Ludwig-Maximilians-Universität München
Institut für Statistik

Diplomarbeit

Bayes-Inferenz in generalisierten geoadditiven gemischten Modellen

(Korrigierte Version)

Betreuung:

Prof. Dr. Ludwig Fahrmeir
Dr. Stefan Lang

Thomas Kneib

Danke

Viele Personen verdienen Dank für ihre direkte oder indirekte Hilfe bei der Erstellung dieser Arbeit:

- Ludwig Fahrmeir und Stefan Lang für die hervorragende Betreuung und stetige Ermunterung,
- Rüdiger Krause für die Durchführung der Schätzungen mit Hilfe des Programms `mgcv` in Kapitel 5.1,
- Susanne Breitner, Ralf Breuninger, Monika Entholzner, Stefan Krieger, Katharina Lensing und David Rummel dafür, dass sie nicht nur ohne Murren die ständige Belegung mehrerer Computer zur Durchführung der Simulationen ertragen haben, sondern auch immer bereit waren, ihre Arbeit für einen Kaffee oder Tee zu unterbrechen,
- meine Eltern, die mich immer unterstützt haben, sogar als ich den Wunsch äußerte, Statistik zu studieren
- und Daniela für Verständnis, Zuneigung und Aufmerksamkeit.

Inhaltsverzeichnis

1 Problemstellung

Eine wesentliche Fragestellung der statistischen Analyse ist es, den Zusammenhang zwischen einer Reihe von Variablen oder Merkmalen zu untersuchen. Häufig ist dabei eine der Variablen als abhängige Variable ausgezeichnet, die von den übrigen Variablen beeinflusst wird. Diese abhängige Variable y wird dann oft auch als Response, die übrigen Variablen $x_1, \ldots, x_p$ als Kovariablen bezeichnet.

Der vermutlich am häufigsten zur Modellierung solcher gerichteten Zusammenhänge verwendete Ansatz ist die Regressionsanalyse. Im klassischen linearen Modell wird die Abhängigkeit zwischen den Kovariablen und dem als normalverteilt angenommenen Response durch einen linearen Zusammenhang beschrieben. Genauer nimmt man an, dass

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta = \eta$$

gilt. Für jede der Kovariablen wird also der Einfluss auf den Erwartungswert des Response als linear betrachtet. Die Summe der Effekte η wird daher als linearer Prädiktor bezeichnet. Durch Nelder & Wedderburn (1972) wurde das klassische lineare Modell auf die Analyse auch nicht normalverteilter Daten erweitert, indem man den Zusammenhang zwischen Kovariablen und Response mit Hilfe einer geeigneten Response-Funktion $h(\cdot)$ über

$$\mathbb{E}(y|x) = \mu = h(\eta)$$

beschreibt. Man spricht in diesem Fall von generalisierten linearen Modellen. Sowohl für das klassische lineare Modell als auch in generalisierten linearen Modellen werden die Beobachtungen des Response als unabhängig angenommen.

Obwohl mit Hilfe von generalisierten linearen Modellen bereits eine große Klasse von Modellen analysiert werden kann, ist man in vielen Datensituationen mit einem oder mehreren der folgenden Probleme konfrontiert:

- Der Zusammenhang zwischen einigen Kovariablen und dem Response y lässt sich nicht angemessen durch eine Gerade beschreiben.
- Die Beobachtungen sind räumlich oder zeitlich korreliert.
- Zwischen den Beobachtungen besteht unbeobachtete Heterogenität, das heißt, die Unterschiede zwischen den Beobachtungen lassen sich nicht aus-

reichend durch die beobachteten Kovariablen erklären. Ursache unbeobachteter Heterogenität ist häufig die Zugehörigkeit der Beobachtungen zu verschiedenen Gruppen.

Um auch solchen Datensituationen gerecht werden zu können, wurde das generalisierte lineare Modell in verschiedener Hinsicht erweitert. Durch Hastie & Tibshirani (1990) wurde eine flexible Modellierung des Zusammenhangs zwischen metrischen Kovariablen und Response eingeführt, mit deren Hilfe sich dieser Zusammenhang über glatte, aber in ihrer funktionalen Form unspezifizierte Funktionen beschreiben lässt. Dabei wird der lineare Prädiktor η ersetzt durch den additiven Prädiktor

$$\eta = f_1(x_1) + \ldots + f_p(x_p).$$

Zur Schätzung der Funktionen f_1 bis f_p existieren eine Vielzahl von Vorschlägen, man vergleiche beispielsweise Fahrmeir & Tutz (2001) Kapitel 5 für einen Überblick. Ein verhältnismäßig einfacher und auch numerisch vorteilhafter Vorschlag, der unter dem Namen P-Splines bekannt ist, stammt von Eilers & Marx (1996) und ermöglicht die simultane Schätzung einer relativ großen Zahl von Funktionen, weil jede Funktion durch eine relativ geringe Zahl von Parametern beschrieben wird. Man spricht in diesem Zusammenhang auch von Low-Rank-Verfahren.

Das Problem zeitlich korrelierter Beobachtungen lässt sich beispielsweise durch die Einbeziehung der Kalenderzeit oder einer anderen Zeitskala als Kovariable lösen. Insbesondere kann es hilfreich sein, auch den zeitlichen Einfluss nichtlinear durch eine glatte Funktion zu schätzen. In ähnlicher Weise lassen sich räumliche Korrelationen durch die Modellierung einer räumlich strukturierten, glatten Funktion berücksichtigen. Solche Funktionen können beispielsweise mit Hilfe von Markov-Zufallsfeldern definiert werden (vergleiche etwa Besag, York & Mollié (1991)).

Zur Modellierung unbeobachteter Heterogenität bieten sich Modelle mit zufälligen Effekten an (Fahrmeir & Tutz (2001), Kapitel 7). Dabei werden einige der Regressionskoeffizienten nicht mehr als feste, unbekannte Parameter, sondern als Zufallsgrößen betrachtet. Mit Hilfe dieses Ansatzes lassen sich beispielsweise gruppenspezifische Effekte auf eine parametersparsame Art und Weise modellieren.

Durch eine Kombination der verschiedenen Ansätze zu einem Modell, das im Folgenden als generalisiertes geoadditives gemischtes Modell bezeichnet werden

soll, lassen sich prinzipiell die oben beschriebenen Probleme lösen. Obwohl die Bestimmung der einzelnen Modellkomponenten mit bekannten Verfahren prinzipiell durchführbar ist, entsteht durch ihre Kombination ein wesentlich komplexeres Schätzproblem. Insbesondere hängt die Schätzung von einer Reihe von Hyperparametern ab. Beispielsweise werden bei der Modellierung des Einflusses einer Kovariablen über P-Splines die Schätzungen durch einen Glättungsparameter beeinflusst, der den Kompromiss zwischen Datentreue und Glattheit der Funktionsschätzung steuert. Ähnlich hängen auch die über Markov-Zufallsfelder definierten räumlichen Funktionen von Glättungsparametern und die zufälligen Effekte von Varianzparametern ab. Während die Schätzung bei gegebenen Hyperparametern relativ einfach möglich ist, war die geeignete Wahl der Hyperparameter lange Zeit eine schwierige Aufgabe.

Im Rahmen dieser Arbeit soll nun eine verhältnismäßig einfache Möglichkeit beschrieben werden, alle Modellkomponenten eines generalisierten geoadditiven gemischten Modells, also insbesondere auch die Hyperparameter, über einen einheitlichen Ansatz zu bestimmen. Dabei wird ausgenutzt, dass sowohl die Schätzung von P-Splines als auch von Markov-Zufallsfeldern auf einer Penalisierung der zur Schätzung verwendeten Log-Likelihood beruht. Diese Tatsache erlaubt die Darstellung des gesamten Modells als generalisiertes lineares Modell mit zufälligen Effekten, in dem dann alle Hyperparameter als Varianzparameter betrachtet werden können. Zur Schätzung dieser Varianzparameter existieren bereits anwendbare Schätzverfahren.

Generalisierte lineare gemischte Modelle bilden damit die Grundlage für die in dieser Arbeit untersuchten Methoden. Daher werden zunächst in Kapitel 2 eine Reihe grundlegender Ergebnisse zu diesen Modellen zusammengetragen. Diese betreffen insbesondere die Schätzung von Varianzparametern in generalisierten linearen gemischten Modellen. In Kapitel 3 wird dann, basierend auf P-Splines und Markov-Zufallsfeldern, das allgemeinere generalisierte geoadditive gemischte Modell eingeführt. Kernpunkt dieses Kapitels ist neben der Beschreibung der einzelnen Modellkomponenten die Darstellung des komplexeren Modells als generalisiertes lineares gemischtes Modell, so dass die Schätzverfahren aus Kapitel 2 anwendbar werden. Es werden aber auch kurz herkömmliche Schätzverfahren behandelt, um die Vorteile des neuen Ansatzes zu verdeutlichen.

Die Darstellung als Modell mit zufälligen Effekten eröffnet zudem die Möglichkeit, zumindest in einigen einfachen Modellen der nonparametrischen Regression und der räumlichen Analyse, Tests der Glättungsparameter mit Hilfe von Likelihood-Quotienten-Tests durchzuführen. Häufig interessiert man sich dabei besonders für den Fall, dass der Glättungsparameter auf dem Rand seines Parameterraums liegt. Dieser Fall ist jedoch nicht durch die Standardtheorie zu Likelihood-Quotienten-Tests abgedeckt, so dass eine genauere Betrachtung der entsprechenden Verteilungsaussagen notwendig ist. In Kapitel 4 werden daher Likelihood-Quotienten-Tests in linearen gemischten Modellen behandelt, die in verschiedener Hinsicht von den in der Standardtheorie zu Likelihood-Quotienten-Tests angenommenen Regularitätsbedingungen abweichen. In einer Reihe von Beispielen werden außerdem mögliche Anwendungen dieser Tests demonstriert. So erhält man etwa die Möglichkeit, die Notwendigkeit einer nonparametrischen Modellierung des Effekts einer Kovariablen zu überprüfen, das heißt einen Test auf Linearität des Effekts durchzuführen.

In Kapitel 5 sollen dann in einer Reihe von Simulationsstudien die Güteeigenschaften sowohl der in Kapitel 3 beschriebenen Schätzverfahren als auch der in Kapitel 4 beschriebenen Tests näher untersucht werden. Die Schätzverfahren werden dazu auch mit anderen Ansätzen verglichen, die jedoch meist nur die Schätzung von generalisierten additiven Modellen ohne zufällige und räumliche Effekte erlauben.

Um konkrete Anwendungsmöglichkeiten der vorgestellten Verfahren aufzuzeigen, werden abschließend in Kapitel 6 mehrere reale Datensätze analysiert. Durch die verschiedenen untersuchten Fragestellungen wird nochmals die Flexibilität und breite Anwendbarkeit der Verfahren demonstriert.

Um die Notation in den folgenden Kapiteln etwas zu erleichtern, soll in den Darstellungen nicht zwischen Zufallsvariablen und ihren Realisationen unterschieden werden. Außerdem werden die zur Schätzung verwendeten Likelihood-Funktionen stets mit $L(\cdot)$ sowie die zugehörigen Log-Likelihoods mit $l(\cdot)$ bezeichnet. Ebenso wird für Dichten stets die Bezeichnung $p(\cdot)$ verwendet. Aus dem Zusammenhang und den Argumenten der jeweiligen Funktionen sollte jedoch in der Regel klar sein, um welche Funktion es sich handelt.

2 Generalisierte lineare gemischte Modelle

2.1 Lineare gemischte Modelle

2.1.1 Modell

Im gewöhnlichen linearen Modell wird der Zusammenhang zwischen der abhängigen Variablen y und Kovariablen $x_1, \ldots, x_p$ durch

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon = x'\beta + \varepsilon$$

modelliert. Dabei nimmt man an, dass der Zusammenhang zwischen Response und Kovariablen nicht deterministisch gilt, sondern eine gewisse, zufällige Streuung aufweist. Genauer nimmt man an, dass sich der Response zerlegen lässt in den deterministischen Anteil $x'\beta$, der aus statistischer Perspektive dem Erwartungswert von y entspricht, und einen zufälligen Fehler ε.

Um die unbekannten Modellparameter $\beta_0, \ldots, \beta_p$ zu schätzen, werden Daten in Form von n unabhängigen Messwiederholungen $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$ erhoben. Für jede dieser Messwiederholungen gilt dann das Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i.$$

Um das Modell vollständig zu spezifizieren, müssen noch Annahmen über die Fehler ε_i getroffen werden. Üblicherweise nimmt man an, dass diese stochastisch unabhängig und identisch verteilt sind mit

$$\mathbb{E}(\varepsilon_i) = 0 \text{ und } \operatorname{Var}(\varepsilon_i) = \sigma^2.$$

In Matrixschreibweise lässt sich das Modell zusammenfassen zu

$$y = X\beta + \varepsilon$$

mit

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Als Annahme für ε erhält man

$$\mathbb{E}(\varepsilon) = 0 \text{ und } \operatorname{Var}(\varepsilon) = \sigma^2 I_n.$$

Häufig fordert man zusätzlich zur identischen Verteilung der ε_i auch noch einen speziellen Verteilungstyp. Dies erlaubt nicht nur die Konstruktion von Tests und Konfidenzbereichen im linearen Modell, sondern auch, dass alle Modellparameter über ein einheitliches Prinzip, das Maximum-Likelihood-Prinzip, beziehungsweise Varianten dieses Prinzips, geschätzt werden können. Mathematisch vorteilhaft und in Datensituationen mit stetigem y zumindest nach einer geeigneten Transformation des Response häufig auch angemessen ist die Annahme der Normalverteilung für die Fehler, das heißt

$$\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n).$$

Im linearen gemischten Modell wird das klassische lineare Modell erweitert zu

$$y = X\beta + Zb + \varepsilon,$$

wobei Z eine $n \times q$ Designmatrix, vergleichbar mit X, ist und b ein q-dimensionaler Parametervektor. Im Unterschied zu β wird b jedoch nicht als fester, unbekannter Parameter betrachtet, sondern als Zufallsgröße, für die

$$\mathbb{E}\begin{pmatrix}\varepsilon \\ b\end{pmatrix} = \begin{pmatrix}0 \\ 0\end{pmatrix} \text{ und } \operatorname{Var}\begin{pmatrix}\varepsilon \\ b\end{pmatrix} = \begin{pmatrix}\sigma^2 I_n & 0 \\ 0 & Q(\nu)\end{pmatrix}$$

gelten soll. Insbesondere gilt also $\operatorname{Var}(b) = Q(\nu)$, wobei $Q(\nu)$ eine symmetrische, positiv semidefinite Kovarianzmatrix ist, die von einem Vektor ν von Varianzparametern abhängt. Außerdem fordert man, dass ε und b unkorreliert sein sollen. Der Parameter b wird aufgrund der Spezifikation als Zufallsgröße auch als zufälliger Effekt bezeichnet, während β auch fixer Effekt genannt wird.

Der Name „Gemischtes Modell“ wird verwendet, weil sowohl fixe als auch zufällige Effekte im Modell vorhanden sind. Eine ebenfalls gebräuchliche Bezeichnung ist „Modell mit zufälligen Effekten“.

Wie β kann auch b in einem gewissen, im folgenden Kapitel näher definierten Sinn ohne Verteilungsannahme optimal geschätzt werden. Um jedoch das einheitliche Maximum-Likelihood-Prinzip anwenden zu können, soll im Folgenden von der Annahme

$$\begin{pmatrix}\varepsilon \\ b\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma^2 I_n & 0 \\ 0 & Q(\nu)\end{pmatrix}\right)$$

ausgegangen werden.

Bedingt auf b erhält man für y ein gewöhnliches lineares Modell und für die bedingte Verteilung von y gegeben b gilt:

$$y|b \sim \mathrm{N}(X\beta + Zb, \sigma^2 I_n).$$

Als marginale Verteilung für y ergibt sich dagegen

$$y \sim \mathrm{N}(X\beta, V(\vartheta))$$

mit der Kovarianzmatrix

$$V(\vartheta) = \sigma^2 I_n + ZQ(\nu)Z'$$

und den im d-dimensionalen Vektor $\vartheta = (\sigma^2, \nu')'$ zusammengefassten Varianzparametern. Marginal erhält man für y also ein allgemeines lineares Modell, das heißt ein lineares Modell, in dem die Annahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ durch die allgemeinere Annahme $\varepsilon \sim \mathrm{N}(0, V(\vartheta))$ ersetzt wird.

Anwendung fand das lineare gemischte Modell lange insbesondere zur Analyse von Daten, deren Beobachtungen eine Gruppierungsstruktur aufweisen, zum Beispiel Longitudinaldaten (Gruppierung nach Individuen) oder räumlich strukturierte Daten (Gruppierung nach Regionen, zum Beispiel in Breslow & Clayton (1993)).

Ein klassisches Beispiel zur Verwendung von linearen gemischten Modellen ist die Analyse von an Schülern erhobenen Daten, wobei mehrere Gruppierungsebenen denkbar sind: Auf einer ersten Ebene könnten die Schüler in Klassen zusammengefasst werden, auf der zweiten Ebene Klassen zu Schulen und auf der dritten Ebene Schulen zu Regierungsbezirken. Aufgrund der Struktur der Daten ist es plausibel, anzunehmen, dass sich Schüler innerhalb einer Klasse in Bezug auf bestimmte Merkmale ähnlicher sind als Schüler verschiedener Klassen. Eine äquivalente Aussage gilt dann wiederum für Klassen einer Schule beziehungsweise für Schulen in einem Regierungsbezirk. Häufig lassen sich die Faktoren, von denen die Ähnlichkeit verursacht wird, wie beispielsweise Eigenschaften des Lehrers einer Klasse, aber nicht oder nur unvollständig erfassen. Man spricht dann auch von unbeobachteter Heterogenität der Daten. Dies ist bei der Analyse mit Hilfe von Regressionsmodellen problematisch, weil durch die unbeobachteten Einflussgrößen Korrelationen zwischen den Schülern einer Klasse entstehen. Man kann die Daten also nicht mehr, wie in den Annahmen des klassischen linearen Modells

gefordert, als unabhängig betrachten. Die Verwendung von Modellen mit zufälligen Effekten bietet nun die Möglichkeit, die Gruppierungsstruktur der Daten und die daraus entstehenden Korrelationen zu berücksichtigen.

Betrachtet man nur eine Stufe der Gruppierung, so erhält man ein Modell, das wie das im Folgenden beschriebene Modell für Longitudinaldaten aufgebaut ist. Bei Betrachtung mehrerer Gruppierungsebenen spricht man auch von Mehrebenenmodellen und genesteten zufälligen Effekten. Auf diesen Spezialfall soll im Folgenden nicht eingegangen werden; Details hierzu findet man beispielsweise in Bryk & Raudenbush (1992) Kapitel 8.

Es soll nun kurz das lineare gemischte Modell für Longitudinaldaten, wie es von Laird & Ware (1982) eingeführt wurde, vorgestellt und gezeigt werden, wie es als Spezialfall des allgemeinen Modells aufgefasst werden kann. Im Modell für Longitudinaldaten weisen Z und $Q(\nu)$ eine besondere Struktur auf, die aus der zugrunde liegenden Gruppierung resultiert. Durch einfache Modifikationen lässt sich dieses Modell auch auf die beiden anderen beschriebenen Datenstrukturen (Gruppierung nach Regionen beziehungsweise Gruppierung nach Schulklassen) oder andere Situationen, in denen die Daten einer Gruppierung unterliegen, übertragen.

Longitudinaldaten bestehen aus wiederholten Messungen von Variablen an N Individuen. Für jedes Individuum erhält man so eine Reihe von Messungen $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, N$, wobei die Zahl der Messungen n_i je nach Individuum verschieden sein kann. Im Gegensatz zum allgemeinen linearen gemischten Modell werden die Beobachtungen hier doppelt indiziert, um die Zugehörigkeit einer Messung zu einem bestimmten Individuum und den Zeitpunkt der Messung anzugeben. Fasst man die Beobachtungen eines Individuums im Vektor $y_i = (y_{i1}, \ldots, y_{in_i})'$ zusammen, so postuliert man auf einer ersten Stufe für jedes Individuum das lineare gemischte Modell

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i.$$

Hier repräsentiert der Vektor β Kovariableneffekte, die für alle Individuen als gleich angenommen werden. Der $\tilde{q}$-dimensionale Vektor $b_i = (b_{i1}, \ldots, b_{i\tilde{q}})'$ enthält dagegen individuenspezifische Effekte. Die $n_i \times \tilde{q}$-Matrix Z_i ist dabei häufig eine Teilmatrix von X_i. Dies ist aber keineswegs Voraussetzung zur Anwendbarkeit des Modells. Für die einzelnen Vektoren b_i nimmt man als Verteilung eine $\mathrm{N}(0, \tilde{Q}(\nu))$-

Verteilung an. Das allgemeine lineare gemischte Modell erhält man dann mit

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, Z = \begin{pmatrix} Z_1 & 0 & \dots & \dots & 0 \\ 0 & Z_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & Z_N \end{pmatrix},$$

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}, Q(\nu) = \begin{pmatrix} \tilde{Q}(\nu) & 0 & \dots & \dots & 0 \\ 0 & \tilde{Q}(\nu) & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \tilde{Q}(\nu) \end{pmatrix}$$

zurück. Dabei gilt $n = \sum_{i=1}^{N} n_i$ und $q = N \cdot \tilde{q}$. Für die marginale Kovarianzmatrix von y erhält man

$$\text{Var}(y) = V(\vartheta) = \sigma^2 I_n + ZQ(\nu)Z' = \text{blockdiag}(\sigma^2 I_{n_i} + Z_i \tilde{Q}(\nu) Z_i').$$

Wie man sieht, sind also die einzelnen Vektoren y_i unabhängig, während zwei Beobachtungen eines Individuums y_{ij} und y_{ik} in der Regel abhängig sind. Durch die Modellierung mit zufälligen Effekten kann also das beschriebene Problem korrelierter Beobachtungen aufgrund unbeobachteter Heterogenität behoben werden.

Lineare gemischte Modelle können jedoch auch in anderen, allgemeineren Datensituationen nützlich sein. Insbesondere erlauben sie die Schätzung generalisierter geoadditiver gemischter Modelle und die automatische Wahl von Glättungsparametern, wie in Kapitel 3 beschrieben. Dazu benötigt man jedoch das Modell in seiner allgemeineren Formulierung. Darum wird im Weiteren nicht mehr der Spezialfall Longitudinaldaten betrachtet, sondern es werden Ergebnisse für das allgemeine Modell vorgestellt. Eine detaillierte Einführung in das allgemeine lineare gemischte Modell bieten beispielsweise Robinson (1991) oder McCulloch & Searle (2001) Kapitel 6, zum Modell für Longitudinaldaten siehe beispielsweise Fahrmeir & Tutz (2001) Kapitel 7, oder Verbeke & Molenberghs (2000). Die weitere Darstellung richtet sich nicht nur nach der Literatur zum allgemeinen linearen gemischten Modell, sondern auch nach den übrigen angegebenen Quellen.

2.1.2 Schätzung aus frequentistischer Sicht

Zunächst soll nun davon ausgegangen werden, dass die Varianzparameter $\vartheta = (\sigma^2, \nu')'$ bekannt sind und Schätzer für die Parameter β und b gesucht werden. Die Schätzung von ϑ wird in Kapitel 2.2 behandelt. Thema dieses Abschnitts ist die Schätzung in frequentistischer Betrachtungsweise, das heißt, β wird als fester unbekannter Parameter betrachtet.

Die Schätzung von β ist unabhängig von Schätzungen für b möglich. Ausgangspunkt hierfür ist die marginale Verteilung von y. Wie im vorigen Abschnitt erwähnt, erhält man bei marginaler Betrachtungsweise ein verallgemeinertes lineares Modell für y, das heißt ein lineares Modell in dem die Annahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ durch die allgemeinere Annahme $\varepsilon \sim \mathrm{N}(0, V(\vartheta))$ ersetzt wird.

