbb{R}}\big(\textstyle\sum_i s_i, \sum_i t_i\big)$$

Literatur

Abe, M. (1999). A Generalized Additive Model for Discrete-Choice Data. *Journal of Business and Economic Statistics* **17**, 271–284.

Adebayo, S. B. (2001). Multinomial Logistic Regression Analysis of Sickle Cell Anaemia Data. *Journal of the Nigerian Statistical Association* **14**, 18–25.

Aerts, M., Claeskens, G., und Wand, M. P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference* **103**, 455–470.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Hrsg.), *Proc. 2nd Int. Symp. on Information Theory*, Budapest: Akadémiai Kiadó.

Amemiya, T. (1981). Qualitative Response Models: A Survey. *Journal of Economic Literature* **19**, 1483–1536.

Armstrong, B. und Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* **129**, 191–204.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika* **67**, 413–418.

Azzalini, A. und Bowman, A. W. (1993). On the Use of Nonparametric Regression for Checking Linear Relationships. *Journal of the Royal Statistical Society* B **55**, 549–557.

Bender, R. und Grouven, U. (1998). Using Binary Logistic Regression Models for Ordinal Data with Non–proportional Odds. *Journal of Clinical Epidemiology* **51**, 809–816.

Bhansali, R. J. und Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 547–551.

Block, H. D. und Marschak, J. (1960). Random orderings and stochastic theories of responses. In O. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Hrsg.), *Contributions to Probability and Statistics*. Stanford: Stanford University Press.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Bowman, A. W. und Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford Statistical Science Series: Clarendon Press.

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics* **46**, 1171–1178.

Brezger, A. und Lang, S. (2003). Generalized structured additive regression based on Bayesian P-Splines. Discussion Paper 321, SFB 386, Institut für Statistik, Universität München.

Burns, P. J. (1998). S Poetry. `URL: http://www.burns-stat.com/`.

Crainiceanu, C. M., Ruppert, D., Claeskens, G., und Wand, M. P. (2003). Exact Likelihood Ratio Tests for Penalized Splines. Preprint.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford: Clarendon Press.

Eilers, P. H. C. und Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science* **11**, 89–121.

Eilers, P. H. C. und Marx, B. D. (2002). Generalized Linear Additive Smooth Structures. *Journal of Computational and Graphical Statistics* **11**, 758–783.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

Fahrmeir, L. (1987). Asymptotic Testing Theory for Generalized Linear Models. *Statistics* **18**, 65–76.

Fahrmeir, L. und Kaufmann, H. (1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics* **13**, 342–368.

Fahrmeir, L., Kneib, T., und Lang, S. (2003). Penalized additive regression for space-time data: a Bayesian perspective. Revised for *Statistica Sinica*.

Fahrmeir, L. und Lang, S. (2001). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics* **53**, 10–30.

Fahrmeir, L. und Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2. Aufl.). New York: Springer.

Fan, J. und Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Farewell, V. T. (1982). A note on regression analysis of ordinal data with variability of classification. *Biometrika* **69**, 533–538.

Firth, D., Glosup, J., und Hinkley, D. V. (1991). Model checking with nonparametric curves. *Biometrika* **78**, 245–252.

Friedman, J. (1991). Multivariate Adaptive Regression Splines (with Discussion). *The Annals of Statistics* **19**, 1–141.

Friedman, J. und Silverman, B. (1989). Flexible parsimonious smoothing and additive modelling (with discussion). *Technometrics* **31**, 3–39.

Gasser, T. und Müller, H. G. (1984). Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scandinavian Journal of Statistics* **11**, 171–185.

Gelfand, I. und Schilow, G. (1967). *Verallgemeinerte Funktionen (Distributionen),* Band I. Berlin: Deutscher Verlag der Wissenschaften.

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

Golub, G. H. (1989). *Matrix Computations.* Baltimore, MD: John Hopkins University Press.

Green, P. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55**, 245–259.

Green, P. und Yandell, B. (1985). Semi-Parametric Generalized Linear Models. In R. Gilchrist, B. Francis, & J. Whittaker (Hrsg.), *Generalized Linear Models.* Heidelberg: Springer Lecture Notes.

Gu, C. (1990). Adaptive Spline Smoothing in Non-Gaussian Regression Models. *Journal of the American Statistical Association* **85**, 801–807.

Hämmerlin, G. und Hoffmann, K. H. (1992). *Numerische Mathematik.* Berlin: Springer-Verlag.

Hastie, T. und Loader, C. (1993). Local Regression: Automatic Kernel Carpentry. *Statistical Science* **8**, 120–143.

Hastie, T. und Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Hastie, T. und Tibshirani, R. (1993). Varying–coefficient Models. *Journal of the Royal Statistical Society* B **55**, 757–796.

Holland, J. (1975). *Adaption in neural and artificial systems.* Ann Arbor: University of Michigan Press.

Hruschka, H. (2002). Market share analysis using semi-parametric attraction models. *European Journal of Operational Research* **138**, 212–225.