Die Dichte der marginalen Verteilung von y kann man aus der bedingten Verteilung von y gegeben b und der Verteilung von b bestimmen:

$$\begin{aligned} p(y) &= \int p(y|b)p(b)db \\ &= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} |V(\vartheta)|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\beta)'V(\vartheta)^{-1}(y - X\beta)\right]. \end{aligned} \tag{2.1}$$

Mit $p(y), p(y|b)$ und $p(b)$ werden dabei die Dichten der jeweiligen Verteilungen bezeichnet. Man beachte, dass sich die Integration im Normalverteilungsfall analytisch durchführen lässt und man als marginale Verteilung für y wieder eine Normalverteilung erhält. In den generalisierten linearen gemischten Modellen aus Kapitel 2.3 wird die analytische Lösung des Integrals dagegen im Allgemeinen nicht mehr möglich sein.

Als Log-Likelihood für β erhält man unter Vernachlässigung additiver Konstanten

$$l(\beta) = -\frac{1}{2}\log(|V(\vartheta)|) - \frac{1}{2}(y - X\beta)'V(\vartheta)^{-1}(y - X\beta).$$

Maximiert man die Log-Likelihood bezüglich β, so ergibt sich als Schätzer für β der als Aitken-Schätzer bezeichnete gewichtete Kleinste-Quadrate-Schätzer

$$\hat{\beta} = (X'V(\vartheta)^{-1}X)^{-1}X'V(\vartheta)^{-1}y.$$

Dieser Schätzer ist (bei gegebenem ϑ) auch ohne die Normalverteilungsannahme für ε bester linearer unverzerrter Schätzer (BLUE) für β. Diese Optimalitätseigen-

schaft von $\hat{\beta}$ folgt unmittelbar aus dem Gauß-Markov-Theorem für das gewöhnliche lineare Modell, indem man im Modell $y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, V(\vartheta))$ die folgenden Substitutionen durchführt: Ersetze X durch $X^* = V(\vartheta)^{-\frac{1}{2}}X$, y durch $y^* = V(\vartheta)^{-\frac{1}{2}}y$ und ε durch $\varepsilon^* = V(\vartheta)^{-\frac{1}{2}}\varepsilon$. Man erhält so das gewöhnliche lineare Modell $y^* = X^*\beta + \varepsilon^*$ mit der Annahme $\varepsilon^* \sim \mathrm{N}(0, I_n)$ und als optimalen Schätzer $\hat{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (X'V(\vartheta)^{-1}X)^{-1}X'V(\vartheta)^{-1}y$ (vergleiche auch Rawlings, Pantula & Dickey (1998) Kapitel 12.5 oder Toutenburg (2003), Kapitel 7.3).

Im Allgemeinen benötigt man nicht nur Schätzungen für β, sondern auch für b. Zur simultanen Schätzung verwendet man als Likelihood die Dichte der gemeinsamen Verteilung von y und b

$$L(\beta, b) = p(y, b) = p(y|b)p(b).$$

Nimmt man für b eine Normalverteilung an, so maximiert man zur Schätzung von β und b also die Log-Likelihood

$$l(\beta, b) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta - Zb)'(y - X\beta - Zb) - \frac{1}{2}b'Q(\nu)^{-1}b, \quad (2.2)$$

die man für b als penalisierte Likelihood betrachten kann. Differenzieren und Nullsetzen liefern die Schätzer $\hat{\beta}$ und $\hat{b}$ als Lösungen des Gleichungssystems

$$\begin{pmatrix} X'W(\vartheta)X & X'W(\vartheta)Z \\ Z'W(\vartheta)X & Z'W(\vartheta)Z + Q(\nu)^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'W(\vartheta)y \\ Z'W(\vartheta)y \end{pmatrix}. \quad (2.3)$$

Dabei ist $W(\vartheta)$ definiert durch $W(\vartheta) = (\mathrm{Var}(y|b))^{-1} = \frac{1}{\sigma^2}I_n$ und wird bei der Schätzung generalisierter linearer gemischter Modelle in veränderter Form wieder verwendet werden. Darum sollen hier auch nicht die möglichen Vereinfachungen, die sich aus der einfachen Gestalt von $W(\vartheta)$ ergeben, ausgenutzt werden, sondern die Verbindung zur allgemeineren Schätzung in Kapitel 2.3.3 betont werden. Ein großer Vorteil der Darstellung in (2.3) ist die Vermeidung der Inversion von $V(\vartheta)$. Für $\hat{b}$ existieren jedoch noch eine Reihe weiterer Darstellungen (Harville (1976), Harville (1977)), die hier ebenfalls kurz vorgestellt werden sollen:

$$\begin{aligned} \hat{b} &= \left(Z'Z + \sigma^2 Q(\nu)^{-1}\right)^{-1} Z'(y - X\hat{\beta}) \quad (2.4) \\ &= Q(\nu)Z'V(\vartheta)^{-1}(y - X\hat{\beta}) \\ &= Q(\nu)Z'P(\vartheta)y \end{aligned}$$

mit

$$P(\vartheta) = V(\vartheta)^{-1} - V(\vartheta)^{-1}X(X'V(\vartheta)^{-1}X)^{-1}X'V(\vartheta)^{-1}.$$

Der Schätzer $\hat{b}$ besitzt mit (2.4) eine Darstellung als Ridge- beziehungsweise Shrinkage-Schätzer. Er stellt einen Kompromiss dar zwischen dem Kleinste-Quadrate-Schätzer für b, basierend auf den Residuen bezüglich der fixen Effekte, und dem Erwartungswert $\mathbb{E}(b) = 0$. Besonders deutlich wird dies bei Betrachtung eines Modells für Longitudinaldaten mit zufälligem Intercept: Modelliert man $\mathbb{E}(y_i)$ als $X_i\beta + b_i$ mit $b_i \sim \mathrm{N}(0, \nu^2)$, so erhält man als Schätzer für b_i:

$$\hat{b}_i = \frac{\nu^2}{\sigma^2/n_i + \nu^2}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - x'_{ij}\hat{\beta}).$$

Man schätzt also b_i aus dem Mittelwert der Residuen bezüglich des fixen Effekts und gewichtet diesen mit dem Faktor $\frac{\nu^2}{\sigma^2/n_i+\nu^2}$. Dieser Faktor ist stets kleiner als 1 und schrumpft $\hat{b}_i$ in Richtung $\mathbb{E}(b_i) = 0$. Die Schrumpfung ist dabei gering, wenn ν^2 im Verhältnis zu σ^2/n_i groß ist, und fällt umso größer aus, je kleiner ν^2 im Verhältnis zu σ^2/n_i ist. Insbesondere hängt die Schrumpfung vom Stichprobenumfang n_i des i-ten Individuums ab. Sind für Individuum i viele Beobachtungen vorhanden, so liegt die Schätzung für b_i nahe am Mittelwert der Residuen, sind dagegen nur wenige Beobachtungen vorhanden, so liegt die Schätzung nahe bei 0. Man spricht auch davon, dass die Schrumpfung umso geringer ausfällt, je reliabler, also verlässlicher die Daten sind.

Die Eigenschaft als Shrinkage-Schätzer findet eine weitere Interpretation in der Tatsache, dass für jede Linearkombination $\lambda'b$, $\lambda \in \mathbb{R}^q$

$$\mathrm{Var}(\lambda'\hat{b}) \leq \mathrm{Var}(\lambda'b)$$

gilt (Verbeke & Molenberghs (2000) Seite 81). Zudem lässt sich $\hat{b}$ auch als Minimax-Schätzer auffassen. Ähnlich wie in einer bayesianischen Betrachtungsweise werden in der Minimax-Schätzung A-Priori-Restriktionen in der Schätzung berücksichtigt. Unter geeigneten zusätzlichen Annahmen erhält man so ebenfalls einen Schätzer der Form (2.4). Man vergleiche hierzu auch Toutenburg (2003), Kapitel 4.13 in dem die Minimax-Schätzung im linearen Modell ausführlich behandelt wird.

Ähnlich wie für $\hat{\beta}$ kann man für $\hat{b}$ auch ohne die Normalverteilungsannahme für ε Optimalitätseigenschaften herleiten. Der Schätzer $\hat{b}$ ist bester linearer Prädiktor

(BLUP), das heißt $\hat{b}$ minimiert $\mathbb{E}((\hat{b}-b)'(\hat{b}-b))$ in der Klasse der unverzerrten linearen Schätzer. Eine allgemeine Lösung dieses Minimierungsproblems erhält man durch $\mathbb{E}(b|y)$ (siehe auch 2.1.3 über die bayesianische Betrachtungsweise des Schätzproblems). Die Eigenschaft der Unverzerrtheit ist für $\hat{b}$ jedoch anders zu verstehen als für $\hat{\beta}$, da b ja als zufällig angesehen wird. Für den festen Parameter β bedeutet Unverzerrtheit, dass $\mathbb{E}_\beta(\hat{\beta}) = \beta$ gilt, das heißt, unter der Annahme, dass β der wahre Parameter ist, ist der Erwartungswert von $\hat{\beta}$ gerade β. Für $\hat{b}$ fordert man dagegen $\mathbb{E}(\hat{b}) = \mathbb{E}(b) = 0$. Dass der oben angegebene allgemeine Schätzer $\mathbb{E}(b|y)$ diese Eigenschaft aufweist, erhält man aus dem Satz vom iterierten Erwartungswert. Die Bezeichnung von $\hat{b}$ als Prädiktor und nicht als Schätzer wird von einigen Autoren als irreführend betrachtet (siehe beispielsweise Robinson (1991) Abschnitt 7.1) und hat historische Gründe. Man vergleiche hierzu auch Robinson (1991) Abschnitt 4.4, in dem die Herleitung von $\hat{b}$ über die Problemstellung der Prädiktion für eine neue Beobachtung im allgemeinen linearen Modell beschrieben wird. Im Folgenden wird $\hat{b}$ in der Regel als Schätzer bezeichnet werden.

Die Herleitung der einzelnen Formeln zur Schätzung von β und b und Nachweise zu den Optimalitätseigenschaften der Schätzer findet man beispielsweise in Harville (1976) und Harville (1977).

2.1.3 Schätzung aus bayesianischer Sicht

Nun soll die Schätzung in linearen gemischten Modellen auch aus bayesianischer Sicht beschrieben werden. Sowohl b als auch β werden jetzt als Zufallsgrößen angesehen, unterscheiden sich jedoch in Bezug auf die gewählte Priori-Verteilung. Für β verwendet man eine nichtinformative Priori $p(\beta) \propto const$, das heißt, man nimmt an, dass über β kein Vorwissen vorliegt. Für b geht man dagegen davon aus, dass Vorwissen vorhanden ist und drückt dieses durch Verteilungsannahme $b \sim \mathrm{N}(0, Q(\nu))$ aus. Dieses Vorwissen spiegelt sich auch in der Eigenschaft von $\hat{b}$ als Shrinkage-Schätzer wieder: Die unpenalisierte Schätzung wird durch die Priori-Kovarianzmatrix hin zum Priori-Erwartungswert 0 geschrumpft. Die Schätzung stellt also einen Kompromiss zwischen Vorwissen und dem aus den Daten gewonnenen Wissen dar. Zusätzlich nimmt man noch an, dass β und b a priori unabhängig sind.

Punktschätzer in der bayesianischen Inferenz sind beispielsweise Posteriori-Erwartungswert oder Posteriori-Modus. Beide können mit Hilfe der Dichte der Posteriori-Verteilung

$$p(\beta, b|y) = \frac{p(y|\beta, b)p(\beta)p(b)}{\int p(y|\beta, b)p(\beta)p(b)d\beta db}$$

bestimmt werden. Für diese gilt aufgrund der flachen Priori-Verteilung für β

$$p(\beta, b|y) \propto p(y|\beta, b)p(b).$$

Man erhält also eine Dichte, die proportional ist zur penalisierten Likelihood (2.2) und die der gemeinsamen Verteilung von y und b entspricht. Da die Posteriori-Verteilung eine Normalverteilung ist, fallen Posteriori-Modus und Posteriori-Erwartungswert zusammen und die bayesianischen Schätzer stimmen auch mit den frequentistisch hergeleiteten Schätzern überein.

Die Optimalitätseigenschaften von $\hat{\beta}$ und $\hat{b}$ (BLUE beziehungsweise BLUP) folgen dann auch aus allgemeinen Resultaten für optimale Bayes-Schätzer. Optimal bezüglich der quadratischen Verlustfunktion $\mathbb{E}((\hat{\beta}-\beta)'(\hat{\beta}-\beta))$ ist der Posteriori-Erwartungswert $\mathbb{E}(\beta|y)$ (Rüger (1999), Seite 300/1), also der oben hergeleitete Schätzer. Dieselbe Aussage erhält man für $\hat{b}$.

2.2 Varianzparameter im linearen gemischten Modell

Nun sollen Verfahren zur Schätzung der Varianzparameter im linearen gemischten Modell vorgestellt werden. Bezeichne dazu wieder ϑ den d-dimensionalen Vektor aller Varianzparameter des Modells, das heißt, ϑ besteht, wie in Kapitel 2.1.1 definiert, aus σ^2 und einem Vektor ν von Parametern, über den $Q(\nu) = \mathrm{Var}(b)$ eineindeutig parametrisiert ist.

Im Folgenden werden die Varianzparameter als feste Parameter betrachtet und nicht wie in einem vollen Bayes-Ansatz als Zufallsgrößen modelliert. Die in Kapitel 2.1 vorgestellten Schätzer können daher in bayesianischer Betrachtungsweise als empirische Bayes-Schätzer angesehen werden. Die empirische Bayes-Schätzung unterscheidet sich von der vollen Bayes-Schätzung dahingehend, dass nicht primär interessierende Hyperparameter, wie im linearen gemischten Modell die Varianzparameter, als fest betrachtet und vorab aus den Daten geschätzt werden. Die

Schätzer der Hyperparameter werden dann in die Formeln zur Schätzung der primär interessierenden Parameter eingesetzt. Problematisch kann dieses Vorgehen sein, weil in der Posteriori-Verteilung der interessierenden Parameter nicht die zusätzliche Variabilität durch die Schätzung der Hyperparameter berücksichtigt wird. Man vergleiche beispielsweise Ruppert & Carroll (2000) für eine kurze Diskussion dieses Problems und weitergehende Literatur. Details zur vollen Bayes-Schätzung im linearen gemischten Modell findet man beispielsweise in Gamerman (1997).

Es sei noch darauf hingewiesen, dass auch der empirische Bayes-Ansatz unter Umständen als voller Bayes-Ansatz betrachtet werden kann, wenn man uneigentliche Verteilungen (siehe Rüger (1999), Seite 211-219) für die Priori-Verteilungen der Varianzparameter zulässt. Beispielsweise erhält man den Maximum-Likelihood-Schätzer für σ^2 auch als Schätzer in einem vollen Bayes-Ansatz durch die Verwendung einer Priori-Verteilung mit Dichte $p(\sigma^2) \propto const > 0$, $\sigma^2 \geq 0$, also einer flachen Priori. Diese Betrachtungsweise wird auf Seite 19 zur Charakterisierung der sich ergebenden Schätzer verwendet werden. Sie ist jedoch problematisch, weil im linearen gemischten Modell bei Verwendung uneigentlicher Priori-Verteilungen die Existenz der gemeinsamen Posteriori aller Parameter nicht garantiert ist (Hobert & Casella 1996). Die bayesianische Interpretation der Parameterschätzer kann also nur unter dem Vorbehalt der Existenz dieser Posteriori erfolgen. Als Ausweg bietet sich hier die Verwendung eigentlicher, aber an den Verlauf nichtinformativer, uneigentlicher Verteilungen angenäherter Priori-Verteilungen in einem voll-bayesianischen Ansatz an. Man vergleiche hierzu auch die Diskussion verschiedener Priori-Verteilungen für Varianzparameter in Kapitel 5.1.2.

Maximum-Likelihood-Schätzer für ϑ erhält man durch Einsetzen des Schätzers $\hat{\beta}$ in die marginale Likelihood für β und Maximierung der sich ergebenden Profile-Likelihood bezüglich ϑ. Die logarithmierte Profile-Likelihood hat (bis auf additive Konstanten) die Form

$$l(\vartheta) = -\frac{1}{2}\log|V(\vartheta)| - \frac{1}{2}(y - X\hat{\beta})'V(\vartheta)^{-1}(y - X\hat{\beta}).$$

Bei Verwendung der Profile-Likelihood zur Schätzung von ϑ bleibt der Verlust von Freiheitsgraden durch die Schätzung von $\hat{\beta}$ unberücksichtigt. Beispielsweise erhält man im gewöhnlichen linearen Modell (das als Spezialfall im linearen gemischten

Modell enthalten ist) als Maximum-Likelihood-Schätzer für σ^2

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

mit den geschätzten Residuen $\hat{\varepsilon}_i = y_i - x_i' \hat{\beta}$. Bekanntlich ist dieser Schätzer nicht erwartungstreu, sondern nach unten verzerrt. Ein erwartungstreuer Schätzer für σ^2 ist

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Diesen Schätzer zeichnet nicht nur seine Erwartungstreue aus, sondern er ist auch nach einem allgemeinen Schätzprinzip erhältlich, das im Weiteren ausführlicher beschrieben werden soll.

Wie im Beispiel ist der Maximum-Likelihood-Schätzer für ϑ in der Regel verzerrt und zwar umso stärker, je größer die Zahl der fixen Effekte p ist. Im Beispiel ist dies unmittelbar aus dem Korrekturfaktor $\frac{n}{n-p-1}$ zu erkennen, mit dem der Maximum-Likelihood-Schätzer multipliziert werden müsste, um den erwartungstreuen Schätzer zu erhalten. Dieser Faktor wird umso größer, je größer p ist. Aus diesem Grund wird der Maximum-Likelihood-Schätzer häufig als zur Schätzung von ϑ weniger geeignet betrachtet (zum Beispiel McCulloch & Searle (2001) Kapitel 6.10).

Um den Verlust an Freiheitsgraden zu berücksichtigen verwendet man eine Idee von Patterson & Thompson (1971). Diese entwickeln eine modifizierte Likelihood-Schätzung, deren Ziel die Herleitung von erwartungstreuen Schätzern für Varianzkomponenten über ein allgemeines, likelihood-basiertes Verfahren ist. Dieses Verfahren beinhaltet als Spezialfälle eine Reihe von Schätzmethoden, die beispielsweise in der Varianzanalyse verwendet werden, ist aber auch in allgemeineren Modellen anwendbar. Man vergleiche hierzu auch Abschnitt 1 aus Patterson & Thompson (1971).

In diesem modifizierten Likelihood-Ansatz wird die Inferenz für ϑ nicht über die Likelihood der gesamten Daten, sondern über die Likelihood einer Menge von Fehlerkontrasten (error contrasts) durchgeführt. Unter Fehlerkontrasten versteht man dabei Linearkombinationen $a'y$, $a \in \mathbb{R}^n$ mit $\mathbb{E}(a'y) = 0$, das heißt die Verteilung von $a'y$ ist $\mathrm{N}(0, a'V(\vartheta)a)$ und damit unabhängig von β.

Beispiele für solche Fehlerkontraste sind etwa die Residuen $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}$. Den Vektor a erhält man dabei als i-te Zeile der Residualmatrix $R = I - X(X'X)^{-1}X'$. Fasst man alle Residuen zum Vektor $\hat{\varepsilon}$ zusammen, so gilt für diesen

$$\hat{\varepsilon} \sim \mathrm{N}(0, R\sigma^2).$$

Die Residuen besitzen also insbesondere die gewünschte Eigenschaft $\mathbb{E}(\hat{\varepsilon}) = 0$. Dennoch verwendet man nicht direkt die Residuen, da die Matrix R nicht vollen Rang besitzt, die Verteilung von $\hat{\varepsilon}$ also singulär ist. Um dies zu vermeiden, wählt man eine Menge linear unabhängiger Fehlerkontraste aus, die dann eine reguläre Verteilung besitzen.

Es gibt maximal $n - p - 1$ linear unabhängige Fehlerkontraste, was aus $\mathrm{rg}(R) = n - p - 1$ folgt. Diese fasst man im Vektor $u = A'y$ zusammen. A ist dabei eine $n \times (n-p-1)$-Matrix mit vollem Spaltenrang und zum Beispiel aus der Zerlegung

$$AA' = I - X(X'X)^{-1}X' \text{ mit } A'A = I \tag{2.5}$$

erhältlich. Eine Alternative wäre die Auswahl von $n - p - 1$ linear unabhängigen Spalten aus R. Üblich ist jedoch die obige Zerlegung, da diese unabhängig von der konkreten Datensituation möglich ist.

Für die Dichte $p(u)$ der Verteilung von u gilt in diesem Fall

$$\begin{aligned} p(u) &= \left(\frac{1}{2\pi}\right)^{\frac{n-p-1}{2}} |X'X|^{\frac{1}{2}} |V(\vartheta)|^{-\frac{1}{2}} |X'V(\vartheta)^{-1}X|^{-\frac{1}{2}} \\ &\quad \cdot \exp\left[-\frac{1}{2}(y - X\hat{\beta})'V(\vartheta)^{-1}(y - X\hat{\beta})\right]. \end{aligned}$$

Diese Darstellung geht zurück auf Harville (1974) und wird in Anhang A.2 detailliert hergeleitet. Man beachte, dass die Dichte nicht in Abhängigkeit von u, sondern in Abhängigkeit von y und $\hat{\beta}$ geschrieben wird. Die Fehlerkontraste müssen daher nicht explizit zur Schätzung von ϑ bestimmt werden. Aus der Herleitung in Anhang A.2 erkennt man auch, wie u durch y und $\hat{\beta}$ festgelegt ist.

Die Dichte $p(u)$ bezeichnet man auch als Restricted-Likelihood, da die Inferenz für ϑ nicht mit allen Daten durchgeführt, sondern auf eine Menge von Fehlerkontrasten eingeschränkt wird. Man beachte, dass in der Herleitung die Matrix A aus Zerlegung (2.5) benutzt wird, die Verteilung sich aber bei Verwendung einer

anderen Matrix nur um eine Normierungskonstante ändert, solange $n - p - 1$ linear unabhängige Fehlerkontraste betrachtet werden (Verbeke & Molenberghs (2000) Kapitel 5.3).

Den Restricted-Maximum-Likelihood-Schätzer $\hat{\vartheta}_{REML}$ erhält man nun durch Maximierung der Funktion

$$\begin{aligned} l^*(\vartheta) &= -\frac{1}{2}\log(|V(\vartheta)|) - \frac{1}{2}\log(|X'V(\vartheta)^{-1}X|) - \frac{1}{2}(y - X\hat{\beta})'V(\vartheta)^{-1}(y - X\hat{\beta}) \\ &= l(\vartheta) - \frac{1}{2}\log(|X'V(\vartheta)^{-1}X|) \end{aligned}$$

bezüglich ϑ. Die Funktion $l^*(\vartheta)$ erhält man aus $p(u)$ durch Logarithmieren und Vernachlässigung von additiven Konstanten. Insbesondere ist $l^*(\vartheta)$ also unabhängig von der speziellen Wahl von A.

Die Restricted-Log-Likelihood $l^*(\vartheta)$ unterscheidet sich nur durch den Term $-\frac{1}{2}\log(|X'V(\vartheta)^{-1}X|)$ von der Log-Likelihood $l(\vartheta)$. Da der zusätzliche Term unabhängig ist von β, wird häufig auch

$$l^*(\beta, \vartheta) = -\frac{1}{2}\log(|V(\vartheta)|) - \frac{1}{2}\log(|X'V(\vartheta)^{-1}X|) - \frac{1}{2}(y - X\beta)'V(\vartheta)^{-1}(y - X\beta)$$

als Restricted-Log-Likelihood bezeichnet und zur simultanen Schätzung von β und ϑ verwendet (Verbeke & Molenberghs (2000) Kapitel 5.3.3).

Eine alternative Herleitung des REML-Schätzers geht auf Harville (1974) zurück und bietet die Möglichkeit einer interessanten Charakterisierung dieses Schätzers. Harville geht von der bayesianischen Betrachtungsweise des Schätzproblems aus, das heißt sowohl β und b als auch ϑ werden als Zufallsgrößen betrachtet, wobei β mit einer flachen Priori und b mit der üblichen Normalverteilungspriori $\mathrm{N}(0, Q(\nu))$ versehen wird. Für ϑ nimmt man die oben beschriebene flache Priori mit Dichte $p(\vartheta) \propto const > 0$ an und zusätzlich werden alle Parameter als unabhängig betrachtet. Die Inferenz für ϑ wird nun über die marginale Posteriori für ϑ durchgeführt. Im Gegensatz zum Vorgehen in 2.1.2 wird dafür nicht nur die Marginalverteilung bezüglich b gebildet, sondern auch bezüglich β. Die marginale

Posteriori-Verteilung für ϑ wird also hergeleitet über

$$\begin{aligned} p(\vartheta|y) &= \int p(\beta, b, \vartheta|y) dbd\beta \\ &= \int \frac{p(y|\beta, b, \vartheta)p(b)p(\beta)p(\vartheta)}{\int p(y|\beta, b, \vartheta)p(b)p(\beta)p(\vartheta)dbd\beta d\vartheta} dbd\beta \\ &\propto \int p(y|\beta, b, \vartheta)p(b)dbd\beta \qquad (2.6) \\ &= \int p(y|\beta, \vartheta)d\beta, \end{aligned}$$

wobei $p(y|\beta, \vartheta)$ mit (2.1) übereinstimmt. Nach dem Satz von Bayes gilt zusammen mit der Unabhängigkeit von β und ϑ

$$p(\beta|y, \vartheta) = \frac{p(y|\beta, \vartheta)p(\beta)}{\int p(y|\beta, \vartheta)p(\beta)d\beta} = \frac{p(y|\beta, \vartheta)}{\int p(y|\beta, \vartheta)d\beta}, \qquad (2.7)$$

wobei die letzte Gleichheit aus der diffusen Priori für β folgt. Beachtet man noch, dass $p(\beta|y, \vartheta)$ mit (A.12) übereinstimmt, so erhält man aus Anhang A.2

$$p(\vartheta|y) \propto \int p(y|\beta, \vartheta)d\beta = \frac{p(y|\beta, \vartheta)}{p(\beta|y, \vartheta)} \propto p(u).$$

Damit ist die Maximierung der marginalen Posteriori $p(\vartheta|y)$ äquivalent zur REML-Schätzung.

Es ergibt sich die folgende Charakterisierung von ML- und REML-Schätzer: $\hat{\vartheta}_{ML}$ ist die ϑ-Komponente des Modus der Posteriori-Verteilung von ϑ und β, während $\hat{\vartheta}_{REML}$ der Modus der marginalen Posteriori-Verteilung für ϑ ist. Es sei nochmals darauf hingewiesen, dass diese Interpretation nur dann zulässig ist, wenn die gemeinsame Posteriori $p(\beta, b, \vartheta|y)$ existiert, weil sonst insbesondere die Proportionalität in (2.6) und die Gleichheit in (2.7) nicht mehr zutreffen. Hobert & Casella (1996) geben Bedingungen an, die die Existenz der gemeinsamen Posteriori sichern.

Fraglich scheint, ob durch den Übergang von y auf u Information über ϑ verloren geht. Dies ist jedoch nicht der Fall, wie Harville (1977) zeigt, da die Fehlerkontraste u marginal suffizient sind für ϑ, wie in Sprott (1975) definiert. Sprott verallgemeinert hier den üblichen Suffizienzbegriff auf Situationen mit Störparametern, wie dem Parameter β bei der Schätzung von Varianzkomponenten im linearen gemischten Modell. Informell bedeutet die marginale Suffizienz von u,

dass in Abwesenheit von a priori erhältlicher Information über β keine Information verloren geht, wenn man die Inferenz für ϑ nur auf die Likelihood von u stützt. Dies zeigt sich auch darin, dass die Profile-Log-Likelihood $l(\vartheta)$ de facto nur von $n-p-1$ Fehlerkontrasten abhängt, also im REML-Ansatz keine Information verloren geht, die im ML-Ansatz tatsächlich verwendet wird. Für eine ausführlichere Diskussion zur Rechtfertigung des REML-Ansatzes vergleiche man Harville (1977).

Da $l^*(\vartheta)$ nichtlinear in ϑ ist, muss $\hat{\vartheta}_{REML}$ iterativ bestimmt werden. Dazu bieten sich der Newton-Raphson-Algorithmus beziehungsweise dessen Modifikation zum Fisher-Scoring an. Zur numerischen Durchführung benötigt man die folgenden Größen:

Die Score-Funktion

$$s^*(\vartheta) = \frac{\partial l^*(\vartheta)}{\partial \vartheta} = (s_j^*(\vartheta))_{j=1,\ldots,d},$$

die beobachtete Fisher-Information

$$F_{obs}^*(\vartheta) = -\frac{\partial s^*(\vartheta)}{\partial \vartheta} = -\frac{\partial^2 l^*(\vartheta)}{\partial \vartheta \partial \vartheta'} = (F_{obs,jk}^*(\vartheta))_{j,k=1,\ldots,d}$$

sowie die erwartete Fisher-Information

$$F^*(\vartheta) = \mathbb{E}(F_{obs}^*(\vartheta)) = ((F_{jk}^*(\vartheta))_{j,k=1,\ldots,d}.$$

Die Score-Funktion $s^*(\vartheta)$ ist ein $d \times 1$-Vektor mit den Elementen

$$\begin{aligned} s_j^*(\vartheta) &= \frac{\partial l^*(\vartheta)}{\partial \vartheta_j} \\ &= -\frac{1}{2}\,\mathrm{spur}\left(P(\vartheta)\frac{\partial V(\vartheta)}{\partial \vartheta_j}\right) + \frac{1}{2}(y - X\hat{\beta})' V(\vartheta)^{-1} \frac{\partial V(\vartheta)}{\partial \vartheta_j} V(\vartheta)^{-1}(y - X\hat{\beta}) \\ &= -\frac{1}{2}\,\mathrm{spur}\left(P(\vartheta)\frac{\partial V(\vartheta)}{\partial \vartheta_j}\right) \\ &\quad + \frac{1}{2}(y - X\hat{\beta} - Z\hat{b})' W(\vartheta) \frac{\partial V(\vartheta)}{\partial \vartheta_j} W(\vartheta)(y - X\hat{\beta} - Z\hat{b}). \end{aligned} \tag{2.8}$$

Wie in 2.1.2 sind $W(\vartheta)$ und $P(\vartheta)$ definiert durch $W(\vartheta) = (\mathrm{Var}(y|b))^{-1} = \frac{1}{\sigma^2} I_n$ und

$$\begin{aligned} P(\vartheta) &= V(\vartheta)^{-1} - V(\vartheta)^{-1} X (X' V(\vartheta)^{-1} X)^{-1} X' V(\vartheta)^{-1} \\ &= W(\vartheta) - W(\vartheta)(X, Z) H^{-1} (X, Z)' W(\vartheta) \end{aligned}$$

mit

$$H = \begin{pmatrix} X'W(\vartheta)X & X'W(\vartheta)Z \\ Z'W(\vartheta)X & Z'W(\vartheta)Z + Q^{-1}(\nu) \end{pmatrix}.$$

Die jeweils zweiten Ausdrücke für $s^*(\vartheta)$ und $P(\vartheta)$, die nicht mehr von $V(\vartheta)^{-1}$ abhängen, ergeben sich aus einem Wechsel von der marginalen in die bedingte Betrachtungsweise des linearen gemischten Modells. Diese Darstellungen sind numerisch vorteilhafter, da sie die Inversion der $n \times n$-Matrix $V(\vartheta)$ vermeiden.

$F^*_{obs}(\vartheta)$ und $F^*(\vartheta)$ sind $d \times d$-Matrizen mit Einträgen

$$\begin{aligned} F^*_{obs,jk}(\vartheta) &= -\frac{\partial^2 l^*(\vartheta)}{\partial\vartheta_j d\vartheta_k} \\ &= \frac{1}{2}\,\text{spur}\left(P(\vartheta)\frac{\partial^2 V(\vartheta)}{\partial\vartheta_j d\vartheta_k} - P(\vartheta)\frac{\partial V(\vartheta)}{\partial\vartheta_j}P(\vartheta)\frac{\partial V(\vartheta)}{\partial\vartheta_k}\right) \\ &\quad -\frac{1}{2}(y - X\hat{\beta})'V(\vartheta)^{-1}\left(\frac{\partial^2 V(\vartheta)}{\partial\vartheta_j d\vartheta_k} - 2\frac{\partial V(\vartheta)}{\partial\vartheta_j}P(\vartheta)\frac{\partial V(\vartheta)}{\partial\vartheta_k}\right)V(\vartheta)^{-1}(y - X\hat{\beta}) \end{aligned}$$

beziehungsweise

$$\begin{aligned} F^*_{jk}(\vartheta) &= \mathbb{E}(F^*_{obs,jk}(\vartheta)) \\ &= \frac{1}{2}\,\text{spur}\left(P(\vartheta)\frac{\partial V(\vartheta)}{\vartheta_j}P(\vartheta)\frac{\partial V(\vartheta)}{\vartheta_k}\right). \end{aligned} \tag{2.9}$$

Details zur Herleitung der einzelnen Größen findet man teilweise in McCulloch & Searle (2001) Seite 178-184 und ausführlich in Anhang A.3 bis A.5. Zur konkreten Bestimmung der angegebenen Größen müssen noch die Ableitungen $\frac{\partial V(\vartheta)}{\partial\vartheta_j}$ und $\frac{\partial^2 V(\vartheta)}{\partial\vartheta_j\partial\vartheta_k}$ $j,k = 1,\ldots,d$ bestimmt werden. Für die Ableitung bezüglich σ^2 erhält man beispielsweise

$$\frac{\partial V(\vartheta)}{\partial\sigma^2} = I_n.$$

Die übrigen Ableitungen (nach den Elementen von ν) hängen von der speziellen Form von $Q(\nu)$ ab und müssen für die einzelnen Modelle gezielt bestimmt werden. Gilt beispielsweise $Q(\nu) = \nu^2 I_q$, so erhält man

$$\frac{\partial V(\vartheta)}{\partial\nu^2} = ZZ'.$$

Der REML-Schätzer $\hat{\vartheta}_{REML}$ erfüllt als globales Maximum der Restricted-Log-Likelihood $s^*(\hat{\vartheta}_{REML}) = 0$ und $F^*_{obs}(\hat{\vartheta}_{REML})$ positiv definit. Die im Folgenden

vorgestellten Algorithmen sind jedoch nur in der Lage, ein lokales Maximum über eine Nullstelle der Score-Funktion zu finden. Geht man von einer konvexen Form der Restricted-Log-Likelihood aus, so fallen lokales und globales Maximum zusammen.

Die Grundidee des Newton-Raphson-Algorithmus besteht darin, $s^*(\vartheta)$ in eine Taylorreihe um $\tilde{\vartheta}$ zu entwickeln, wobei $\tilde{\vartheta}$ eine geeignete Näherung von $\hat{\vartheta}$ sei. Man erhält so

$$\begin{aligned} 0 &= s^*(\hat{\vartheta}) \\ &\approx s^*(\tilde{\vartheta}) - F^*_{obs}(\tilde{\vartheta})(\hat{\vartheta} - \tilde{\vartheta}). \end{aligned}$$

Durch Umformungen ergibt sich für $\hat{\vartheta}$ der Ausdruck

$$\hat{\vartheta} = \tilde{\vartheta} + (F^*_{obs}(\tilde{\vartheta}))^{-1} s^*(\tilde{\vartheta}).$$

Ausgehend von einem Startwert $\hat{\vartheta}^{(0)}$ bestimmt man nun iterativ Werte

$$\hat{\vartheta}^{(k+1)} = \hat{\vartheta}^{(k)} + (F^*_{obs}(\hat{\vartheta}^{(k)}))^{-1} s^*(\hat{\vartheta}^{(k)}),$$

bis ein geeignetes Abbruchkriterium erfüllt ist. Als solches kann man beispielsweise die relative Veränderung der Schätzwerte $\frac{||\hat{\vartheta}^{(k+1)} - \hat{\vartheta}^{(k)}||}{||\hat{\vartheta}^{(k)}||}$ verwenden und abbrechen, wenn diese einen bestimmten Wert unterschreitet.

Häufig wird aufgrund der einfacheren Berechnung das Fisher-Scoring dem Newton-Raphson-Algorithmus vorgezogen. Dabei wird die beobachtete Fisher-Information durch die erwartete Fisher-Information ersetzt. Man erhält so die Iterationsvorschrift

$$\hat{\vartheta}^{(k+1)} = \hat{\vartheta}^{(k)} + (F^*(\hat{\vartheta}^{(k)}))^{-1} s^*(\hat{\vartheta}^{(k)}).$$

Zu beachten ist sowohl beim Newton-Raphson-Algorithmus als auch beim Fisher-Scoring die Parametrisierung von $V(\vartheta) = \mathrm{Var}(y)$ über den Parametervektor ϑ. Nimmt man beispielsweise an, dass $Q(\nu)$ eine Diagonalmatrix mit Elementen $\nu_1^2, \ldots, \nu_q^2$ ist, so wäre eine mögliche Parametrisierung $\vartheta = (\sigma^2, \nu_1^2, \ldots, \nu_q^2)$. Diese Parametrisierung weist aber einen entscheidenden Nachteil auf: Für jeden der Parameter gilt die Beschränkung auf Werte aus dem Intervall $[0, \infty)$. Diese Beschränkung wird jedoch bei der iterativen Berechnung über den Newton-Raphson-Algorithmus oder Fisher-Scoring nicht berücksichtigt. Insbesondere bei zufälligen

Effekten mit kleinen Varianzen oder der Modellierung ‚überflüssiger' zufälliger Effekte können so negative Varianzen als REML-Schätzer entstehen. Wie Lindstrom & Bates (1988) feststellen, ist es jedoch häufig notwendig, zunächst ein überparametrisiertes Modell zu schätzen, um von diesem ausgehend zu einem reduzierten Modell zu gelangen. Gerade bei der Schätzung solcher überparametrisierter Modelle kann es dann zur Schätzung negativer Varianzen kommen.

Da die Nebenbedingungen, die bei der Schätzung von ϑ zu beachten sind, nichtlinear in ϑ sind, können sie nicht durch Standardverfahren zur Lösung von Gleichungssystemen unter Nebenbedingungen im Schätzverfahren berücksichtigt werden. Stattdessen ist es in der Regel möglich, $V(\vartheta)$ geeigneter zu parametrisieren. Im obigen Beispiel kann man etwa die Parametrisierung $\vartheta = (\sigma, \nu_1, \ldots, \nu_q)$ wählen, wobei mit σ beziehungsweise ν_j nicht die zugehörigen Standardabweichungen gemeint sind, da für diese wieder entsprechende Einschränkungen des Parameterraums zu beachten wären, sondern beliebige, also positive oder negative Wurzeln der entsprechenden Parameter. Betrachtet man eine unstrukturierte Kovarianzmatrix $Q(\nu)$ für die zufälligen Effekte, so kann man $Q(\nu)$ über die Elemente des Cholesky-Faktors von $Q(\nu)$ parametrisieren. Ein weiterer Vorteil dieser Parametrisierung ohne Restriktionen ist nach Lindstrom & Bates (1988), dass der Algorithmus schneller konvergiert als in der ‚naiven' Parametrisierung. In Pourahmadi (1999) und Pourahmadi (2000) findet man eine ausführlichere Diskussion und weitergehende Informationen zu geeigneten Parametrisierungen von Kovarianzmatrizen.

Bei der konkreten Schätzung von $\hat{\beta}$, $\hat{b}$ und $\hat{\vartheta}$ ergibt sich nun das Problem, dass $\hat{\beta}$ und $\hat{b}$ von $\hat{\vartheta}$ abhängen und umgekehrt $\hat{\vartheta}$ von $\hat{\beta}$ und $\hat{b}$ abhängt. Daher bestimmt man ausgehend von Startwerten für alle Parameter abwechselnd neue Schätzer für $\hat{\beta}$ und $\hat{b}$ beziehungsweise $\hat{\vartheta}$ in Abhängigkeit von den jeweils aktuellen Schätzungen der übrigen Parameter. Genauer verwendet man den folgenden Algorithmus:

Algorithmus 1 (Schätzung im linearen gemischten Modell)

(i) Wähle Startwerte $\hat{\beta}^{(0)}$, $\hat{b}^{(0)}$ und $\hat{\vartheta}^{(0)}$, die maximale Iterationszahl *maxit* sowie ein Abbruchkriterium ϵ, beispielsweise $\epsilon = 0.00001$. Setze $k = 0$.

(ii) Bestimme $\hat{\beta}^{(k+1)}$ und $\hat{b}^{(k+1)}$ durch Lösen des Gleichungssystems

$$\begin{pmatrix} X'W(\hat{\vartheta}^{(k)})X & X'W(\hat{\vartheta}^{(k)})Z \\ Z'W(\hat{\vartheta}^{(k)})X & Z'W(\hat{\vartheta}^{(k)})Z + Q(\hat{\nu}^{(k)})^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'W(\hat{\vartheta}^{(k)})y \\ Z'W(\hat{\vartheta}^{(k)})y \end{pmatrix}.$$

(iii) Bestimme $\hat{\vartheta}^{(k+1)}$ als

$$\hat{\vartheta}^{(k+1)} = \hat{\vartheta}^{(k)} + F^*(\hat{\vartheta}^{(k)})^{-1} s^*(\hat{\vartheta}^{(k)}),$$

wobei $s^*(\hat{\vartheta}^{(k)})$ und $F^*(\hat{\vartheta}^{(k)})$ aus (2.8) und (2.9) in Abhängigkeit von $\hat{\beta}^{(k+1)}$ und $\hat{b}^{(k+1)}$ berechnet werden.

(iv) Bestimme die Abbruchkriterien

$$d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) = \frac{||\hat{\vartheta}^{(k+1)} - \hat{\vartheta}^{(k)}||}{||\hat{\vartheta}^{(k)}||}$$

und

$$d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) = \frac{||\hat{\delta}^{(k+1)} - \hat{\delta}^{(k)}||}{||\hat{\delta}^{(k)}||}$$

mit $\delta = (\beta', b')'$.

Falls $d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) > \epsilon$ oder $d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) > \epsilon$ und $k < maxit$, setze $k = k + 1$ und gehe zurück zu (ii).
Falls $d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) < \epsilon$ und $d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) < \epsilon$ und $k < maxit$, so sind $\hat{\beta}^{(k+1)}$, $\hat{b}^{(k+1)}$ und $\hat{\vartheta}^{(k+1)}$ die endgültigen Schätzer.
Falls $k = maxit$, breche die Schätzung ohne Ergebnis ab.

Problematisch bei der Bestimmung der Score-Funktion und der Fisher-Informations-Matrizen ist die Abhängigkeit vom Stichprobenumfang n. Auch wenn die Inversion von $V(\vartheta)$ zur Berechnung von Score-Funktion und erwarteter Fisher-Informationsmatrix durch die Verwendung der Formeln, die sich aus der bedingten Betrachtungsweise ergeben und die nur von $W(\vartheta)$ abhängen, vermeiden lässt, so ist doch die Berechnung der Spuren der $n \times n$-Matrizen $P(\vartheta)\frac{\partial V(\vartheta)}{\partial \vartheta_j}$ zur Bestimmung der Score-Funktion und $P(\vartheta)\frac{\partial V(\vartheta)}{\partial \vartheta_j} P(\vartheta)\frac{\partial V(\vartheta)}{\partial \vartheta_k}$ zur Bestimmung der erwarteten Fisher-Information notwendig. Diese Spuren hängen darüber hinaus von den aktuellen Schätzwerten $\hat{\vartheta}^{(k)}$ ab, so dass sie in jeder Iteration neu bestimmt werden müssen. Die Anwendung der vorgestellten Modelle ist daher aufgrund von Speicherplatzproblemen derzeit auf Modelle mit Stichprobenumfängen von circa 2000-3000 Beobachtungen beschränkt.

Als Alternative zur Bestimmung der Varianzparameter über den Newton-Raphson-Algorithmus oder Fisher-Scoring bietet sich der EM-Algorithmus an, wie ursprünglich von Laird & Ware (1982) vorgeschlagen. Lindstrom & Bates (1988) vergleichen Newton-Raphson- und EM-Algorithmus und kommen zu

dem Schluss, dass der Newton-Raphson-Algorithmus im Allgemeinen dem EM-Algorithmus vorzuziehen ist. Obwohl ein einzelner Iterationsschritt des Newton-Raphson-Algorithmus numerisch aufwändiger ist, als ein Schritt des EM-Algorithmus, weist der Newton-Raphson-Algorithmus häufig eine wesentlich geringere Iterationszahl auf. Zudem ist nach Konvergenz die beobachtete Fisher-Information (Newton-Raphson) beziehungsweise die erwartete Fisher-Information (Fisher-Scoring) direkt erhältlich. Ein Nachteil des Newton-Raphson-Algorithmus ist dagegen, dass er nicht gegen ein lokales Maximum der Log-Likelihood konvergieren muss, während die Konvergenz beim EM-Algorithmus theoretisch garantiert ist.

Zur Schätzung der Varianzparameter im Rahmen der Simulationsstudien in Kapitel 5 und der Datenanalysen in Kapitel 6 wird eine auf dem Fisher-Scoring-Algorithmus basierende Implementation verwendet, weil dieser zum einen einfacher zu implementieren ist als der Newton-Raphson-Algorithmus und zum anderen mit Hilfe der erwarteten Fisher-Information asymptotische Konfidenzintervalle für die geschätzten Varianzparameter konstruiert werden können.

2.3 Generalisierte lineare gemischte Modelle

2.3.1 Modell

In vielen Datensituationen ist eine Analyse mit Hilfe des linearen Modells nicht möglich, weil die abhängige Variable nicht als (approximativ) normalverteilt angenommen werden kann. Beispiele hierfür sind etwa binäre Daten, Häufigkeiten oder nichtnegative Daten. Um auch solche Daten in vergleichbarer Weise wie im linearen Modell analysieren zu können, verwendet man das von Nelder & Wedderburn (1972) eingeführte generalisierte lineare Modell. Zu einer ausführlicheren Beschreibung generalisierter linearer Modelle vergleiche man beispielsweise Fahrmeir & Tutz (2001) oder McCullagh & Nelder (1989). Hier soll nun nicht näher auf das generalisierte lineare Modell eingegangen werden, sondern sofort das allgemeinere Modell mit zufälligen Effekten beschrieben werden. Eine Einführung zu generalisierten linearen gemischten Modellen findet man etwa in Diggle, Liang & Zeger (1994) Kapitel 9 oder Fahrmeir & Tutz (2001) Kapitel 7.

Wie im generalisierten linearen Modell benötigt man zur Spezifikation eines generalisierten linearen gemischten Modells eine Verteilungsannahme und eine struk-

turelle Annahme. Zusätzlich sind noch Annahmen über die zufälligen Effekte zu treffen.

Im linearen gemischten Modell nimmt man an, dass $y_i|b$ normalverteilt ist mit Erwartungswert $\mu_i = \mathbb{E}(y_i|b) = x_i'\beta + z_i'b$. Diese Annahme wird im generalisierten linearen gemischten Modell ersetzt durch die Annahme, dass die Verteilung von $y_i|b$ aus einer einfachen Exponentialfamilie stammt. Die Dichte von $y_i|b$ soll sich also schreiben lassen als

$$p(y_i|b) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i + c(y_i, \phi, \omega_i)\right), \tag{2.10}$$

wobei θ_i der so genannte natürliche Parameter ist, ϕ der Skalen- oder Dispersionsparameter und $b(\cdot)$ und $c(\cdot)$ bekannte Funktionen, die vom Typ der Exponentialfamilie abhängen. Die $\omega_i \geq 0$ sind Gewichte, die insbesondere bei gruppierten Daten verwendet werden, prinzipiell aber beliebig wählbar sind. Für ungruppierte Daten gilt meist $\omega_i = 1$, für gruppierte Daten $\omega_i = n_i$, falls der Gruppenmittelwert als abhängige Variable betrachtet wird, und $\omega_i = 1/n_i$, falls die Summe der Beobachtungen in einer Gruppe als abhängige Variable betrachtet wird. Typische Beispiele für Exponentialfamilien sind die Binomialverteilung, die Poissonverteilung und die Gammaverteilung. Auch die Normalverteilung gehört zu den Exponentialfamilien, so dass sich das lineare Modell als Spezialfall eines generalisierten linearen Modells und das lineare Modell mit zufälligen Effekten als Spezialfall eines generalisierten linearen gemischten Modells auffassen lässt. Man vergleiche Fahrmeir & Tutz (2001) Seite 21 für eine Übersicht über die Darstellung verschiedener Verteilungen als Exponentialfamilien.