Hurvich, C. M., Simonoff, J. S., und Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society* B **60**, 271–293.

Jörgens, V., Grüsser, M., Bott, U., Mühlhauser, I., und Berger, M. (1993). Effective and safe translation of intensified insulin therapy to general internal medicine departments. *Diabetologia* **36**, 99–105.

Kaufmann, H. (1988). On Existence and Uniqueness of Maximum Likelihood Estimates in Quantal and Ordinal Response Models. *Metrika* **35**, 291–313.

Kay, R. und Little, S. (1986). Assessing the Fit of the Logistic Model: A Case Study of Children with the Haemolytic Uraemic Syndrome. *Applied Statistics* **35**, 16–30.

Krause, R. und Tutz, G. (2003). Additive Modelling with Penalized Regression Splines and Genetic Algorithms. Discussion Paper 312, SFB 386, Institut für Statistik, Universität München.

Lang, S. und Brezger, A. (2003). Bayesian P–Splines. *Journal of Computational and Graphical Statistics* (to appear).

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

McCullagh, P. (1980). Regression Models for Ordinal Data (with Discussion). *Journal of the Royal Statistical Society* B **42**, 109–127.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Hrsg.), *Frontiers in Econometrics*. New York: Academic Press.

Mühlhauser, I., Bender, R., Bott, U., Jörgens, V., Grüsser, M., und Wagner, W. (1996). Cigarette smoking and progression of retinopathy and nephropathy in type 1 diabetes. *Diabetic Medicine* **13**, 536–543.

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Berlin, Heidelberg: Springer.

Nelder, J. A. und Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society* A **135**, 370–384.

O'Sullivan, F., Yandell, B., und Raynor, W. (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association* **81**, 96–103.

Peterson, B. und Harrell, F. E. (1990). Partial Proportional Odds Models for Ordinal Response Variables. *Applied Statistics* **39**, 205–217.

Pruscha, H. und Göttlein, A. (2002). Regression analysis of forest inventory data with time and space dependencies. *Environmental and Ecological Statistics* **9**, 43–56.

Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik* **10**, 177–183.

Ruppert, D. (2002). Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, D. und Carroll, R. J. (2000). Spatially–adaptive Penalties for Spline Fitting. *Australian and New Zealand Journal of Statistics* **42**, 205–223.

Ruppert, D. und Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics* **22**, 1346–1370.

Ruppert, D., Wand, M. P., und Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Santner, T. J. und Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*. New York: Springer.

Schwarz, G. (1978). Estimating the Dimensions of a Model. *The Annals of Statistics* **6**, 461–464.

Seber, G. A. F. und Wild, C. J. (1989). *Nonlinear Regression*. New York: Wiley.

Silverman, B. W. (1985). Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting (with Discussion). *Journal of the Royal Statistical Society* B **47**, 1–52.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer.

Staniswalis, J. G. (1989). The Kernel Estimate of a Regression Function in Likelihood-Based Models. *Journal of the American Statistical Association* **84**, 276–283.

Stone, C., Hansen, M., Kooperberg, C., und Truong, Y. (1997). Polynomial Splines and their Tensor Products in Extended Linear Modeling. *The Annals of Statistics* **25**, 1371–1470.

Tutz, G. (2000). *Die Analyse kategorialer Daten*. München: Oldenbourg Verlag.

Tutz, G. (2003a). Generalized semiparametrically structured mixed models. Accepted for *Computational Statistics & Data Analysis*.

Tutz, G. (2003b). Generalized semiparametrically structured ordinal models. *Biometrics* **59**, 263–273.

Venables, W. N. und Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*. New York: Springer.

Wand, M. P. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika* **86**, 936–940.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics* (in press).

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27–32.

Yee, T. W. und Wild, C. J. (1996). Vector Generalized Additive Models. *Journal of the Royal Statistical Society* B **58**, 481–493.

Lebenslauf

Torsten Scholz
geboren am 01.01.1973 in Räckelwitz

Schulbildung

09/1979–08/1987	Polytechnische Oberschulen in Oranienburg und Teltow
09/1987–06/1991	ESOS „Georg Thiele“ in Kleinmachnow

Wehrdienst

04/1992–03/1993	Wehrdienst in Berlin und Dahmsdorf

Studium

10/1991–10/1998	Studium der Wirtschaftsmathematik am Fachbereich Mathematik der Technischen Universität Berlin
04/1995	Vordiplom in Wirtschaftsmathematik
10/1998	Diplom in Wirtschaftsmathematik (Dipl.–Math. oec)

Berufliche Tätigkeiten

11/1998–10/2003	vollbeschäftigter wissenschaftlicher Mitarbeiter am Institut für Statistik der Ludwig–Maximilians–Universität München bei Prof. Dr. Gerhard Tutz
11/2003–02/2004	teilzeitbeschäftigt im Sonderforschungsbereich 386 am Institut für Statistik der Ludwig–Maximilians–Universität München bei Prof. Dr. Gerhard Tutz