Zusätzlich nimmt man an, dass die einzelnen Beobachtungen bedingt auf b unabhängig sind. Wie im Normalverteilungsfall gilt damit wieder, dass die Beobachtungen bedingt unabhängig sind, während sie marginal abhängig sind.

Die strukturelle Annahme $\mu_i = \mathbb{E}(y_i|b) = x_i'\beta + z_i'b$ aus dem linearen Modell mit zufälligen Effekten wird im generalisierten linearen gemischten Modell ersetzt durch die Annahme $\mu_i = h(x_i'\beta + z_i'b) = h(\eta_i)$ beziehungsweise $\eta_i = g(\mu_i)$. Die Funktion $h(\cdot)$ soll dabei eine bekannte, hinreichend glatte und eineindeutige Funktion sein, die Responsefunktion genannt wird. Die Funktion $g(\cdot)$ ist die Umkehrfunktion zu $h(\cdot)$ und wird Linkfunktion genannt. Die Summe von fixem und zufälligem Effekt η_i bezeichnet man als linearen Prädiktor.

Für b nimmt man wie im Normalverteilungsfall an, dass

$$b \sim \mathrm{N}(0, Q(\nu)) \tag{2.11}$$

gilt.

Von Interesse ist häufig noch, dass man Erwartungswert und Varianz von $y_i|b$ aus den Komponenten der Exponentialfamilie erhalten kann. Es gilt nämlich

$$\mu_i = \mathbb{E}(y_i|b) = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta}$$

und

$$\mathrm{Var}(y_i|b) = \phi v(\mu_i)/\omega_i,$$

wobei $v(\mu_i) = b''(\theta_i)$ die so genannte Varianzfunktion der entsprechenden Exponentialfamilie bezeichnet (siehe Pruscha (2000) Kapitel II 2.3 für Beweise der Aussagen). Aus der ersten Beziehung erkennt man auch, dass der natürliche Parameter eine Funktion des Erwartungswertes ist, das heißt, es gilt $\theta_i = \theta(\mu_i)$.

Für jede Exponentialfamilie existiert eine so genannte natürliche Linkfunktion, die den natürlichen Parameter direkt mit dem linearen Prädiktor verbindet. Für diese gilt $g(\mu_i) = \theta(\mu_i) = \eta_i$. Die Verwendung der natürlichen Link-Funktion vereinfacht die zur Schätzung notwendigen Formeln und ist daher besonders aus mathematischer Sicht vorteilhaft.

2.3.2 Schätzung aus frequentistischer Sicht

Zunächst soll wieder die Schätzung der Modellparameter β und b aus frequentistischer Sicht beschrieben werden. Ist man nur an der Schätzung von β interessiert, so kann man diese theoretisch wie im Normalverteilungsfall über die marginale Likelihood

$$L(\beta) = \int p(y|b)p(b)db$$

durchführen. Im Gegensatz zum Normalverteilungsfall lässt sich dieses Integral jedoch im Allgemeinen nicht analytisch lösen. Stattdessen muss die Integration numerisch durchgeführt werden, was nur für eine geringe Zahl von zufälligen Effekten möglich ist.

Einfacher ist dagegen die simultane Schätzung von β und b. Dazu verwendet man wieder die gemeinsame Verteilung von y und b als Likelihood, das heißt,

man erhält $\hat{\beta}$ und $\hat{b}$ durch Maximieren von $p(y|b)p(b)$. Dies ist wieder äquivalent zur Bestimmung des Posteriori-Modus bei Verwendung einer flachen Priori für β, wie im nächsten Abschnitt gezeigt wird. Algorithmische Details zur Maximierung von $p(y|b)p(b)$ werden ebenfalls im nächsten Abschnitt vorgestellt.

2.3.3 Schätzung aus bayesianischer Sicht

Im Folgenden soll das Schätzproblem wieder aus einer bayesianischen Perspektive betrachtet werden, das heißt, sowohl β als auch b werden als Zufallsgrößen betrachtet. Dabei nimmt man für β wieder eine flache Priori-Verteilung und für b die Normalverteilung aus (2.11) als Priori-Verteilung an, so dass sich die Posteriori für $\delta = (\beta', b')'$ bestimmen lässt als

$$p(\delta|y) = \frac{p(y|\delta)p(\delta)}{\int p(y|\delta)p(\delta)d\delta}$$

mit $p(\delta) = p(\beta)p(b) \propto p(b)$. Als Punktschätzer bieten sich nun Erwartungswert oder Modus der Posteriori-Verteilung an. Der Posteriori-Erwartungswert lässt sich berechnen durch

$$\mathbb{E}(\delta|y) = \int \delta p(\delta|y)d\delta = \frac{\int \delta p(y|\delta)p(\delta)d\delta}{\int p(y|\delta)p(\delta)d\delta}.$$

Dazu müssen jeweils die in der Regel hochdimensionalen Integrale in Zähler und Nenner ausgewertet werden, was außer im Normalverteilungsfall nicht analytisch möglich ist. Somit ergeben sich Beschränkungen der Dimension von δ, die eine generelle Verwendung des Posteriori-Erwartungswertes als Schätzer behindern. Eine Alternative zur numerischen Integration bieten Markov-Chain-Monte-Carlo-Verfahren, mit deren Hilfe Zufallszahlen aus der Posteriori $p(\delta|y)$ gezogen werden können. Der Posteriori-Erwartungswert lässt sich dann, mit im Prinzip beliebiger Genauigkeit, durch das arithmetische Mittel der realisierten Stichprobe schätzen (vergleiche Gamerman (1997)).

Der Modus der Posteriori-Verteilung lässt sich dagegen ohne numerische Integration oder MCMC-Verfahren über einen modifizierten Fisher-Scoring-Algorithmus bestimmen. Diese Möglichkeit ergibt sich aus der Tatsache, dass die Posteriori $p(\delta|y)$ proportional ist zur Dichte $p(y|\delta)p(\delta)$ der gemeinsamen Verteilung von y und δ. Die Normierungskonstante $\int p(y|\delta)p(\delta)d\delta$ muss zur Maximierung von $p(\delta|y)$ nicht bekannt sein.

Äquivalent zur Maximierung der Posteriori ist die Maximierung der logarithmierten Posteriori $\log(p(y|\delta)) + \log(p(\delta))$. Beim ersten Term handelt es sich um die Log-Likelihood für δ, der zweite Term reduziert sich für bekanntes ϑ und unter der Vernachlässigung konstanter Terme zu dem in b quadratischen Strafterm $-\frac{1}{2}b'Q(\nu)^{-1}b$. Insgesamt maximiert man also

$$l(\delta) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i - \frac{1}{2}b'Q(\nu)^{-1}b. \tag{2.12}$$

Analog zum Vorgehen im generalisierten linearen Modell bestimmt man zunächst die Score-Funktion, also die Ableitung der Log-Posteriori nach δ. Wegen des Strafterms, der allein b betrifft, ist es sinnvoll, die Ableitungen nach β und b getrennt zu betrachten. Man erhält (Fahrmeir & Tutz (2001), Seite 298/99) $s(\delta) = \partial l(\delta)/\partial\delta = (s_\beta(\delta)', s_b(\delta)')'$ mit

$$s_\beta(\delta) = \frac{\partial l(\delta)}{\partial\beta} = X'D(\delta)\Sigma(\delta)^{-1}(y - \mu(\delta)) \tag{2.13}$$

und

$$s_b(\delta) = \frac{\partial l(\delta)}{\partial b} = Z'D(\delta)\Sigma(\delta)^{-1}(y - \mu(\delta)) - Q(\nu)^{-1}b \tag{2.14}$$

wobei

$$D(\delta) = \mathrm{diag}(D_i(\delta)) = \mathrm{diag}(\partial h(\eta_i)/\partial\eta) \tag{2.15}$$

und

$$\Sigma(\delta) = \mathrm{Var}(y|\delta) = \mathrm{diag}(\sigma_i^2(\delta)) = \mathrm{diag}(\phi v(\mu_i)/\omega_i) \tag{2.16}$$

gelten. Man beachte, dass $\Sigma(\delta)$ auch vom Skalenparameter ϕ abhängt. Diese Abhängigkeit soll jedoch in diesem Abschnitt notationell unterdrückt werden, da ja von gegebenen Varianzparametern ausgegangen wird. Bei der Schätzung der Varianzkomponenten im nächsten Abschnitt wird diese Abhängigkeit dann wieder ein Rolle spielen.

Für die erwartete Fisher-Information erhält man

$$F(\delta) = \begin{pmatrix} F_{\beta\beta}(\delta) & F_{\beta b}(\delta) \\ F_{b\beta}(\delta) & F_{bb}(\delta) \end{pmatrix}$$

mit

$$F_{\beta\beta}(\delta) = X'D(\delta)\Sigma(\delta)^{-1}D(\delta)X$$

$$F_{\beta b}(\delta) = F_{b\beta}(\delta)' = X'D(\delta)\Sigma(\delta)^{-1}D(\delta)Z$$

$$F_{bb}(\delta) = Z'D(\delta)\Sigma(\delta)^{-1}D(\delta)Z + Q(\nu)^{-1}.$$

Der Posteriori-Modus-Schätzer $\hat{\delta}$ erfüllt $s(\hat{\delta}) = 0$ und lässt sich iterativ bestimmen durch

$$\hat{\delta}^{(k+1)} = \hat{\delta}^{(k)} + F(\hat{\delta}^{(k)})^{-1}s(\hat{\delta}^{(k)}).$$

Mit Hilfe der Arbeitsbeobachtungen (working observations)

$$\tilde{y}^{(k)} = \tilde{y}(\hat{\delta}^{(k)}) = X\hat{\beta}^{(k)} + Z\hat{b}^{(k)} + D(\hat{\delta}^{(k)})^{-1}(y - \mu(\hat{\delta}^{(k)})) \tag{2.17}$$

erhält man $\hat{\delta}^{(k+1)}$ auch durch das Lösen des Gleichungssystems

$$\begin{pmatrix} X'W(\hat{\delta}^{(k)})X & X'W(\hat{\delta}^{(k)})Z \\ Z'W(\hat{\delta}^{(k)})X & Z'W(\hat{\delta}^{(k)})Z + Q(\nu)^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta}^{(k+1)} \\ \hat{b}^{(k+1)} \end{pmatrix} = \begin{pmatrix} X'W(\hat{\delta}^{(k)})\tilde{y}^{(k)} \\ Z'W(\hat{\delta}^{(k)})\tilde{y}^{(k)} \end{pmatrix}$$

mit

$$W(\delta) = \mathrm{diag}(w_i(\delta)) = D(\delta)\Sigma(\delta)^{-1}D(\delta). \tag{2.18}$$

Man beachte auch hier wieder die Abhängigkeit der Matrix $W(\delta)$ vom Skalenparameter ϕ über $\Sigma(\delta)$, die ebenfalls notationell unterdrückt wird. Im Spezialfall mit normalverteiltem Response erhält man für die Gewichte $W(\delta)$ die Matrix $W(\vartheta) = \frac{1}{\sigma^2}I$ aus Abschnitt 2.1.2 zurück.

Wie man sieht, kann $\hat{\delta}$ durch das wiederholte Schätzen eines linearen gemischten Modells für $\tilde{y}$ bestimmt werden. Im Gegensatz zu Kapitel 2.1 muss man jedoch für ε die $\mathrm{N}(0, W(\delta)^{-1})$-Verteilung annehmen, deren Kovarianzmatrix in jeder Iteration verändert wird. Die Schätzung von $\hat{\delta}$ über das iterative Schätzen eines linearen gemischten Modells für die Arbeitsbeobachtungen $\tilde{y}$ verallgemeinert die iterativ gewichtete Kleinste-Quadrate-Schätzung, die häufig zur Bestimmung eines generalisierten linearen Modells verwendet wird (vergleiche Fahrmeir & Tutz (2001) Seite 42).

Bei der Anwendung des beschriebenen Schätzverfahrens im Rahmen der Simulationsstudien in Kapitel 5 trat bei poissonverteiltem Response in einigen Fällen das Problem auf, dass die erwartete Fisher-Information $F(\delta)$ numerisch nicht vollen Rang besaß. Dieses Problem war dabei in der Regel schon bei den ersten Iterationen zu beobachten. Um dennoch die Schätzung des Modells zu ermöglichen, wurden bei Vorliegen eines Rangdefizits die Diagonalelemente von $F(\delta)$ mit 1.0005 multipliziert. Der Wert 1.0005 ist dabei ein willkürlich gewählter Wert, der

im Verhältnis zu den Daten nur eine kleine Veränderung von $F(\delta)$ verursacht, aber bewirkt, dass das Rangdefizit behoben wird, $F(\delta)$ also wieder invertierbar ist.

Man vergleiche beispielsweise Kennedy Jr. & Gentle (1980) Seite 443-450 für ausgefeiltere Strategien mit denen man dem Rangdefizit begegnen kann. Die Grundidee bleibt jedoch stets eine Vergrößerung der Diagonalelemente von $F(\delta)$, so dass die Matrix wieder vollen Rang besitzt. Da die von Kennedy Jr. & Gentle vorgestellten Verfahren auf einer modifizierten Cholesky-Zerlegung beruhen und damit nur unter verhältnismäßig großem Aufwand in S-Plus zu implementieren gewesen wären, wurde im Rahmen dieser Arbeit darauf verzichtet.

In den beobachteten problematischen Fällen mussten in der Regel in mehreren aufeinander folgenden Iterationen die Diagonalelemente von $F(\delta)$ vergrößert werden, bis der Schätzer für δ nahe genug zur Lösung $\hat{\delta}$ konvergiert war. Ab einem bestimmten Zeitpunkt besaß die Fisher-Informationsmatrix dann wieder vollen Rang, so dass man davon ausgehen kann, dass die Schätzergebnisse von der Modifikation unbeeinflusst bleiben.

2.4 Varianzparameter im generalisierten linearen gemischten Modell

Überträgt man die Konzepte zur Schätzung der Varianzparameter aus dem linearen Modell mit zufälligen Effekten auf das generalisierte lineare gemischte Modell, so erhält man Maximum-Likelihood-Schätzer durch das Maximieren der marginalen Likelihood, die sich als

$$L(\vartheta) = \int p(y|b)p(b)db$$

berechnen lässt. Im Allgemeinen ist dieses Integral jedoch nicht wie in Kapitel 2.2 analytisch lösbar, sondern muss numerisch ausgewertet werden. Eine Alternative bietet eine Approximation, wie sie unten für den REML-Schätzer vorgestellt wird.

Um wieder den Verlust an Freiheitsgraden zu berücksichtigen, soll nun das Konzept der Restricted-Likelihood auf das generalisierte lineare gemischte Modell übertragen werden. Dabei ergibt sich das Problem, dass die ursprüngliche Idee der Verwendung von Fehlerkontrasten $u = A'y$ und der Maximierung der Marginalverteilung (bezüglich der zufälligen Effekte) dieser Fehlerkontraste nicht direkt

im generaliserten linearen Modell anwendbar ist. Zum einen ist die Marginalverteilung der Fehlerkontraste wieder nur numerisch zu bestimmen und zum anderen erhält man durch Linearkombinationen $a'y$ im Allgemeinen keine Fehlerkontraste, deren Verteilung von β unabhängig ist. Dies liegt an der nichtlinearen Abhängigkeit des Erwartungswertes $\mu_i = \mathbb{E}(y_i|b)$ von β.

Es lässt sich jedoch zeigen, dass man die Restricted-Likelihood in einem generalisierten linearen gemischten Modell approximieren kann durch die Restricted-Likelihood in einem linearen gemischten Modell für die wie in (2.17) definierten Arbeitsbeobachtungen $\tilde{y}(\delta)$ (Breslow & Clayton (1993), Lin & Zhang (1999)).

Dazu approximiert man zunächst die Log-Likelihood $l(\delta) = \log(p(y|\delta))$ durch die Pearson χ^2-Statistik, wobei nun explizit die Abhängigkeit der Matrizen $\Sigma(\delta, \phi)$ und $W(\delta, \phi)$ vom Skalenparameter ϕ beachtet werden soll:

$$\begin{aligned} l(\delta) &\approx \sum_{i=1}^{n} \frac{(y_i - \mu_i(\delta))^2}{\omega_i v(\mu_i)/\phi} \\ &= (y - \mu(\delta))' \Sigma(\delta, \phi)^{-1} (y - \mu(\delta)). \end{aligned}$$

Dies entspricht der Laplace-Approximation der Log-Likelihood wie sie in Tierney & Kadane (1986) vorgestellt wird. Aus der Definition von $\tilde{y}(\delta)$ in (2.17) erhält man:

$$(y - \mu(\delta)) = D(\delta)(\tilde{y}(\delta) - X\beta - Zb).$$

Damit ergibt sich

$$\begin{aligned} l(\delta) &\approx (\tilde{y}(\delta) - X\beta - Zb)' D(\delta) \Sigma(\delta, \phi)^{-1} D(\delta) (\tilde{y}(\delta) - X\beta - Zb) \\ &= (\tilde{y}(\delta) - X\beta - Zb)' W(\delta, \phi) (\tilde{y}(\delta) - X\beta - Zb). \end{aligned}$$

Nimmt man nun noch an, dass die GLM-Gewichte $w_i(\delta, \phi)$ nur langsam in Abhängigkeit von μ variieren (Breslow & Clayton 1993), so erkennt man, dass $l(\delta)$ durch die Log-Likelihood eines linearen gemischten Modells für $\tilde{y}(\delta)$ approximiert werden kann. Genauer nimmt man, dass

$$\tilde{y}(\delta)|b \overset{a}{\sim} \mathrm{N}(X\beta + Zb, W(\delta, \phi)^{-1})$$

gilt und damit

$$\tilde{y}(\delta) \overset{a}{\sim} \mathrm{N}(X\beta, V(\delta, \vartheta))$$

mit $V(\delta,\vartheta) = W(\delta,\phi)^{-1} + ZQ(\nu)Z'$ und $\vartheta = (\phi,\nu')'$.

Im linearen gemischten Modell für die Arbeitsbeobachtungen $\tilde{y}(\delta)$ kann man die Restricted-Likelihood dann wie in Anhang A.2 herleiten. Man ersetzt lediglich die Annahme $\mathrm{Var}(y|b) = \sigma^2 I_n$ durch $\mathrm{Var}(\tilde{y}(\delta)|b) = W(\delta,\phi)^{-1}$ und erhält damit als zu maximierenden Ausdruck

$$\begin{aligned} l^*(\vartheta) &= -\frac{1}{2}\log(|V(\delta,\vartheta)|) - \frac{1}{2}\log(|X'V(\delta,\vartheta)^{-1}X|) \\ &\quad -\frac{1}{2}(\tilde{y}(\delta) - X\hat{\beta})'V(\delta,\vartheta)^{-1}(\tilde{y}(\delta) - X\hat{\beta}). \end{aligned}$$

Die Score-Funktion und die beobachtete beziehungsweise erwartete Fisher-Information besitzen dann ebenfalls ähnliche Darstellungen wie in Kapitel 2.2. Genauer erhält man für die Elemente $s_j^*(\vartheta)$ des Score-Vektors und $F_{jk}^*(\vartheta)$ der erwarteten Fisher-Information:

$$\begin{aligned} s_j^*(\vartheta) &= -\frac{1}{2}\,\mathrm{spur}\left(P(\delta,\vartheta)\frac{\partial V(\delta,\vartheta)}{\partial\vartheta_j}\right) \\ &\quad +\frac{1}{2}(\tilde{y}(\delta) - X\hat{\beta})'V(\delta,\vartheta)^{-1}\frac{\partial V(\delta,\vartheta)}{\partial\vartheta_j}V(\delta,\vartheta)^{-1}(\tilde{y}(\delta) - X\hat{\beta}) \\ &= -\frac{1}{2}\,\mathrm{spur}\left(P(\delta,\vartheta)\frac{\partial V(\delta,\vartheta)}{\partial\vartheta_j}\right) \qquad (2.19) \\ &\quad +\frac{1}{2}(\tilde{y}(\delta) - X\hat{\beta} - Z\hat{b})'W(\delta,\phi)\frac{\partial V(\delta,\vartheta)}{\partial\vartheta_j}W(\delta,\phi)(\tilde{y}(\delta) - X\hat{\beta} - Z\hat{b}) \end{aligned}$$

und

$$F_{jk}^*(\vartheta) = \frac{1}{2}\,\mathrm{spur}\left(P(\delta,\vartheta)\frac{\partial V(\delta,\vartheta)}{\vartheta_j}P(\delta,\vartheta)\frac{\partial V(\delta,\vartheta)}{\vartheta_k}\right) \qquad (2.20)$$

mit

$$\begin{aligned} P(\delta,\vartheta) &= V(\delta,\vartheta)^{-1} - V(\delta,\vartheta)^{-1}X(X'V(\delta,\vartheta)^{-1}X)^{-1}X'V(\delta,\vartheta)^{-1} \\ &= W(\delta,\phi) - W(\delta,\phi)(X,Z)H^{-1}(X,Z)'W(\delta,\phi) \end{aligned}$$

und

$$H = \begin{pmatrix} X'W(\delta,\phi)X & X'W(\delta,\phi)Z \\ Z'W(\delta,\phi)X & Z'W(\delta,\phi)Z + Q^{-1}(\nu) \end{pmatrix}.$$

Wie im linearen gemischten Modell werden zur Schätzung wieder δ und ϑ ausgehend von Startwerten jeweils abwechselnd durch neue Approximationen ersetzt, bis die relativen Veränderungen von $\hat{\delta}$ und $\hat{\vartheta}$ klein sind. Das genaue Vorgehen wird im folgenden Algorithmus zusammengefasst:

Algorithmus 2 (Schätzung im generalisierten linearen gemischten Modell)

(i) Wähle Startwerte $\hat{\beta}^{(0)}$, $\hat{b}^{(0)}$ und $\hat{\vartheta}^{(0)}$, die maximale Iterationszahl $maxit$ sowie ein Abbruchkriterium ϵ . Setze $k = 0$.

(ii) Bestimme die Arbeitsbeobachtungen $\tilde{y}^{(k)} = \tilde{y}(\hat{\delta}^{(k)}) = X\hat{\beta}^{(k)} + Z\hat{b}^{(k)} + D(\hat{\delta}^{(k)})^{-1}(y - \mu(\hat{\delta}^{(k)}))$ und Arbeitsgewichte $W^{(k)} = W(\hat{\delta}^{(k)}, \hat{\vartheta}^{(k)})$ und berechne $\hat{\beta}^{(k+1)}$ und $\hat{b}^{(k+1)}$ durch Lösen des Gleichungssystems

$$\begin{pmatrix} X'W^{(k)}X & X'W^{(k)}Z \\ Z'W^{(k)}X & Z'W^{(k)}Z + Q(\hat{\nu}^{(k)})^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'W^{(k)}\tilde{y}^{(k)} \\ Z'W^{(k)}\tilde{y}^{(k)} \end{pmatrix}.$$

(iii) Bestimme $\hat{\vartheta}^{(k+1)}$ als

$$\hat{\vartheta}^{(k+1)} = \hat{\vartheta}^{(k)} + F^*(\hat{\vartheta}^{(k)})^{-1} s^*(\hat{\vartheta}^{(k)}),$$

wobei $s^*(\hat{\vartheta}^{(k)})$ und $F^*(\hat{\vartheta}^{(k)})$ aus (2.19) und (2.20) in Abhängigkeit von $\hat{\beta}^{(k+1)}$ und $\hat{b}^{(k+1)}$ berechnet werden.

(iv) Bestimme die Abbruchkriterien

$$d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) = \frac{||\hat{\vartheta}^{(k+1)} - \hat{\vartheta}^{(k)}||}{||\hat{\vartheta}^{(k)}||}$$

und

$$d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) = \frac{||\hat{\delta}^{(k+1)} - \hat{\delta}^{(k)}||}{||\hat{\delta}^{(k)}||}.$$

Falls $d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) > \epsilon$ oder $d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) > \epsilon$ und $k < maxit$, setze $k = k + 1$ und gehe zurück zu (ii).
Falls $d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) < \epsilon$ und $d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) < \epsilon$ und $k < maxit$, so sind $\hat{\vartheta}^{(k+1)}$ und $\hat{\delta}^{(k+1)}$ die endgültigen Schätzer.
Falls $k = maxit$, breche die Schätzung ohne Ergebnis ab.

Alternativ zur iterativen Schätzung von $\hat{\vartheta}_{REML}$ über die direkte Maximierung der Restricted-Likelihood wäre auch die indirekte Maximierung über den EM-Algorithmus (Fahrmeir & Tutz (2001) Seite 301/2) denkbar. Eine alternative Schätzmethode bietet ein voller Bayes-Ansatz wie beispielsweise in Gamerman (1997).

2.5 Quasi-Likelihood-Modelle

In den Modellen, die in Kapitel 2.3 vorgestellt wurden, ist stets die Annahme enthalten, dass die Verteilung von $y_i|b$ aus einer Exponentialfamilie stammt. Diese Annahme impliziert für einen bestimmten Verteilungstyp eine bestimmte Varianzfunktion, die man aus der Beziehung $v(\mu) = b''(\theta)$ erhalten kann. Mit Hilfe dieser Varianzfunktion ist dann die Varianz der abhängigen Variablen festgelegt durch $\text{Var}(y_i|b) = \phi v(\mu_i)/\omega_i$. Für einige Verteilungen ist der Skalenparameter ϕ dabei vorab festgelegt. Beispielsweise gilt für Poisson- und Binomialverteilung $\phi = 1$. In vielen konkreten Datensituationen ist man jedoch mit dem Problem konfrontiert, dass die Daten eine zusätzliche Variabilität besitzen, die nicht durch die Verteilungsannahme erklärt werden kann. Dieses Phänomen nennt man Überdispersion (siehe beispielsweise Fahrmeir & Tutz (2001) Seite 35 für eine Beschreibung des Problems in einem binären Modell).

Eine einfache Möglichkeit, diesem Problem zu begegnen, besteht nun darin, den Skalenparameter nicht zu fixieren, sondern als zusätzlichen Dispersionsparameter ebenfalls aus den Daten zu schätzen. Dieser Dispersionsparameter wird dann in die Formeln der Score-Funktion und der erwarteten Fisher-Information, wie sie im Posteriori-Modus-Ansatz in Kapitel 2.3.3 bestimmt wurden, eingesetzt. Die Schätzung der Regressionsparameter β und b kann dann mittels Fisher-Scoring wie beschrieben durchgeführt werden.

Dabei ist aber zu beachten, dass durch die Einführung des zusätzlichen Dispersionsparameters die Score-Funktion nicht mehr die Ableitung der Likelihood einer Poisson- oder Binomialverteilung ist. Es ist jedoch möglich, eine Quasi-Likelihood $q(\beta, b, \phi, \vartheta)$ zu definieren, für die $\partial q(\beta, b, \phi, \vartheta)/\partial\beta$ die Form (2.13) und $\partial q(\beta, b, \phi, \vartheta)/\partial b$ die Form (2.14) besitzen (vergleiche Fahrmeir & Tutz (2001) Kapitel 2.3.1).

Nach Lin & Zhang (1999) kann auch die so konstruierte Quasi-Likelihood mit Hilfe eines linearen gemischten Modells für die Arbeitsbeobachtungen approximiert werden. Somit ist die REML-Schätzung der Varianzparameter und des zusätzlichen Dispersionsparameters ϕ auch in Quasi-Likelihood-Modellen möglich.

3 Generalisierte geoadditive gemischte Modelle

3.1 Modell

In vielen Datensituationen erweist sich die lineare Modellierung des Einflusses metrischer Kovariablen auf die abhängige Variable y, wie sie in generalisierten linearen gemischten Modellen angenommen wird, als zu restriktiv. Betrachtet man etwa die Nettomieten pro Quadratmeter, die zur Bestimmung des Münchner Mietspiegels verwendet werden (vergleiche Kapitel 6.3 für eine genaue Beschreibung und Analyse dieser Daten), so werden diese beispielsweise durch die Wohnfläche einer Wohnung und das Baujahr des entsprechenden Gebäudes beeinflusst.

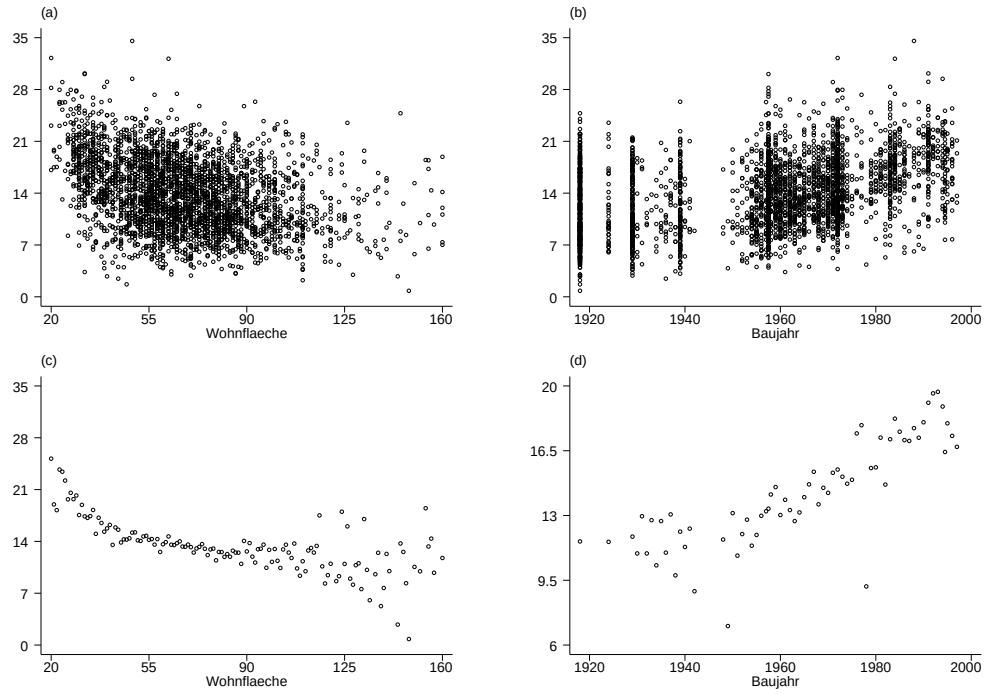

Abbildung 3.1: Scatterplots der Nettomiete pro Quadratmeter gegen Wohnfläche und Baujahr ((a) und (b)) sowie der mittleren Nettomiete pro Quadratmeter gegen Wohnfläche und Baujahr ((c) und (d)).

In Abbildung 3.1 (a) und (b) sind die Nettomieten gegen die Wohnfläche beziehungsweise das Baujahr aufgetragen. Zusätzlich sind in den Grafiken (c) und (d) die mittleren Nettomieten für die verschiedenen Ausprägungen der beiden Kovariablen visualisiert. Für den Einfluss der Wohnfläche ergeben sich daraus

deutliche Hinweise auf einen nichtlinearen Zusammenhang, die man insbesondere aus Grafik (c) ablesen kann. Für den Einfluss des Baujahrs erscheint die Beurteilung dagegen etwas schwieriger, man erhält aber auch leichte Hinweise auf einen nichtlinearen Effekt. Man beachte dabei auch, dass in Grafik (d) eine andere Achseneinteilung gewählt wurde, um den Verlauf des Zusammenhangs deutlicher erkennen zu können.

Prinzipiell lassen sich auch nichtlineare Zusammenhänge zwischen Kovariablen und abhängiger Variable mit Hilfe von generalisierten linearen gemischten Modellen schätzen, solange das Modell in den Parametern β und b linear ist. Dazu ist aber ein gewisses Vorwissen über die funktionale Form der Abhängigkeit notwendig. Nimmt man etwa an, dass der Einfluss einer Kovariablen x einem Polynom vom Grad l entspricht, so lässt sich dieses über ein generalisiertes lineares Modell mit dem in β linearen Prädiktor

$$\eta_i = \beta_0 + x_i\beta_1 + \ldots + x_i^l\beta_l$$

bestimmen. Da man jedoch in der Regel kein Vorwissen über die genaue Form des Einflusses der Kovariablen besitzt, ist man an einer flexibleren Möglichkeit der Modellierung interessiert. Hierzu bietet sich die additive Modellierung des Einflusses metrischer Kovariablen an, wie sie von Hastie & Tibshirani (1990) vorgeschlagen und beispielsweise von Lin & Zhang (1999) auf Modelle mit zufälligen Effekten erweitert wurde. Während man die Annahmen über die Verteilung der abhängigen Variable aus Kapitel 2.3 beibehält, ersetzt man den linearen Prädiktor

$$\eta_i = x_i'\beta + z_i'b$$

durch den additiven Prädiktor

$$\eta_i = x_{i,par}'\beta_{par} + f_1(x_{i1}) + \ldots + f_s(x_{is}) + z_{i,ran}'b_{ran}.$$

Mit f_1 bis f_s werden dabei Funktionen bezeichnet, die in ihrer funktionalen Form unspezifiziert bleiben und nicht unmittelbar parametrisiert werden. Man spricht daher auch von einer nonparametrischen Modellierung des Effekts der Kovariablen x_{i1} bis x_{is}. Man fordert lediglich, dass die Funktionen f_1 bis f_s gewissen Glattheitsansprüchen genügen, also etwa stetig, stetig differenzierbar oder mehrfach stetig differenzierbar sein sollen. Häufig wird diese Forderung noch dadurch eingeschränkt, dass man annimmt, dass die Funktionen aus einem bestimmten,

endlich dimensionalen Teilraum beispielsweise des Raums der zweimal stetig differenzierbaren Funktionen stammen. Die Funktionen lassen sich dann als Linearkombinationen einer endlichen Menge von Basisfunktionen darstellen, so dass man von Basisfunktionenansätzen spricht. In Abschnitt 3.1.1 werden einige auf Basisfunktionenansätzen beruhende Möglichkeiten zur nonparametrischen Modellierung der Funktionen f_1 bis f_s vorgestellt, wobei der Schwerpunkt auf den von Eilers & Marx (1996) eingeführten P-Splines liegen wird.

Im $(p_{par}+1)$-dimensionalen Vektor $x_{i,par}$ werden Kovariablen zusammengefasst, die in der Regel kategorialer Natur sind und deren Einfluss auch im weiteren parametrisch modelliert und geschätzt werden soll. Insbesondere soll $x_{i,par}$ stets die Konstante und dementsprechend β_{par} den Intercept enthalten. Der q_{ran}-dimensionale Vektor $z_{i,ran}$ beinhaltet Kovariablen, deren Regressionskoeffizienten b_{ran} als zufällig angenommen werden sollen. Wie in Kapitel 2 soll für den Parameter b_{ran} wieder gelten

$$b_{ran} \sim \mathrm{N}(0, Q_{ran}(\nu_{ran})).$$

Bei der Bestimmung nonparametrischer Effekte kann es zu Identifizierbarkeitsproblemen kommen, da häufig zwar die funktionale Form der zu schätzenden Funktionen identifzierbar ist, aber nicht das Niveau der einzelnen Funktionen. Betrachtet man als Beispiel ein Modell mit zwei Funktionen, also

$$\eta_i = f_1(x_{i1}) + f_2(x_{i2}),$$

so verändert sich das Modell nicht, wenn man von f_1 eine Konstante abzieht und diese zu f_2 dazu addiert. Um die Identifizierbarkeit zu gewährleisten, müssen gewisse Nebenbedingungen an das Niveau der einzelnen Funktionen erfüllt sein. In der Regel fordert man, dass die Funktionen f_1 bis f_s zentriert sein sollen, das heißt

$$\sum_{i=1}^{n_j} f_j(x_{(i)j}) = 0.$$

Mit $x_{(1)j}, \ldots, x_{(n_j)j}$ werden dabei die n_j verschiedenen Ausprägungen der j-ten nonparametrisch modellierten Kovariable bezeichnet. Der additive Prädiktor ändert sich durch die Zentrierungsbedingung nicht, solange er den Intercept enthält, wovon im Folgenden stets ausgegangen werden soll.

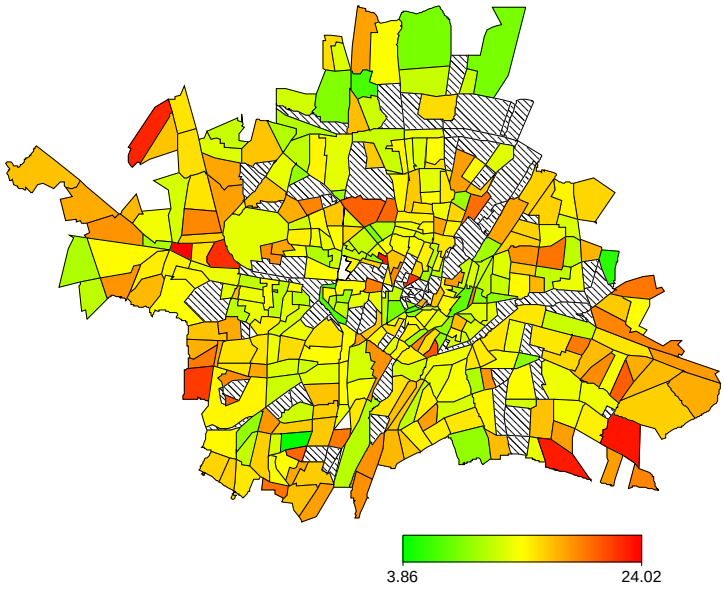

Abbildung 3.2: Mittlere Nettomieten pro Quadratmeter in den Münchner Bezirksvierteln.

Eine zusätzliche Besonderheit stellen Datensituationen dar, in denen die Daten eine räumliche Struktur aufweisen. Häufig muss man dann davon ausgehen, dass die Beobachtungen $y_1, \ldots, y_n$ nicht mehr als unabhängig angenommen werden können, sondern dass sie als räumlich korreliert betrachtet werden müssen. Beispielsweise lassen sich die Wohnungen, die für den Münchner Mietspiegel untersucht wurden, unterschiedlichen Bezirksvierteln innerhalb Münchens zuordnen. Dabei ist es plausibel, anzunehmen, dass sich Wohnungen benachbarter Bezirksviertel in Bezug auf ihre Nettomiete pro Quadratmeter ähnlicher sind, als Wohnungen aus weit voneinander entfernten Bezirksvierteln, so dass Korrelationen der Beobachtungen zu erwarten sind. In Abbildung 3.2 sind die mittleren Mieten der Münchner Bezirksviertel wiedergegeben. Man erkennt relativ deutliche Hinweise auf eine räumliche Struktur der Nettomieten, beispielsweise mit erhöhten Mieten im Münchner Osten.

Das Ziel ist es nun, analog zur glatten, nonparametrischen Modellierung des Effekts metrischer Kovariablen eine räumlich glatte Funktion als Schätzung des räumlichen Effekts zu erhalten. Dazu erweitert man den additiven Prädiktor um die Funktion f_{spat} zu

$$\eta_i = x'_{i,par}\beta_{par} + f_1(x_{i1}) + \ldots + f_s(x_{is}) + f_{spat}(R_i) + z'_{i,ran}b_{ran},$$

wobei mit R_i die Region bezeichnet sei, zu der Beobachtung i gehört. Man beachte, dass dabei auch an f_{spat} geeignete Nebenbedingungen zu stellen sind, um die Identifizierbarkeit des Modells zu gewährleisten. Im Folgenden wird darum davon ausgegangen, dass auch f_{spat} zentriert ist.

Man beachte außerdem, dass in der Regel nicht die Zugehörigkeit einer Beobachtung zu einer bestimmten Region an sich einen Effekt auf die abhängige Variable ausübt. Stattdessen versucht man, über die Einbeziehung eines räumlichen Effekts unbeobachtete, mit der räumlichen Struktur in Zusammenhang stehende Kovariablen zu berücksichtigen. In ähnlicher Weise wird häufig ein glatter, zeitabhängiger Trend geschätzt, um unbeobachtete, strukturiert über die Zeit variierende Kovariablen berücksichtigen zu können. Häufig dient die räumliche Analyse auch dazu, Hinweise auf zugrunde liegende, bisher unbekannte Einflussfaktoren zu finden, die die räumliche Variation der Daten zur Folge haben (vergleiche Kapitel 6.1).

In einigen Fällen kann es sinnvoll sein, den räumlichen Effekt in einen strukturierten und einen unstrukturierten Anteil aufzuspalten. Diesem Vorgehen liegt die Idee zugrunde, dass man in der Regel nicht vorab weiß, ob die den räumlichen Effekt verursachenden Kovariablen eine räumliche Struktur aufweisen oder räumlich unstrukturiert variieren. Indem man beide Effekte im Modell berücksichtigt, erhält man die Möglichkeit, dies zu beurteilen. Während der räumlich strukturierte Anteil über die Funktion f_{spat} modelliert wird, kann man für den unstrukturierten Anteil einen zufälligen Intercept annehmen, für den die Gruppierungsstruktur auf der Zugehörigkeit zu den Regionen beruht. Man vergleiche auch Kapitel 6.1 in dem die Aufspaltung des räumlichen Effekts in einen strukturierten und einen unstrukturierten Teil zur Analyse des Mortalitätsrisikos bezüglich Mundhöhlenkrebs verwendet wird.

In Matrixdarstellung erhält man nun insgesamt das Modell

$$\eta = X_{par}\beta_{par} + f_1 + \ldots + f_s + f_{spat} + Z_{ran}b_{ran} \tag{3.1}$$

mit

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \; X_{par} = \begin{pmatrix} x'_{par,1} \\ \vdots \\ x'_{par,n} \end{pmatrix}, \; f_j = \begin{pmatrix} f_j(x_{1j}) \\ \vdots \\ f_j(x_{nj}) \end{pmatrix}, \; j = 1, \ldots, s,$$

$$f_{spat} = \begin{pmatrix} f_{spat}(R_1) \\ \vdots \\ f_{spat}(R_n) \end{pmatrix}, \; Z_{ran} = \begin{pmatrix} z'_{ran,1} \\ \vdots \\ z'_{ran,n} \end{pmatrix}.$$

Ein auf Prädiktor (3.1) basierendes Modell wird im Weiteren generalisiertes geoadditives gemischtes Modell genannt werden.

Nun sollen die einzelnen Verfahren, mit denen die zwei neuen Modellkomponenten f_1 bis f_s beziehungsweise f_{spat} modelliert werden, zunächst anhand von einfachen Modellen vorgestellt werden, in denen jeweils nur einer der Effekte vertreten ist. Anschließend wird dann wieder auf das vollständige, durch (3.1) beschriebene Modell eingegangen.

3.1.1 P-Splines

Der folgende Abschnitt beruht im Wesentlichen auf den Darstellungen zu P-Splines in Brezger (2000), Eilers & Marx (1996) und Marx & Eilers (1998). An einzelnen Stellen wird aber auch zusätzlich auf weiterführende Literatur verwiesen. Aus der Vielzahl von Möglichkeiten, die Funktionen f_1 bis f_s zu modellieren, sollen im Folgenden nur solche herausgegriffen werden, die auf der Verwendung von Basisfunktionenansätzen beruhen oder sich auf einen solchen Ansatz zurückführen lassen. Speziell werden Polynom-Splines und P-Splines, sowie kurz auch Glättungssplines behandelt werden. Man betrachte Kapitel 2 in Hastie & Tibshirani (1990) oder Kapitel 5 in Fahrmeir & Tutz (2001) für einen weitergehenden Überblick zu anderen nonparametrischen Modellierungsmöglichkeiten der Funktionen f_1 bis f_s.

Basisfunktionenansätze sind nicht im eigentlichen Sinne nonparametrisch, da die Schätzung der Funktionen f_1 bis f_s zurückgeführt wird auf die Schätzung einer Menge von Regressionsparametern, die allerdings möglicherweise sehr groß sein kann. Eilers & Marx (1996) schlagen daher vor, von überparametrisierten Modellen oder anonymen Modellen zu sprechen, weil die geschätzten Regressionskoeffizienten für sich keine direkte statistische Interpretation mehr besitzen. Im Folgenden soll dennoch stets von einer nonparametrischen Modellierung der glatten Funktionen gesprochen werden, weil diese Bezeichnung für die Beschreibung des Einflusses metrischer Kovariablen über glatte Funktionen allgemein üblich ist.

Um die Modellierung der Funktionen f_1 bis f_s vorzustellen, soll nun zunächst das Modell mit Prädiktor

$$\eta_i = f(x_i)$$

behandelt werden. Dazu definiert man zunächst eine geeignete Klasse von Funktionen, deren Elemente dann mit Hilfe von Basisfunktionen dargestellt werden können.

Polynom-Splines

Sei $a = \xi_1 < \ldots < \xi_m = b$ eine Zerlegung des Intervalls $[a, b]$. Eine Funktion $g : [a, b] \to \mathbb{R}$ heißt dann Polynom-Spline vom Grad l, $l \in \mathbb{N}_0$, zur Knotenmenge $\Omega_m = \{\xi_1, \ldots, \xi_m\}$, falls gilt

1. $g(x)$ ist $(l-1)$-mal stetig differenzierbar und

2. $g(x)$ ist für $x \in [\xi_j, \xi_{j+1})$, $j = 1, \ldots, m-1$, ein Polynom vom Grad l.

Ein Polynom-Spline g ist also insbesondere stückweise stetig, für $l > 0$ stetig und für $l > 1$ stetig differenzierbar. Die einzelnen Elemente von Ω_m werden als Knoten bezeichnet.

Im Folgenden soll weitgehend auf Beweise verzichtet werden, um die Darstellung nicht unnötig zu verkomplizieren. Stattdessen werden eine Reihe von Ergebnissen zu Polynom-Splines vorgestellt und eventuell plausibel gemacht. Zu einer detaillierteren und theoretischeren Einführung zu Splines vergleiche man etwa Hämmerlin & Hoffmann (1994) Kapitel 6.

Die Menge aller Polynom-Splines vom Grad l zur Knotenmenge Ω_m bildet einen $(m+l-1)$-dimensionalen Teilraum des Vektorraums der $(l-1)$-mal stetig differenzierbaren Funktionen, der häufig mit $S_l(\Omega_m)$ bezeichnet wird. Ein Polynom-Spline lässt sich demzufolge mit Hilfe einer Menge von $m + l - 1$ linear unabhängigen Basisfunktionen $B_1^l(x), \ldots, B_{m+l-1}^l(x)$ aus $S_l(\Omega_m)$ darstellen.

Eine mögliche Basis von $S_l(\Omega_m)$ bilden Polynome und abgeschnittene Potenzen (Truncated Power Series-Basis):

$$B_1^l(x) = 1, B_2^l(x) = x, \ldots, B_{l+1}^l(x) = x^l,$$
$$B_{l+2}^l(x) = (x - \xi_2)_+^l, \ldots, B_{m+l-1}^l(x) = (x - \xi_{m-1})_+^l$$

mit

$$(x - \xi_j)_+^l = \begin{cases} (x - \xi_j)^l & \text{für } x \geq \xi_j \\ 0 & \text{für } x < \xi_j. \end{cases}$$

Nimmt man nun an, dass die zu schätzende Funktion f ein Polynom-Spline ist oder zumindest durch einen Polynom-Spline angenähert werden kann, so gilt mit gewissen $\zeta_1, \ldots, \zeta_{m+l-1}$ für $x \in [a, b]$

$$f(x) = \sum_{j=1}^{m+l-1} \zeta_j B_j^l(x).$$

Damit lässt sich die Schätzung von f auf die Schätzung der Koeffizienten $\zeta = (\zeta_1, \ldots, \zeta_{m+l-1})'$ zurückführen. Der Parametervektor ζ beinhaltet nämlich die Regressionskoeffizienten eines generalisierten linearen Modells mit linearem Prädiktor

$$\eta = B\zeta,$$

wobei die $n \times (m + l - 1)$ Designmatrix B gegeben ist durch

$$B = \begin{pmatrix} B_1^l(x_1) & \ldots & B_{m+l-1}^l(x_1) \\ \vdots & \ddots & \vdots \\ B_1^l(x_n) & \ldots & B_{m+l-1}^l(x_n) \end{pmatrix}. \tag{3.2}$$

Die Schätzung von ζ kann nun, falls B vollen Spaltenrang hat, mit Hilfe von Standardverfahren aus der generalisierten linearen Regression wie in Kapitel 2.3 beschrieben erfolgen. Eine Matrix B mit Rangdefizit kann beispielsweise auftreten, wenn in einem der Intervalle $[\xi_j, \xi_{j+1})$ keine Daten beobachtet wurden. Im Folgenden soll aber stets davon ausgegangen werden, dass B vollen Spaltenrang besitzt. Eine Schätzung für die Funktion f erhält man aus der Schätzung $\hat{\zeta}$ durch $\hat{f} = B\hat{\zeta}$.

Die aus Polynomen und abgeschnittenen Potenzen bestehende Basis des Splineraums $S_l(\Omega_m)$ ist zwar einfach zu verstehen und zu implementieren, weist aber zwei Nachteile auf: Zum einen sind die Basisfunktionen nicht lokal definiert, was sich beim später zur Schätzung von P-Splines eingeführten Penalisierungskonzept als problematisch erweisen wird. Zum anderen sind die einzelnen Basisfunktionen (mit Ausnahme von $B_1^l(x)$) nicht nach oben beschränkt. Darum führt die Verwendung der Truncated Power Series-Basis häufig zu numerischen Problemen, die aus den möglicherweise sehr großen Werten in der Matrix B resultieren. Darum soll nun eine weitere Basis des Splineraums $S_l(\Omega_m)$ vorgestellt werden, die nicht nur numerisch vorteilhafter, sondern auch für das erwähnte Penalisierungskonzept wesentlich ist.

B-Splines

Definiert man rekursiv für $l \geq 0$ die Funktionen

$$B_j^l(x) = \frac{x - \xi_j}{\xi_{j+l} - \xi_j} B_j^{l-1}(x) + \frac{\xi_{j+l+1} - x}{\xi_{j+l+1} - \xi_{j+1}} B_{j+1}^{l-1}(x)$$

mit

$$B_j^0(x) = \mathbf{1}_{[\xi_j, \xi_{j+1})}(x) = \begin{cases} 1 & \xi_j \leq x < \xi_{j+1} \\ 0 & \text{sonst,} \end{cases}$$

so bilden $B_{-l+1}^l(x), \ldots, B_{m-1}^l(x)$ ebenfalls eine Basis von $S_l(\Omega_m)$. Diese Basis wird als B(asic)-Spline-Basis bezeichnet. Es lassen sich auch explizite Formeln für die Basisfunktionen mit $l \geq 1$ herleiten, worauf an dieser Stelle jedoch verzichtet werden soll. Man vergleiche beispielsweise Brezger (2000) Seite 20 für Darstellungen der B-Spline-Basisfunktionen vom Grad $l = 1$ und $l = 2$.

Zur Konstruktion der Basisfunktionen benötigt man zusätzliche Knoten $\xi_{-l+1} < \ldots < \xi_{-1} < \xi_0 < a$ und $b < \xi_{m+1} < \xi_{m+2} < \ldots < \xi_{m+l}$, die man mit den ursprünglichen Knoten zur so genannten erweiterten Partition zusammenfasst. Geht man dabei von äquidistanten Knoten $\xi_1, \ldots, \xi_m$ aus, so lassen sich die zusätzlichen Knoten einfach bestimmen, indem außerhalb des Intervalls $[a, b]$ der gleiche Abstand für die Knoten gewählt wird. Eine alternative Möglichkeit zur Wahl der Knoten besteht in der Verwendung empirischer Quantile der beobachteten Ausprägungen der zugrunde liegenden Kovariablen. Da im später eingeführten Penalisierungskonzept stets äquidistante Knoten verwendet werden, sollen auch die weiteren Betrachtungen auf äquidistante Knoten beschränkt bleiben.

Nun sollen kurz einige Eigenschaften der B-Spline-Basis zusammengefasst werden, die für die weitere Darstellung relevant sind. Für eine Reihe zusätzlicher Eigenschaften betrachte man Eilers & Marx (1996). Zur folgenden Aufzählung vergleiche man auch jeweils Abbildung 3.3, die die Basisfunktionen für äquidistante Knoten aus dem Intervall $[-3, 3]$ und $l = 1, 2, 3$ wiedergibt.

- Es gilt $\sum_{j=-l+1}^{m+l} B_j^l(x) = 1$ für $x \in [a, b]$ (Zerlegung der Einheit).
- Jede Basisfunktion ist nur in einem Bereich von $l + 2$ Knoten positiv, das heißt, B-Splines bilden eine lokale Basis.

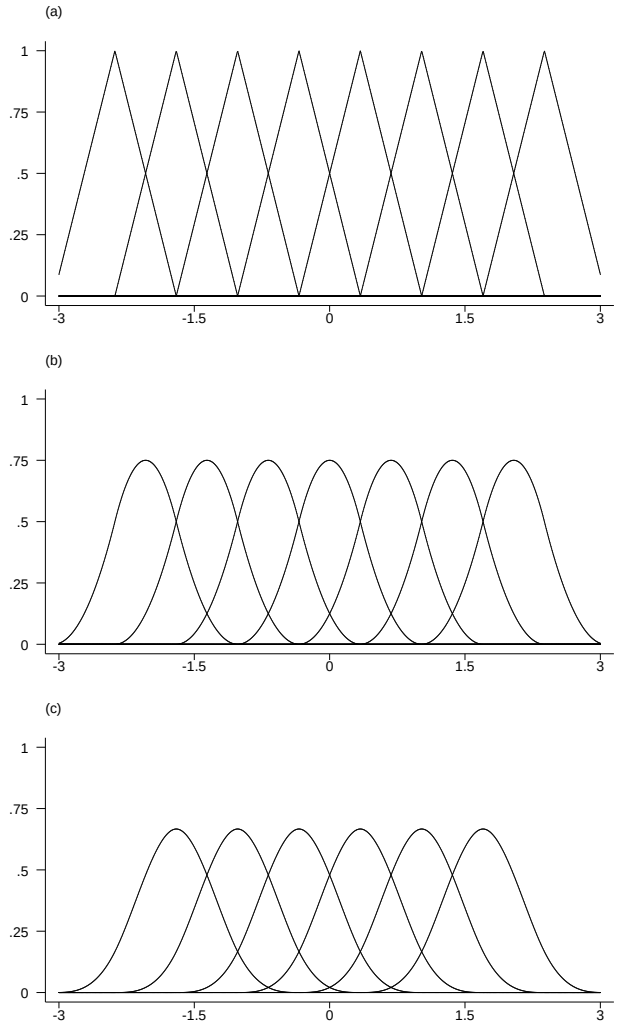

Abbildung 3.3: B-Spline-Basisfunktionen vom Grad $l = 1$ (a), $l = 2$ (b) und $l = 3$ (c).

- Alle Basisfunktionen besitzen (für äquidistante Knoten) die gleiche funktionale Form und sind lediglich auf der x-Achse verschoben.
- Die Basisfunktionen sind (für eine gegebene Knotenmenge) nach oben beschränkt.

Nimmt man wieder an, dass f aus dem Splineraum $S_l(\Omega_m)$ stammt, beziehungsweise sich durch eine Funktion aus $S_l(\Omega_m)$ approximieren lässt, so kann man f darstellen als

$$f(x) = \sum_{j=-l+1}^{m-1} \zeta_j B_j^l(x).$$

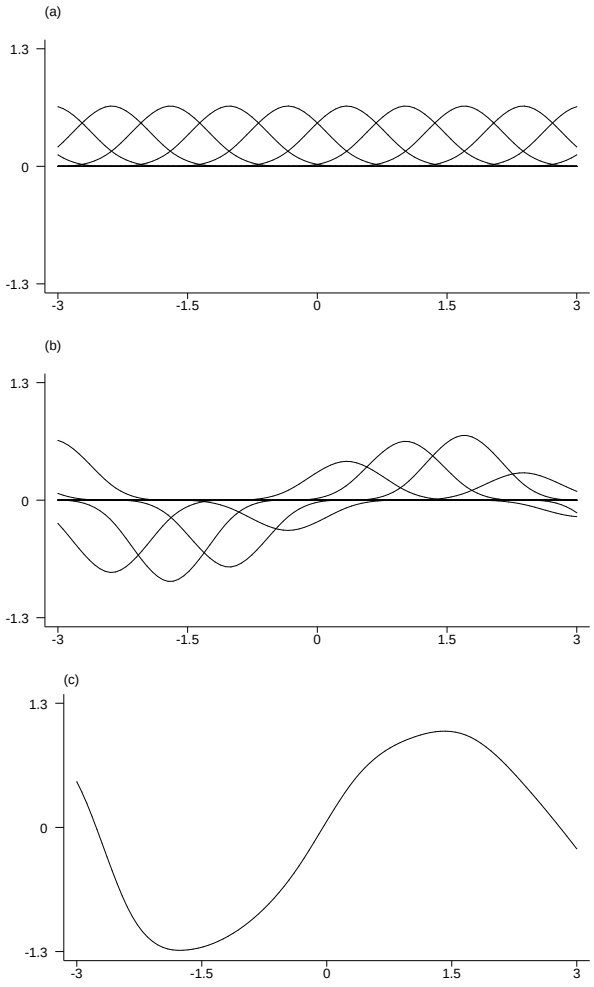

Abbildung 3.4: Nonparametrische Regression mit B-Splines: Verwendete B-Spline-Basis zum Grad 3 (a), mit ζ_j skalierte Basisfunktionen (b) und $\hat{f}$ =Summe der skalierten Basisfunktionen (c).

Die Beschränkung des Wertebereichs der Basisfunktionen führt nun zu numerischen Vorteilen bei der Schätzung der Koeffizienten $\zeta = (\zeta_{-l+1}, \ldots, \zeta_{m-1})'$, die erneut mit Hilfe von Standardverfahren aus dem generalisierten linearen Modell durchgeführt werden kann. Eine Schätzung für f erhält man wieder durch $\hat{f} = B\hat{\zeta}$, wobei B wie in (3.2) aus den Basisfunktionen besteht, ausgewertet an den beobachteten Ausprägungen der Kovariablen. Man vergleiche Abbildung 3.4, die das Vorgehen zur nonparametrischen Schätzung einer Funktion über B-Splines zusammenfasst. Ausgehend von einer Menge von Basisfunktionen (Abbildung 3.4 (a)) bestimmt man die Schätzung $\hat{\zeta}$ der Regressionskoeffizienten dieser

Basisfunktionen. Basierend auf diesen Schätzungen erhält man die skalierten Basisfunktionen $\zeta_j B_j^l(x)$ (Abbildung 3.4 (b)), die sich dann zur Schätzung $\hat{f} = B\hat{\zeta}$ aufsummieren (Abbildung 3.4 (c)).

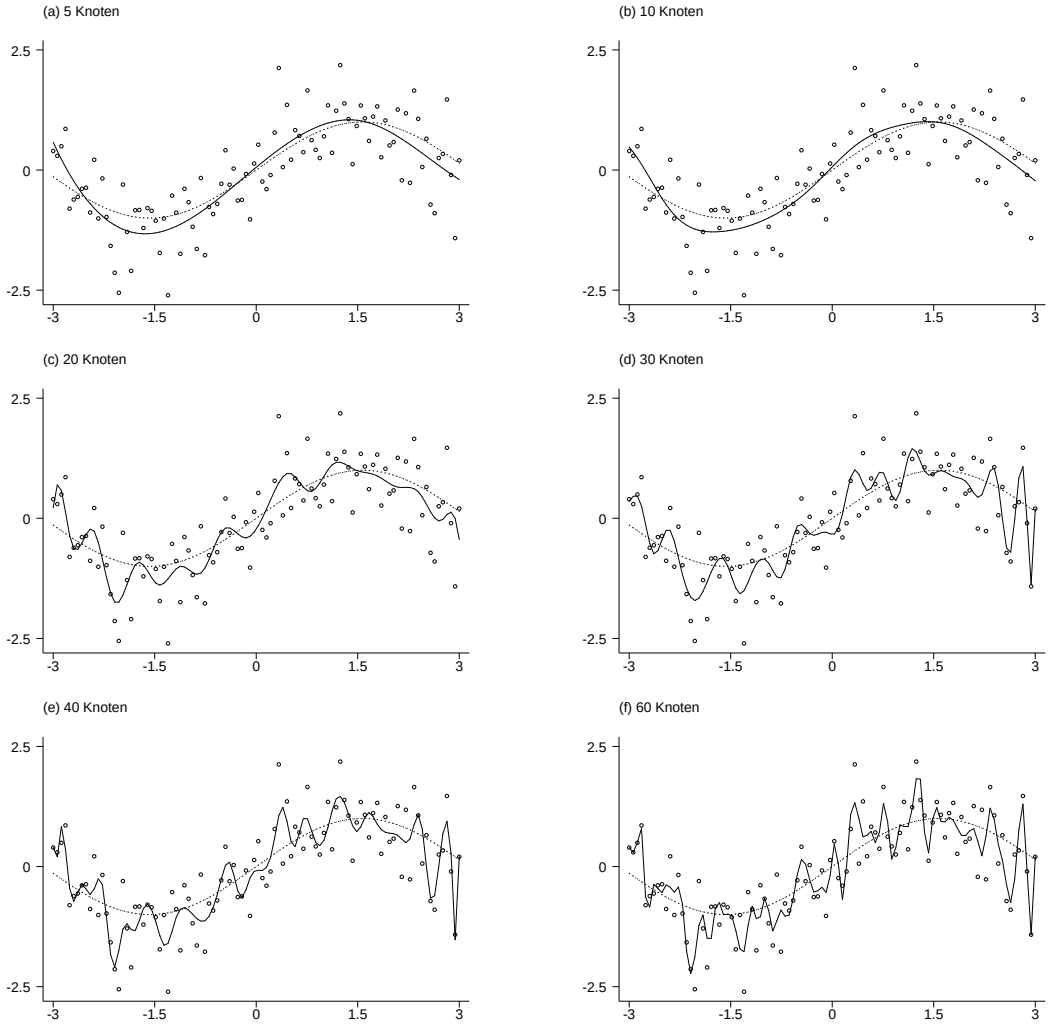

Abbildung 3.5: Einfluss der Knotenzahl auf die Schätzung. Die wahre Funktion ist jeweils gepunktet, die Schätzung als durchgezogene Linie wiedergegeben.

Approximiert man die zu schätzende Funktion f durch einen Polynom-Spline, so hängt die Schätzung neben dem Grad des Polynom-Splines l stark von der Zahl und der Position der ausgewählten Knoten ab. Insbesondere beeinflusst die Zahl der Knoten die Glattheit der Schätzung: Eine geringe Zahl von Knoten führt zu einer verhältnismäßig glatten Funktion, während eine große Zahl von Knoten zu einer raueren Schätzung führt. Man vergleiche hierzu Abbildung 3.5, in der für

verschiedene Anzahlen äquidistanter Knoten Schätzungen durchgeführt wurden. Wie man sieht, erreicht man für fünf bis zehn Knoten annehmbare Resultate, während die Verwendung einer größeren Knotenzahl eine zu große Variation beziehungsweise Datentreue der Schätzungen zur Folge hat.

Problematisch ist dabei, dass zur Schätzung also eine große Zahl von Parametern, nämlich sowohl Zahl als auch Position der Knoten, vom Anwender zu wählen sind und damit die Schätzung stark von subjektiven Entscheidungen abhängt. Um diesem Problem zu begegnen bieten sich prinzipiell zwei Auswege an: Einerseits die adaptive, das heißt datengesteuerte Wahl der Zahl und Position der Knoten und zum anderen die Verhinderung einer zu rauen Schätzung durch Penalisierungsansätze. Im Folgenden soll ein solcher Penalisierungsansatz genauer beschrieben werden. Frequentistische Ansätze zur adaptiven Knotenwahl findet man etwa in Friedman (1991) oder Stone, Hansen, Kooperberg & Troung (1997). Bayesianische Varianten dieses Prinzips werden beispielsweise in Denison, Mallick & Smith (1998) oder Biller (2000a) beschrieben. Das in Biller (2000a) dargestellte Verfahren wird auch zur Schätzung in generalisierten additiven Modellen im Rahmen einer Simulationsstudie in Kapitel 5.1 verwendet.

P-Splines

In Basisfunktionenansätzen wird die Schätzung einer nonparametrisch modellierten Funktion zurückgeführt auf die Schätzung eines generalisierten linearen Modells mit Parameter ζ. Dies geschieht in der Regel durch die Maximierung der entsprechenden Log-Likelihood. Die Grundidee von Penalisierungsansätzen und insbesondere von P-Splines besteht nun darin, eine große Zahl äquidistanter Knoten zu verwenden, um die notwendige Flexibilität der Schätzung zu gewährleisten, aber statt der Log-Likelihood eine penalisierte Log-Likelihood zu maximieren. Die Penalisierung soll dabei zu starke Schwankungen der Funktion bestrafen und durch einen einzelnen Glättungsparameter gesteuert werden. Man maximiert also die penalisierte Log-Likelihood

$$l_p(\zeta) = l(\zeta) - \frac{1}{2}\alpha J(\zeta), \tag{3.3}$$

wobei mit $J(\zeta)$ der Penalisierungsterm und mit $\alpha \geq 0$ der Glättungsparameter bezeichnet wird. Über den Glättungsparameter α wird der Kompromiss zwischen

Anpassung an die gegebenen Daten und der durch die Penalisierung beschriebenen Glattheit gesteuert. Für $\alpha = 0$ erhält man die unpenalisierte Schätzung zurück und für große Werte von α eine im Sinne der Penalisierung glatte Funktion. Der Faktor $\frac{1}{2}$ wird lediglich verwendet, um mathematisch ‚schönere' Formeln zu erhalten. Man beachte, dass für normalverteilten Response häufig auch das Produkt $\lambda = \alpha\sigma^2$ als Glättungsparameter bezeichnet wird.

Mögliche Penalisierungen können nun beispielsweise basierend auf Informationskriterien oder Ableitungen der zu schätzenden Funktion f definiert werden (vergleiche O'Sullivan (1986) und den folgenden Abschnitt zu Glättungssplines). Mathematisch besonders vorteilhaft sind Penalisierungen, die sich als quadratische Formen im zugrunde liegenden Parametervektor ζ schreiben lassen, das heißt Penalisierungen der Form

$$J(\zeta) = \zeta' K \zeta,$$

wobei mit K eine Strafmatrix bezeichnet wird. Dann besitzt die penalisierte Likelihood nämlich eine ähnliche Form wie die Log-Posteriori in (2.12), wobei $Q(\nu)^{-1}$ durch K zu ersetzen ist, so dass zur Maximierung ein ähnlicher modifizierter Fisher-Scoring-Algorithmus wie in Kapitel 2.3.3 verwendet werden kann. Man beachte, dass die, zunächst rein formale, Ähnlichkeit der penalisierten Likelihood und der Log-Posteriori nicht nur zufälliger Natur ist, sondern sich auch inhaltlich interpretieren lässt. In Kapitel 3.3 wird dies ausgenutzt werden, um generalisierte geoadditive gemischte Modelle zu generalisierten linearen gemischten Modellen umzuschreiben.

Eine einfache Penalisierung, basierend auf Differenzen benachbarter B-Spline-Koeffizienten, schlagen Eilers & Marx (1996) vor und erweitern diesen Ansatz in Marx & Eilers (1998) auf die Modellierung und simultane Schätzung mehrerer nonparametrischer Funktionen. Zur Vereinfachung der Notation soll im Folgenden die Zahl der Basisfunktionen, mit deren Hilfe die Funktion f dargestellt werden kann, als r bezeichnet werden, das heißt, es gilt $r = m + l - 1$. Man erhält so für f die Darstellung

$$f(x) = \sum_{j=1}^{r} \zeta_j B_j^l(x) \text{ für } x \in [a, b]. \tag{3.4}$$

Zusätzlich ändert man dabei auch die entsprechenden Indizes der Regressionsparameter und Basisfunktionen.

Als Penalisierung erhält man bei Verwendung von Differenzen k-ter Ordnung

$$J(\zeta) = \sum_{j=k+1}^{r} (\Delta_k(\zeta_j))^2,$$

wobei der Differenzenoperator $\Delta_k(\zeta_j)$ rekursiv definiert ist durch

$$\begin{aligned} \Delta_1(\zeta_j) &= \zeta_j - \zeta_{j-1}, \\ \Delta_k(\zeta_j) &= \Delta_{k-1}(\zeta_j) - \Delta_{k-1}(\zeta_{j-1}). \end{aligned}$$

Speziell für $k = 2$ erhält man $\Delta_2(\zeta_j) = \zeta_j - 2\zeta_{j-1} + \zeta_{j-2}$.

In Matrixschreibweise lässt sich der Penalisierungsterm darstellen als

$$J(\zeta) = \zeta' D_k^{r\prime} D_k^r \zeta$$

mit den Differenzenmatrizen

$$\begin{aligned} D_1^r &= \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix} \qquad (r-1 \times r) \\ D_2^r &= \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix} \qquad (r-2 \times r) \\ D_k^r &= D_1^{r-k+1} D_{k-1}^r \qquad (r-k \times r). \end{aligned}$$

Insbesondere erhält man also, wie gewünscht, eine quadratische Form in ζ als Penalisierung mit Strafmatrix $K = D_k^{r\prime} D_k^r$.

Eine basierend auf dieser Penalisierung geschätzte Funktion f wird im Folgenden als P-Spline bezeichnet. Obwohl der Name P-Spline natürlich auch bei allgemeineren Penalisierungen anwendbar wäre, soll darunter stets der auf Differenzen basierende Penalisierungsansatz verstanden werden. Ein P-Spline lässt sich somit über den Grad l der zugrunde liegenden B-Spline-Basis, die Zahl m der verwendeten Knoten und die Ordnung k der zur Penalisierung verwendeten Differenzen charakterisieren.

Um die Verwendung von Differenzen als Penalisierung zu motivieren, soll noch einmal die Darstellung von f als Polynom-Spline mit Hilfe der B-Spline-Basis genauer betrachtet werden. Die Funktion f lässt sich demnach wie in (3.4) darstellen als Summe der mit den Koeffizienten ζ_j skalierten Basisfunktionen. Da alle B-Spline-Basisfunktionen für äquidistante Knoten die gleiche funktionale Gestalt haben und nur auf der x-Achse verschoben sind, führen nahezu gleiche Koeffizienten zur gleichen Skalierung aller Basisfunktionen und damit aufgrund der Zerlegung der Einheit durch B-Splines zu einer annähernd horizontalen Schätzung von f. Penalisiert man also mit Hilfe der quadrierten Differenzen erster Ordnung, so erhält man bei großem Glättungsparameter α eine glatte, nahezu horizontale Funktion. Penalisiert man durch die quadrierten Differenzen zweiter Ordnung, dann wird der Penalisierungsterm minimiert, wenn sich die benachbarten B-Spline-Koeffizienten jeweils nur um einen konstanten Wert unterscheiden. Dies führt dann bei großem Glättungsparameter zu einer Schätzung von f nahe der Regressionsgeraden von y bezüglich x. Allgemein erhält man das folgende Ergebnis: Für $\alpha \to \infty$ erhält man ein Polynom vom Grad $k-1$ als Schätzung für f. Man vergleiche Eilers & Marx (1996) für einen Beweis dieser Eigenschaft.

Durch die Schätzung von f mit Hilfe von P-Splines wird das Problem der Knotenwahl zurückgeführt auf die Wahl zweier Parameter, nämlich der Zahl der Knoten und des Glättungsparameters. Die Position der Knoten ist bereits durch die Knotenzahl eindeutig festgelegt, da bei P-Splines von äquidistanten Knoten ausgegangen wird. Die Zahl der Knoten spielt bei der Schätzung nur eine untergeordnete Rolle, solange genügend Knoten verwendet werden, um eine ausreichende Flexibilität der Schätzung zu erreichen. In der Regel werden zwischen 20 und 40 Knoten verwendet, bei stark oszillierenden Funktionen kann auch die Verwendung einer größeren Knotenzahl sinnvoll sein. Man vergleiche Ruppert (2002) für eine Möglichkeit die Knotenzahl optimal zu wählen und eine kurze Diskussion des Einflusses der Knotenzahl auf die Qualität der Schätzung.

Zusätzlich zum Glättungsparameter sind noch der Grad der verwendeten B-Spline-Basis und die Ordnung der Differenzen zu wählen, auf denen die Penalisierung beruht. In den allermeisten Datensituationen dürften auf einer B-Spline-Basis vom Grad 3 basierende P-Splines mit zweiten Differenzen geeignete Resultate liefern. Die Verwendung erster Differenzen führt dagegen häufig zu recht rauen Schätzungen. Eine ähnliche Aussage gilt für den Grad des P-Splines: je ge-

ringer der Grad, desto rauer fällt in der Regel die Schätzung aus. Für P-Splines vom Grad 0 erhält man nämlich eine Treppenfunktion und für P-Splines vom Grad 1 einen Polygonzug. Man vergleiche hierzu auch die Simulationsstudie in Brezger (2000).

Man beachte, dass sich über P-Splines vom Grad $l = 0$ auch die von Whittaker (1922/23) eingeführten und häufig in der Analyse von Zeitreihen verwendeten Random Walk-Modelle als P-Splines auffassen lassen. Um dies zu erkennen, betrachte man äquidistante Ausprägungen der Kovariablen und verwende ebensoviele Knoten, wie Beobachtungen vorhanden sind. Mit P-Splines vom Grad 0 erhält man dann bei Verwendung einer auf ersten Differenzen beruhenden Penalisierung ein RW(1)-Modell und bei Verwendung zweiter Differenzen ein RW(2)-Modell.

Glättungssplines

Nun soll noch die Modellierung von f mit Hilfe von Glättungssplines dargestellt werden, weil diese ebenfalls auf einem Penalisierungsansatz beruhen. Gleichzeitig soll gezeigt werden, worin die Vorteile von P-Splines gegenüber Glättungssplines bestehen.

Bei der Herleitung von Glättungssplines geht man zunächst von der folgenden Optimierungsaufgabe aus: Gesucht ist diejenige Funktion aus der Menge aller zweimal stetig differenzierbaren Funktionen, die das penalisierte Log-Likelihood-Kriterium

$$l_p(f) = l(f) - \frac{1}{2}\alpha \int (f''(x))^2 dx$$

maximiert. Starke Schwankungen der Funktion f werden hier mit Hilfe der zweiten Ableitung penalisiert, da das Integral der quadrierten zweiten Ableitung als Maß für die Krümmung einer Funktion interpretiert werden kann. Eilers & Marx (1996) zeigen, dass sich die Penalisierung mit Hilfe quadrierter zweiter Differenzen als Approximation der Penalisierung über die zweite Ableitung auffassen lässt, so dass man eine weitere Rechtfertigung der Verwendung von auf Differenzen beruhenden Penalisierungen erhält.

Als Lösung des Optimierungsproblems erhält man einen natürlichen kubischen Spline mit Knotenmenge $\Omega = \{x_{(1)}, \ldots, x_{(r)}\}$, wobei mit $a < x_{(1)} < \ldots < x_{(r)} < b$

die r verschiedenen, geordneten Werte von x bezeichnet werden. Unter einem natürlichen kubischen Spline versteht man dabei einen kubischen Polynom-Spline g, der zusätzlich die Randbedingungen $g''(a) = g'''(a) = 0$ und $g''(b) = g'''(b) = 0$ erfüllt, das heißt in den Intervallen $[a, x_{(1)}]$ und $[x_{(r)}, b]$ ist g linear. Aufgrund der Randbedingungen bilden die natürlichen kubischen Splines einen r-dimensionalen Unterraum des Vektorraums der zweimal stetig differenzierbaren Funktionen und auch einen Unterraum von $S_3(\{x_{(1)}, \ldots, x_{(r)}\})$, dem Vektorraum der kubischen Polynom-Splines mit Knoten an allen verschiedenen Designpunkten. Wegen der zweiten Eigenschaft lässt sich ein natürlicher kubischer Spline mit Hilfe einer modifizierten B-Spline-Basis darstellen (vergleiche Biller (2000c) Kapitel 3.1).

Alternativ lässt sich f auch über den Vektor $\zeta = (\zeta_1, \ldots, \zeta_r)'$ der r verschiedenen Funktionswerte von f an den Designpunkten $x_{(1)}, \ldots, x_{(r)}$ parametrisieren. Da man so die Möglichkeit erhält, auch Glättungssplines in Kapitel 3.3 zu einem Modell mit zufälligen Effekten zu reparametrisieren, soll hier nur diese Parametrisierung betrachtet und nicht näher auf die Darstellung mit Hilfe der modifizierten B-Spline-Basis eingegangen werden.

Für f ergibt sich die Darstellung $f = B\zeta$ mit der $(n \times r)$-Designmatrix $B = (b_{ij})_{i=1,\ldots,n, j=1,\ldots,r}$ und

$$b_{ij} = \begin{cases} 1 & \text{falls } x_i = x_{(j)} \\ 0 & \text{sonst.} \end{cases}$$

Der Strafterm $\int (f''(x))^2 dx$ lässt sich dann umschreiben zu $\zeta' K \zeta$ mit der $r \times r$-Strafmatrix $K = E' C^{-1} E$ und den auf den Differenzen $h_j = x_{(j+1)} - x_{(j)}$, also auf den Abständen der verschiedenen, geordneten x-Werte basierenden Matrizen E und C, für die gilt (vergleiche Reinsch (1967) und Fahrmeir & Tutz (2001) Seite 179/180): E ist eine obere Tridiagonalmatrix der Dimension $(r-2 \times r)$,

$$E = \begin{pmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & & & \\ & \frac{1}{h_2} & -\left(\frac{1}{h_2} + \frac{1}{h_3}\right) & \frac{1}{h_3} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \frac{1}{h_{r-2}} & -\left(\frac{1}{h_{r-2}} + \frac{1}{h_{r-1}}\right) & \frac{1}{h_{r-1}} \end{pmatrix}$$

und C ist eine symmetrische $(r-2 \times r-2)$ Bandmatrix der Bandweite 1

$$C = \frac{1}{6}\begin{pmatrix} 2(h_1+h_2) & h_2 & & & & & \\ h_2 & 2(h_2+h_3) & h_3 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & h_{r-3} & 2(h_{r-3}+h_{r-2}) & h_{r-2} & \\ & & & & h_{r-2} & 2(h_{r-2}+h_{r-1}) \end{pmatrix}.$$

Mit dieser Darstellung des Penalisierungsterms erhält man den bezüglich ζ zu maximierenden Ausdruck

$$l_p(\zeta) = l(\zeta) - \frac{1}{2}\alpha\zeta' K\zeta.$$

Der wesentliche Unterschied zwischen Glättungssplines und P-Splines besteht nun weniger in der unterschiedlichen Art der Penalisierung, auch wenn die Strafmatrix für P-Splines deutlich einfacher zu bestimmen ist, als vielmehr in der Dimension der zu bestimmenden Matrizen und Vektoren. Während für P-Splines die Dimension der Strafmatrix im Wesentlichen durch die Zahl der verwendeten Knoten bestimmt ist und damit nahezu unabhängig ist vom Stichprobenumfang, wird die Dimension der Strafmatrix für Glättungssplines durch die Zahl der unterschiedlichen Beobachtungen von x bestimmt. Häufig ist diese in etwa von der gleichen Größenordnung wie der Stichprobenumfang n, was bei einer größeren Beobachtungszahl zu erheblichen numerischen Problemen führen kann. Diese entstehen insbesondere dann, wenn nicht nur eine Funktion zu schätzen ist, sondern die Abhängigkeit von mehreren Kovariablen nonparametrisch modelliert werden soll.

P-Splines sind also ein so genanntes Low-Rank-Verfahren der nonparametrischen Regression im Sinne von Hastie (1996). Dort werden Low-Rank-Verfahren über eine Approximation der Schätzung von Glättungssplines mit Hilfe der dominierenden Eigenwerte der Glättungsmatrix $B(B'B + \lambda K)B'$ definiert, was eine mathematisch anspruchsvollere Herleitung zur Folge hat als bei P-Splines. Das Verfahren beruht aber auf einer ähnlichen Idee: Die Schätzung der Funktion f wird zurückgeführt auf die Schätzung eines Parameters relativ geringer Dimension. Insbesondere ist die Dimension des zu schätzenden Parametervektors nahezu unabhängig von der Zahl der Beobachtungen.

Wie man sowohl für P-Splines als auch für Glättungssplines gesehen hat, kann die Schätzung einer nonparametrischen modellierten Funktion auf die Schätzung

eines Parametervektors und damit auf die Maximierung einer penalisierten Likelihood zurückgeführt werden. Dies ist auch in additiver Modellierung, das heißt bei der nonparametrischen Modellierung mehrerer Funktionen möglich. Bei Verwendung von Glättungssplines ist dann die direkte Lösung der resultierenden Gleichungssysteme über Standardverfahren aufgrund der großen Dimension, die ungefähr von der Ordnung $s \cdot n$ ist, nicht mehr möglich. Stattdessen maximiert man die penalisierte Likelihood indirekt über Backfitting, wie in Abschnitt 3.2 beschrieben. Bei Verwendung eines Low-Rank-Verfahrens wie P-Splines ist es dagegen möglich, die penalisierte Likelihood direkt und mit Standardverfahren zu maximieren. Die Verwendung von P-Splines oder ähnlicher Low-Rank-Verfahren besitzt damit die folgenden Vorteile (Marx & Eilers 1998):

- Die Schätzung generalisierter additiver Modelle wird zurückgeführt auf die Schätzung eines generalisierten linearen Modells mit penalisierter Likelihood.
- Die resultierenden Gleichungssysteme besitzen eine (relativ) geringe Dimension. Damit können alle nonparametrischen Funktionen direkt, ohne Backfitting geschätzt werden.
- Die Schätzung für das gesamte Modell wird in relativ wenigen Parametern zusammengefasst, so dass insbesondere die Vorhersage für neue Beobachtungen erleichtert wird.
- Standardfehler und Regressionsdiagnostika können vergleichsweise einfach berechnet werden. Man vergleiche hierzu auch Marx & Eilers (1998) Abschnitt 5.

3.1.2 Markov-Zufallsfelder

Nun soll eine Möglichkeit vorgestellt werden, in den Daten vorhandene räumliche Informationen geeignet bei der Analyse zu berücksichtigen und damit dem Problem räumlich korrelierter Daten zu begegnen. Dabei soll zunächst wieder von dem einfachen Modell

$$\eta_i = f_{spat}(R_i)$$

ausgegangen werden. Zusätzlich wird angenommen, dass nur eine endliche Menge von möglichen räumlichen Positionen R_i vorhanden ist, deren Anzahl mit r

bezeichnet wird. Der Einfachheit halber soll $R_i \in \{1, \ldots, r\}$ gelten, das heißt, die räumlichen Positionen sind durch die Ziffern 1 bis r bezeichnet. Bei den Positionen R_i soll es sich entweder um Punkte auf einem diskreten, zweidimensionalen Gitter oder um eine Menge zusammenhängender Regionen handeln. Beispielsweise sind die Positionen einzelner Bäume in der Analyse der Waldschadensdaten in Kapitel 6.2 Punkte eines irregulären Gitters, während die räumliche Information in der Analyse des Münchner Mietspiegels durch die Zugehörigkeit zu bestimmten Bezirksvierteln, also Regionen gegeben ist.

Es soll nun davon ausgegangen werden, dass der räumliche Effekt an einer bestimmten räumlichen Position durch den Parameter ζ_j beschrieben wird. Die einzelnen Parameter werden zusammengefasst im r-dimensionalen Vektor $\zeta = (\zeta_1, \ldots, \zeta_r)'$. Der räumliche Effekt $f_{spat} = (f_{spat}(R_1), \ldots, f_{spat}(R_n))'$ lässt sich dann schreiben als $f_{spat} = B\zeta$, wobei $B = (b_{ij})_{i=1,\ldots,n, j=1,\ldots,r}$ eine $n \times r$-Designmatrix ist, mit

$$b_{ij} = \begin{cases} 1 & \text{falls } R_i = j \\ 0 & \text{sonst.} \end{cases}$$

Um eine räumliche Struktur der Daten zu modellieren, geht man nun davon aus, dass sich benachbarte Positionen relativ wenig in ihrem räumlichen Effekt unterscheiden. Dazu benötigt man zunächst geeignete Definitionen benachbarter Positionen für die verschiedenen Datensituationen.

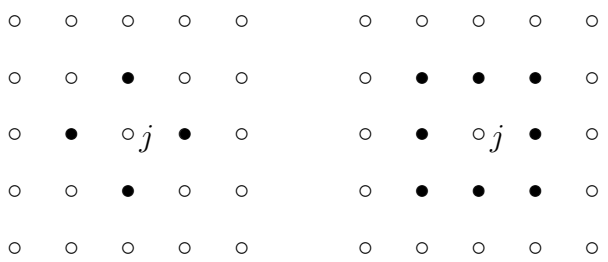

Abbildung 3.6: Nachbarschaftsdefinitionen für Punktdaten auf einem regulären Gitter. Die Nachbarn der Position j sind definiert durch die vier beziehungsweise acht nächsten Positionen auf dem Gitter und sind als schwarze Punkte wiedergegeben.

Zwei Nachbarschaftsdefinitionen für Punktdaten auf regulären Gittern sind in Abbildung 3.6 wiedergegeben. Dabei wurden offensichtliche Möglichkeiten zur

Definition benachbarter Standorte über nächste Nachbarn verwendet. Handelt es sich dagegen um irreguläre Gitter, so definiert man in der Regel zwei Standorte als benachbart, wenn ihre Entfernung einen bestimmten Wert unterschreitet. Eine solche Definition wird auch bei der Analyse der Waldschadensdaten in Kapitel 6.2 zugrunde gelegt.

Zur Definition der Nachbarschaft zweier Regionen lässt sich ebenfalls die Distanz der Zentroide der Regionen heranziehen. Zwei Regionen gelten dann als benachbart, wenn die Entfernung ihrer Zentroide einen bestimmten Wert unterschreitet. Eine alternative Definition beruht auf gemeinsamen Grenzen: Zwei Regionen werden als benachbart betrachtet, wenn sie gemeinsame Grenzen besitzen. Dieses Verfahren wird in der Regel bevorzugt und auch zur Analyse des Münchner Mietspiegels verwendet. Problematisch sind dabei lediglich Regionen, die mit keiner anderen Region gemeinsame Grenzen besitzen und die gegebenenfalls aus der Analyse auszuschließen sind. Dies wird beispielsweise im Rahmen der Analyse des Mortalitätsrisikos bezüglich Mundhöhlenkrebs in Deutschland in Kapitel 6.1 für die Insel Rügen der Fall sein.

In einer bayesianischen Betrachtungsweise erreicht man nun durch geeignete Priori-Annahmen für die Parameter $\zeta_1, \ldots, \zeta_r$, dass sich die Parameterwerte benachbarter Regionen nicht zu stark unterscheiden. Der bayesianische Ansatz lässt sich jedoch in Analogie zu P-Splines auch frequentistisch interpretieren. Im Folgenden wird zunächst die bayesianische Betrachtungsweise angenommen, die üblicherweise in der Herleitung von Markov-Zufallsfeldern verwendet wird.

Bezeichnet man mit δ_j die Menge aller Nachbarn von Region j und mit $n_j = |\delta_j|$ die Anzahl der Nachbarn von Region j, so nimmt man an, dass der bedingte Erwartungswert von ζ_j, gegeben die Parameter der Nachbarn, das gewichtete arithmetische Mittel der Nachbarparameter ist. Dies stellt eine Verallgemeinerung eines Random Walks erster Ordnung auf zweidimensionale Daten dar und wird als Annahme eines Markov-Zufallsfeldes für die Parameter ζ_j bezeichnet. Genauer nimmt man an, dass

$$\zeta_j | \zeta_l, l \in \delta_j \sim \mathrm{N}\left(\sum_{l \in \delta_j} \frac{w_{jl}}{w_{j+}} \zeta_l, \frac{\tau}{w_{j+}} \right) \tag{3.5}$$

gilt, wobei w_{jl} bekannte Gewichte mit $w_{jl} = w_{lj}$ sind und $w_{j+} = \sum_{l \in \delta_j} w_{jl}$ die Summe dieser Gewichte bezeichnet. Der Parameter $\tau \geq 0$ ist ein zusätzlicher

Varianzparameter, der die Glattheit des räumlichen Effekts steuert. Der inverse Varianzparameter $\alpha = 1/\tau$ lässt sich dann wie der Glättungsparameter bei P-Splines interpretieren: Bei großem α sind die Abweichungen des Parameters ζ_j vom bedingten Erwartungswert nur sehr gering, für den Grenzfall $\alpha \to \infty$ erhält man einen konstanten räumlichen Effekt. Ist α dagegen klein und somit umgekehrt der Varianzparameter τ groß, so sind starke Abweichungen des Parameters ζ_j vom bedingten Erwartungswert möglich, was zu einer eher rauen Schätzung führt.

Zur Wahl der Gewichte w_{jl} existieren verschiedene Vorschläge:

- $w_{jl} = 1$, das heißt, ζ_j ist das ungewichtete arithmetische Mittel der Nachbarparameter und $w_{j+} = n_j$.
- $w_{jl} = c \cdot \exp(-d(j,l))$ mit $d(j,l)$ als euklidischem Abstand der Zentroide von Region j und Region l und der Normierungskonstanten c. Die Gewichte sind also umgekehrt proportional zum Abstand der Zentroide der benachbarten Regionen.
- $w_{jl} = c \cdot$gemeinsame Grenzlänge, wobei c wieder eine Normierungskonstante ist.

Häufig wählt man die Normierungskonstante c in Analogie zum ersten Vorschlag so, dass $\sum_{l \in \delta_j} w_{jl} = n_j$ gilt. Im Folgenden wird meist von $w_{jl} = 1$ ausgegangen.

Durch (3.5) sind die bedingten Priori-Verteilungen der Parameter ζ_j gegeben. Es stellt sich nun die Frage, inwiefern die gemeinsame Priori-Verteilung durch Vorgabe der bedingten Prioris festgelegt ist. Man kann zeigen (Besag 1974), dass bereits durch die Vorgabe der bedingten Priori-Verteilungen die gemeinsame Priori-Verteilung bis auf eine Normierungskonstante eindeutig festgelegt ist, solange bei der Wahl der bedingten Prioris bestimmte Restriktionen beachtet werden. Im hier betrachteten Spezialfall des durch (3.5) beschriebenen Markov-Zufallsfelds wurde die einzige Restriktion $w_{jl} = w_{lj}$ bereits berücksichtigt.

Als gemeinsame Priori-Verteilung erhält man aus (3.5) eine singuläre Normalverteilung, deren Dichte proportional ist zu

$$\exp\left(-\frac{1}{2\tau}\zeta' K \zeta\right),$$

wobei die $r \times r$-Matrix $K = (k_{jl})_{j,l=1,\ldots,r}$ bestimmt ist durch

$$\begin{aligned} k_{jj} &= w_{j+} \\ k_{jl} &= \begin{cases} -w_{jl} & l \in \delta_j \\ 0 & \text{sonst.} \end{cases} \end{aligned}$$

Die Verteilung ist singulär, da im Allgemeinen $\mathrm{rg}(K) = r-1$ gilt. Treten Regionen ohne Nachbarn auf, so besitzt K sogar ein entsprechend größeres Rangdefizit.

Als Präzisionsmatrix der Priori-Verteilung erhält man die Matrix $\frac{1}{\tau}K$, auf deren Hauptdiagonalen die bedingten inversen Varianzen $\frac{w_{j+}}{\tau}$ stehen. Die Nebendiagonalen geben die Stärke der bedingten Korrelation von ζ_j und ζ_l an. Insbesondere sind ζ_j und ζ_l bedingt auf die übrigen Parameter unkorreliert und aufgrund der gemeinsamen Normalverteilung auch bedingt unabhängig, wenn $k_{jl} = 0$ gilt.

Als bayesianische Schätzer für ζ bieten sich wieder der Modus oder der Erwartungswert der Posteriori-Verteilung an. Wie im generalisierten linearen gemischten Modell ist die Posteriori aber nur im Normalverteilungsfall analytisch und nur bei einer geringen Zahl von Regionen numerisch zu bestimmen. Einen Ausweg zur Bestimmung des Posteriori-Erwartungswertes bieten MCMC-Verfahren (Fahrmeir & Lang (2001a,2001b)), die zusätzlich auch die simultane Schätzung des Varianzparameters τ erlauben. In dieser Arbeit soll jedoch wieder der Posteriori-Modus als Schätzer verwendet werden, der sich durch Maximierung der Log-Posteriori

$$\log(p(\zeta|y)) = l(\zeta) - \frac{1}{2\tau}\zeta' K \zeta \tag{3.6}$$

ohne numerische Integration auch für nicht normalverteilte Daten bestimmen lässt.

Betrachtet man die Log-Posteriori (3.6), so besitzt diese die Form der penalisierten Likelihood $l_p(\zeta)$, die zur Schätzung von P-Splines verwendet wurde. Man kann daher Markov-Zufallsfelder auch in einem frequentistischen Ansatz betrachten, indem man, analog zu $D_k^{r\prime} D_k^r$ bei P-Splines, die Matrix K als Strafmatrix betrachtet, die zu starke Abweichungen zwischen den Parametern benachbarter Regionen penalisiert. Eventuell richtet sich die Stärke der Penalisierung dabei noch nach den Gewichten w_{jl}. Die Schätzung erfolgt dann wie bei P-Splines über die Maximierung der penalisierten Likelihood

$$l_p(\zeta) = l(\zeta) - \frac{1}{2}\alpha\zeta' K \zeta$$

mit dem Glättungsparameter $\alpha = \frac{1}{\tau}$.

Vergleicht man Markov-Zufallsfelder mit anderen Verfahren zur Schätzung räumlicher Effekte, beispielsweise dem Kriging (Cressie 1993), so erfüllen diese eine ähnliche Eigenschaft wie P-Splines, nämlich die Unabhängigkeit der Dimension des Parameters ζ vom Stichprobenumfang. Während diese Dimension beim Kriging in der Regel nahe beim Stichprobenumfang n liegt, ist bei Verwendung eines Markov-Zufallsfeldes stets nur der r-dimensionale Vektor ζ zu schätzen. Markov-Zufallsfelder können also ebenso wie P-Splines als ein Low-Rank-Verfahren angesehen werden, was ähnliche Vorteile wie bei der Schätzung von P-Splines mit sich bringt. Eine weitere Alternative zur Schätzung räumlicher Effekte bietet die in Lang & Brezger (2002) beschriebene Erweiterung von P-Splines auf die Analyse von Interaktionen zweier metrischer Kovariablen, die ebenfalls als Low-Rank-Verfahren aufgefasst werden kann.

3.2 Schätzung bei gegebenen Hyperparametern

Nun soll die Schätzung aller Modellkomponenten eines generalisierten geoadditiven gemischten Modells bei fest vorgegebenen Glättungs- und Varianzparametern behandelt werden. Dies soll lediglich durch einen sehr knappen Überblick geschehen, da in Abschnitt 3.3 und 3.4 ein alternativer Schätzansatz präsentiert wird, der nicht nur die Bestimmung der Regressionsparameter, sondern auch der Glättungs- und Varianzparameter erlaubt.

Zunächst stellt man fest, dass sich mit Hilfe von P-Splines und Markov-Zufallsfeldern der Prädiktor

$$\eta = X_{par}\beta_{par} + f_1 + \ldots + f_s + f_{spat} + Z_{ran}b_{ran}$$

umschreiben lässt zu

$$\eta = X_{par}\beta_{par} + B_1\zeta_1 + \ldots + B_s\zeta_s + B_{spat}\zeta_{spat} + Z_{ran}b_{ran}, \tag{3.7}$$

wobei die Matrizen B_j, $j = 1, \ldots, s$ aus den B-Spline-Basisfunktionen bestehen, die an den beobachteten Ausprägungen von x_j ausgewertet werden. Die Parameter ζ_j, $j = 1, \ldots, s$ sind jeweils r_j-dimensionale Parametervektoren, die die entsprechenden Regressionskoeffizienten der Basisfunktionen enthalten. Die

Matrix B_{spat} ist, wie im Abschnitt zu Markov-Zufallsfeldern definiert, eine Pseudodesignmatrix, die der jeweiligen Beobachtung den entsprechenden Wert der räumlichen Funktion zuordnet und ζ_{spat} enthält die r_{spat} verschiedenen Funktionswerte der räumlichen Funktion.

Sowohl P-Splines als auch Markov-Zufallsfelder werden über eine Penalisierung der jeweiligen Likelihood modelliert. Berücksichtigt man zusätzlich noch die zufälligen Effekte b_{ran}, so erhält man insgesamt als zu maximierenden Ausdruck

$$l(\beta_{par}, \zeta_1, \ldots, \zeta_s, \zeta_{spat}, b_{ran}) \\ - \frac{1}{2}\sum_{j=1}^{s} \alpha_j \zeta_j' K_j \zeta_j - \frac{1}{2}\alpha_{spat}\zeta_{spat}' K_{spat}\zeta_{spat} - \frac{1}{2} b_{ran}' Q_{ran}(\nu_{ran})^{-1} b_{ran},$$

wobei mit K_j, $j = 1, \ldots, s$ und K_{spat} die Penalisierungsmatrizen und mit α_j, $j = 1, \ldots, s$ sowie α_{spat} die zugehörigen Glättungsparameter der jeweiligen Effekte bezeichnet werden.

Die Ableitung dieser penalisierten Likelihood nach den einzelnen Parametern liefert die Schätzgleichungen

$$\begin{aligned} X_{par}' W(\eta) D(\eta)^{-1}(y - \mu(\eta)) &= 0 \\ B_j' W(\eta) D(\eta)^{-1}(y - \mu(\eta)) - \alpha_j K_j \zeta_j &= 0 \qquad j = 1, \ldots, s \\ B_{spat}' W(\eta) D(\eta)^{-1}(y - \mu(\eta)) - \alpha_{spat} K_{spat}\zeta_{spat} &= 0 \\ Z_{ran}' W(\eta) D(\eta)^{-1}(y - \mu(\eta))) - Q_{ran}(\nu_{ran})^{-1} b_{ran} &= 0, \end{aligned}$$

wobei $D(\eta)$ und $W(\eta)$ wie in (2.15) und (2.18) definiert sind.

Lösungen dieser Gleichungen erhält man ähnlich wie bei der Verwendung des Fisher-Scorings in generalisierten linearen gemischten Modellen über die wiederholte Lösung des folgenden Gleichungssystems (vergleiche Lin & Zhang (1999)):

$$\begin{pmatrix} I & S_{par}B_1 & \ldots & S_{par}B_s & S_{par}B_{spat} & S_{par}Z_{ran} \\ S_1 X_{par} & I & \ldots & S_1 B_s & S_1 B_{spat} & S_1 Z_{ran} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ S_s X_{par} & S_s B_1 & \ldots & I & S_s B_{spat} & S_s Z_{ran} \\ S_{spat} X_{par} & S_{spat} B_1 & \ldots & S_{spat} B_s & I & S_{spat} Z_{ran} \\ S_{ran} X_{par} & S_{ran} B_1 & \ldots & S_{ran} B_s & S_{ran} B_{spat} & I \end{pmatrix} \begin{pmatrix} \beta_{par} \\ \zeta_1 \\ \vdots \\ \zeta_s \\ \zeta_{spat} \\ b_{ran} \end{pmatrix} = \begin{pmatrix} S_{par}\tilde{y} \\ S_1\tilde{y} \\ \vdots \\ S_s\tilde{y} \\ S_{spat}\tilde{y} \\ S_{ran}\tilde{y} \end{pmatrix},$$

wobei mit $\tilde{y} = \tilde{y}(\eta)$ die in (2.17) definierten Arbeitsbeobachtungen bezeichnet

werden und die jeweiligen Schätzmatrizen definiert sind durch

$$\begin{aligned} S_{par} &= (X'_{par}W(\eta)X_{par})^{-1}X'_{par}W(\eta) \\ S_j &= (B'_jW(\eta)B_j + \alpha_jK_j)^{-1}B'_jW(\eta), \qquad j = 1,\ldots,s \\ S_{spat} &= (B'_{spat}W(\eta)B_{spat} + \alpha_{spat}K_{spat})^{-1}B'_{spat}W(\eta) \\ S_{ran} &= (Z'_{ran}W(\eta)Z_{ran} + Q_{ran}(\nu_{ran})^{-1})^{-1}Z'_{ran}W(\eta). \end{aligned}$$

Bei Verwendung von Glättungssplines oder Kriging, die ähnliche Schätzgleichungen zur Folge haben, kann das Gleichungssystem aufgrund der hohen Dimension in der Regel nicht direkt gelöst werden. Stattdessen verwendet man iterative Verfahren wie das Backfitting, um Schätzer für die Parameter zu erhalten. Dabei werden wiederholt mit Hilfe der Matrizen S_j Schätzungen der einzelnen Modellkomponenten bestimmt, die auf den bezüglich allen anderen Effekten gebildeten partiellen Residuen eines additiven Modells für die Arbeitsbeobachtungen $\tilde{y}(\eta)$ beruhen. Um den zugehörigen Algorithmus definieren zu können, soll zunächst die Schätzung eines gewichteten geoadditiven gemischten Modells unter der Normalverteilungsannahme für den Response y in einem Algorithmus zusammengefasst werden:

Algorithmus 3 (Gewichtetes Backfitting im geoadditiven gemischten Modell)

(i) Bestimme Startwerte $\hat{\beta}_{par}^{(0)}$, $\hat{\zeta}_1^{(0)},\ldots,\hat{\zeta}_s^{(0)}$, $\hat{\zeta}_{spat}^{(0)}$, $\hat{b}^{(0)}$ und setze $k = 0$.

(ii) Bilde

$$\begin{aligned} \hat{\beta}_{par}^{(k+1)} &= S_{par}\left(y - B_1\hat{\zeta}_1^{(k)} - \ldots - B_s\hat{\zeta}_s^{(k)} - B_{spat}\hat{\zeta}_{spat}^{(k)} - Z_{ran}\hat{b}_{ran}^{(k)}\right) \\ \hat{\zeta}_j^{(k+1)} &= S_j\Big(y - X_{par}\hat{\beta}_{par}^{(k+1)} - B_1\hat{\zeta}_1^{(k+1)} - \ldots - B_{j-1}\hat{\zeta}_{j-1}^{(k+1)} - B_{j+1}\hat{\zeta}_{j+1}^{(k)} - \\ &\quad \ldots - B_s\hat{\zeta}_s^{(k)} - B_{spat}\hat{\zeta}_{spat}^{(k)} - Z_{ran}\hat{b}_{ran}^{(k)}\Big), \qquad j = 1,\ldots,s \\ \hat{\zeta}_{spat}^{(k+1)} &= S_{spat}\left(y - X_{par}\hat{\beta}_{par}^{(k+1)} - B_1\hat{\zeta}_1^{(k+1)} - \ldots - B_s\hat{\zeta}_s^{(k+1)} - Z_{ran}\hat{b}_{ran}^{(k)}\right) \\ \hat{b}_{ran}^{(k+1)} &= S_{ran}\left(y - X_{par}\hat{\beta}_{par}^{(k+1)} - B_1\hat{\zeta}_1^{(k+1)} - \ldots - B_s\hat{\zeta}_s^{(k+1)} - B_{spat}\hat{\zeta}_{spat}^{(k+1)}\right). \end{aligned}$$

(iii) Bestimme ein geeignetes Abbruchkriterium $d(\eta^{(k)}, \eta^{(k+1)})$. Falls sich die Schätzer im letzten Schritt noch verändert haben, setze $k = k + 1$ und gehe zurück zu (ii). Falls keine Veränderung mehr eingetreten ist, beende den Algorithmus.

Um zentrierte Funktionsschätzungen $\hat{f}_j$, $j = 1, \ldots, s$ und $\hat{f}_{spat}$ zu erhalten, müssen in Schritt (ii) jeweils die sich aus den aktuellen Parameterschätzungen ergebenden Funktionsschätzungen geeignet zentriert werden. Man beachte, dass dabei auch die Schätzung für den Intercept so verändert werden muss, dass sich der additive Prädiktor η nicht verändert. Man vergleiche auch Hastie & Tibshirani (1990) Abschnitt 4.4 für eine detailliertere Behandlung des Backfitting-Algorithmus in additiven Modellen.

Mit Hilfe des Algorithmus zum gewichteten Backfitting lässt sich nun leicht die Schätzung eines generalisierten geoadditiven gemischten Modells zusammenfassen:

Algorithmus 4 (Backfitting im generalisierten geoadditiven gemischten Modell)

(i) Bestimme Startwerte $\hat{\beta}_{par}^{(0)}$, $\hat{\zeta}_1^{(0)}, \ldots, \hat{\zeta}_s^{(0)}$, $\hat{\zeta}_{spat}^{(0)}$, $\hat{b}_{ran}^{(0)}$ und setze $k = 1$.

(ii) Berechne die Arbeitsbeobachtungen

$$\tilde{y}(\eta^{(k)}) = \eta^{(k)} + D(\eta^{(k)})^{-1}(y - \mu(\eta^{(k)}))$$

und die Arbeitsgewichte

$$W(\eta^{(k)}) = D(\eta^{(k)})\Sigma(\eta^{(k)})^{-1}D(\eta^{(k)}).$$

(iii) Schätze mit Hilfe von Algorithmus 3 ein gewichtetes additives Modell, mit $\tilde{y}(\eta^{(k)})$ als abhängiger Variable und $W(\eta^{(k)})$ als Gewichten. Die Schätzung liefert $\hat{\beta}_{par}^{(k+1)}$, $\hat{\zeta}_1^{(k+1)}, \ldots, \hat{\zeta}_s^{(k+1)}$, $\hat{\zeta}_{spat}^{(k+1)}$, $\hat{b}_{ran}^{(k+1)}$.

(iv) Bestimme ein geeignetes Abbruchkriterium $d(\eta^{(k)}, \eta^{(k+1)})$. Falls sich die Schätzer im letzten Schritt noch verändert haben, setze $k = k + 1$ und gehe zurück zu (ii). Falls keine Veränderung mehr eingetreten ist, beende den Algorithmus.

Verwendet man statt Glättungssplines Low-Rank-Verfahren wie beispielsweise P-Splines, so ist die Dimension der resultierenden Gleichungssysteme in der Regel gering genug, um eine direkte Lösung zu ermöglichen. Man vergleiche Marx & Eilers (1998) für eine genauere Beschreibung der direkten Schätzung aller Modellparameter.

3.3 Reparametrisierung

Im vorigen Abschnitt wurde die Schätzung eines generalisierten geoadditiven gemischten Modells für gegebene Hyperparameter, also bei gegebenen Glättungs- und Varianzparametern behandelt. In der Regel sind diese Parameter aber unbekannt und müssen ebenfalls aus den Daten bestimmt werden.

Marx & Eilers (1998) empfehlen beispielsweise, die Glättungsparameter eines generalisierten additiven Modells mit Prädiktor

$$\eta = f_1 + \ldots + f_s = B_1\zeta_1 + \ldots + B_s\zeta_s$$

optimal bezüglich eines Informationskriteriums zu wählen. Speziell werden Akaikes Informationskriterium (AIC) und das Bayessche Informationskriterium (BIC) vorgeschlagen. Beide lassen sich in der Form

$$IC(\alpha) = \text{dev}(y, \zeta, \alpha) + \delta\,\text{df}(\alpha),$$

schreiben, wobei im Vektor $\zeta = (\zeta_1', \ldots, \zeta_s')'$ alle Regressionskoeffizienten des generalisierten additiven Modells und in $\alpha = (\alpha_1, \ldots, \alpha_s)'$ alle Glättungsparameter zusammengefasst seien. Mit $\text{dev}(y, \zeta, \alpha)$ wird die Devianz des Modells (Fahrmeir & Tutz (2001), Seite 50) bezeichnet und $\text{df}(\alpha)$ gibt die effektive Dimension des Modells wieder. Die effektive Dimension ist dabei definiert durch (Hastie & Tibshirani (1990), Seite 52-55)

$$\text{df}(\alpha) = \text{spur}(H),$$

mit

$$H = B(B'W(\zeta)B + K)^{-1}B'W(\zeta)$$

und $B = (B_1, \ldots, B_s)$ sowie $K = \text{blockdiag}(\alpha_1 K_1, \ldots, \alpha_s K_s)$. Für $\delta = 2$ erhält man Akaikes Informationskriterium und für $\delta = \log(n)$ das Bayessche Informationskriterium.

Sowohl $AIC(\alpha)$ als auch $BIC(\alpha)$ stellen einen Kompromiss zwischen Datentreue, gemessen durch die Devianz, und Komplexität des Modells, gemessen durch die effektive Dimension, dar. Das Vorgehen ist also der Idee der Penalisierung bei P-Splines verwandt. Als optimal bezüglich eines der beiden Informationskriterien werden diejenigen Glättungsparameter α bezeichnet, die $AIC(\alpha)$ beziehungsweise $BIC(\alpha)$ minimieren. Bei Verwendung des Bayesschen Informationskriteriums

werden also tendenziell weniger komplexe Modelle bevorzugt, da der Strafterm $\log(n) \cdot \mathrm{df}(\alpha)$ für $n > 7$ ein größeres Gewicht besitzt als $2 \cdot \mathrm{df}(\alpha)$.

Um das optimale Modell zu bestimmen, ist es in der Regel erforderlich, für eine große Zahl von Glättungsparametern die in Kapitel 3.2 beschriebene Schätzung durchzuführen und $AIC(\alpha)$ beziehungsweise $BIC(\alpha)$ zu bestimmen. Dieses Vorgehen kann jedoch bei einer größeren Zahl von Beobachtungen und mehreren nonparametrisch modellierten Funktionen zu erheblichen Rechenzeitproblemen führen, da die Suche eines optimalen Wertes auf einem s-dimensionalen Gitter notwendig ist.

Ein weiteres Maß, das häufig zur Bestimmung optimaler Glättungsparameter in generalisierten additiven Modellen verwendet wird, ist das generalisierte Kreuzvalidierungskriterium (GCV), das als

$$GCV(\alpha) = \frac{1}{n} \sum_{i=1}^{n} w_i(\hat{\eta}_i) \left(\frac{y_i - \mu_i(\hat{\eta}_i)}{1 - \bar{h}} \right)^2 .$$

mit $\bar{h} = \frac{1}{n} \mathrm{spur}(H)$ definiert ist. Man vergleiche Hastie & Tibshirani (1990) Seite 42-52 für eine theoretische Begründung des generalisierten Kreuzvalidierungskriteriums. Wood (2000) präsentiert eine effiziente Möglichkeit $GCV(\alpha)$-optimale Glättungsparameter in generalisierten additiven Modellen zu bestimmen, die allerdings einen relativ großen Implementierungsaufwand benötigt.

Die Bestimmung optimaler Glättungsparameter in generalisierten additiven Modellen über Informationskriterien oder über das generalisierte Kreuzvalidierungskriterium ist prinzipiell auch auf die Bestimmung von Glättungsparametern räumlicher Funktionen erweiterbar. In generalisierten geoadditiven gemischten Modellen bleibt aber weiterhin die Bestimmung der Varianzparameter der zufälligen Effekte fraglich. Im Folgenden soll daher eine Möglichkeit aufgezeigt werden, die beschriebenen Probleme bei der Bestimmung der Hyperparameter zu vermeiden, das heißt, es soll ein Verfahren beschrieben werden, das relativ einfach zu implementieren ist und die direkte Suche eines optimalen Wertes auf einem s-dimensionalen Gitter vermeidet. Darüberhinaus wird dieses Verfahren nicht nur die Bestimmung der Glättungsparameter nonparametrisch modellierter Funktionen, sondern auch die Schätzung des Glättungsparameters der räumlichen Funktion und der Varianzparameter der zufälligen Effekte ermöglichen. Über eine Reparametrisierung des generalisierten geoadditiven gemischten Modells zu einem

generalisierten linearen Modell mit unabhängigen zufälligen Effekten werden dabei die Glättungs- und Varianzparameter mit Hilfe von Methoden aus Kapitel 2 bestimmbar.

Die Darstellung nonparametrischer Funktionsschätzer über Modelle mit zufälligen Effekten wurde seit Ende der 90er Jahre in einer großen Zahl von Veröffentlichungen beschrieben und verwendet. Dazu gehören beispielsweise Lin & Zhang (1999), die generalisierte additive gemischte Modelle mit Hilfe von Glättungssplines behandeln, Kammann & Wand (2003), die geoadditive Modelle im Normalverteilungsfall mit Hilfe von linearen P-Splines und Kriging untersuchen oder Currie & Durban (2002), die eine Reihe von Ansätzen im Normalverteilungsfall mit Hilfe von P-Splines betrachten. Weitere Referenzen zur Darstellung von Glättungssplines als gemischte Modelle sind beispielsweise Wang (1998a), Wang (1998b) oder Zhang, Lin, Raz & Sowers (1998).

3.3.1 Einführendes Beispiel

In diesem Abschnitt soll nun nach Wand (2002) ein einfaches Beispiel mit Hilfe von stückweise konstanten P-Splines und einer Penalisierung über Differenzen erster Ordnung deutlich machen, wie nahe sich die nonparametrische Modellierung glatter Funktionen mit Hilfe von Penalisierungsansätzen und generalisierte lineare gemischte Modelle sind. Dazu wird im Folgenden nur der Normalverteilungsfall betrachtet und erst im nächsten Abschnitt wieder das allgemeinere Modell behandelt.

Betrachtet man das einfache Modell

$$y_i = f(x_i) + \varepsilon_i,$$

wobei für ε_i die gleichen Annahmen gelten sollen, wie im gewöhnlichen linearen Modell, und nimmt man an, dass die Funktion f durch einen Polynom-Spline angenähert werden kann, so lässt sich f mit Hilfe der Truncated Power Series-Basis darstellen als

$$f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_l x^l + \sum_{j=1}^{m-1} b_j (x - \xi_j)_+^l.$$

Dabei werden mit $\xi_1, \ldots, \xi_{m-1}$ wieder die Knoten des Polynom-Splines und mit $\beta_0, \ldots, \beta_l$ sowie $b_1, \ldots, b_{m-1}$ die zu den Basisfunktionen gehörenden Parameter

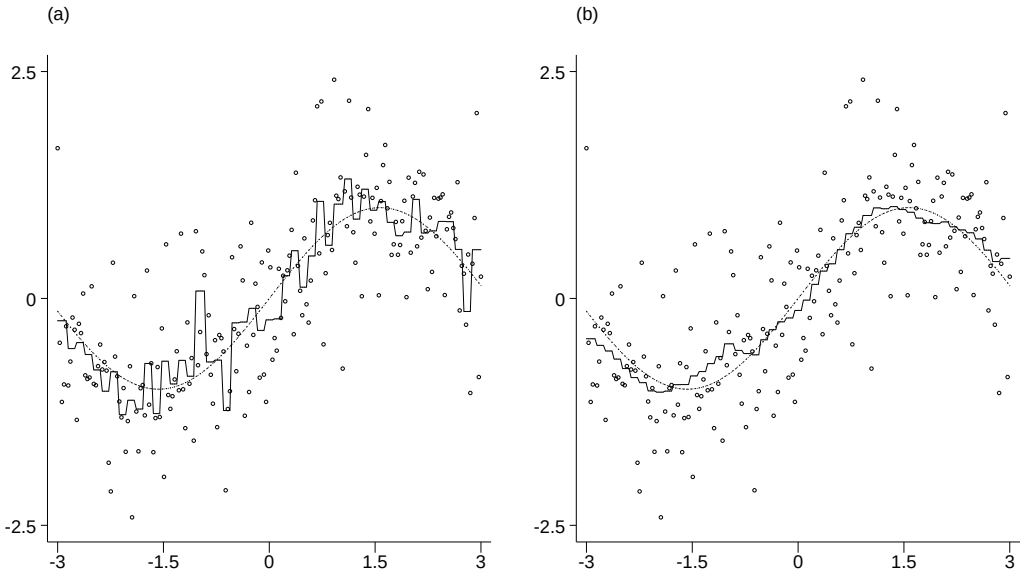

Abbildung 3.7: Nonparametrische Funktionsschätzung mit Polynom-Splines vom Grad 0 in Truncated Power Series-Darstellung. In Grafik (a) werden alle Parameter als fest betrachtet, in Grafik (b) werden die Parameter der abgeschnittenen Potenzen als zufällige Effekte betrachtet. Die durchgezogene Linie gibt die Funktionsschätzung, die gepunktete Linie die wahre Funktion wieder.

bezeichnet. Speziell für $l = 0$ erhält man

$$f(x) = \beta_0 + \sum_{j=1}^{m-1} b_j \mathbf{1}_{[\xi_j,\infty)}(x),$$

als Darstellung für f wobei, $\mathbf{1}_{[\xi_j,\infty)}(x)$ die Indikatorfunktion des Intervalls $[\xi_j, \infty)$ bezeichnet. In Matrixschreibweise ergibt sich so das Modell

$$y = X\beta_0 + Zb + \varepsilon$$

mit den Designmatrizen

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ und } Z = \begin{pmatrix} \mathbf{1}_{[\xi_1,\infty)}(x_1) & \dots & \mathbf{1}_{[\xi_{m-1},\infty)}(x_1) \\ \vdots & \ddots & \vdots \\ \mathbf{1}_{[\xi_1,\infty)}(x_n) & \dots & \mathbf{1}_{[\xi_{m-1},\infty)}(x_n) \end{pmatrix}.$$

sowie $b = (b_1, \dots, b_{m-1})'$.

Abbildung 3.7 (a) zeigt eine auf diesem Modell (ohne Penalisierung) basierende Schätzung mit $m = 50$ Knoten und $f(x) = \sin(x)$. Die Designpunkte wurden

äquidistant aus dem Intervall $[-3, 3]$ gewählt, der Stichprobenumfang beträgt $n = 200$ und die Varianz des Fehlerterms ist $\sigma^2 = 0.49$. Wie man sieht, fällt die Schätzung relativ rau aus. Betrachtet man den Parameter b dagegen als zufälligen Effekt mit der Annahme

$$b \sim \mathrm{N}(0, \tau I_{m-1})$$

und schätzt den Varianzparameter $\tau \geq 0$ über den Restricted-Maximum-Likelihood-Ansatz aus Kapitel 2, so erhält man die in Abbildung 3.7 (b) wiedergegebene Schätzung, die deutlich glatter ist und auch den Verlauf der wahren Funktion wesentlich genauer wiedergibt. Dass auch die Schätzung in Abbildung 3.7 (b) relativ rau ausfällt, liegt an der Verwendung von Polynom-Splines vom Grad $l = 0$, die eine Schätzung als stückweise konstante Funktion zur Folge haben. Dennoch erkennt man deutlich, wie die Annahme des Parameters b als zufälliger Effekt, die Schätzung wesentlich glatter werden lässt.

Die Verbindung von linearen gemischten Modellen und der Schätzung von P-Splines besteht nun in der Tatsache, dass die Behandlung des Parameters b als zufälliger Effekt äquivalent ist zur Penalisierung mit Differenzen erster Ordnung in einer Darstellung des obigen Modells über B-Spline-Basisfunktionen, wobei der Glättungsparameter α gegeben ist durch $\alpha = \frac{1}{\tau}$.

Die Äquivalenz der zwei Darstellungen des Modells über die Truncated Power Series-Basis vom Grad 0 und Behandlung der Parameter $b_1, \ldots, b_{m-1}$ als unabhängige zufällige Effekte beziehungsweise der Verwendung erster Differenzen als Penalisierung für B-Splines vom Grad 0 beruht auf der Tatsache, dass sich die B-Spline-Basisfunktionen in diesem Fall als Differenzen der Basisfunktionen der Truncated Power Series-Basis schreiben lassen. Man vergleiche hierzu die Definition der B-Spline-Basisfunktionen vom Grad 0 durch $B_j^0(x) = \mathbf{1}_{[\xi_j, \xi_{j+1})}$.

3.3.2 Allgemeiner Fall

Nun soll, basierend auf Green (1987), eine allgemeine Möglichkeit aufgezeigt werden, das durch den Prädiktor (3.7) beschriebene Modell, so zu reparametrisieren, dass eine Schätzung über Modelle mit zufälligen Effekten möglich ist. Dazu schreibt man die r_j-dimensionalen Parametervektoren ζ_j, $j = 1, \ldots, s, spat$ als Summe zweier Effekte

$$\zeta_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j,$$

wobei β_j den p_j-dimensionalen, nicht durch die Strafmatrix K_j penalisierten Anteil von ζ_j und b_j den $q_j = (r_j - p_j)$-dimensionalen, penalisierten Anteil von ζ_j enthalten soll. Zusätzlich soll die Reparametrisierung von ζ_j so gewählt werden, dass man die Elemente aus b_j als unabhängige und identisch verteilte zufällige Effekte interpretieren kann. Daher werden im Folgenden β_j und b_j auch als fixer beziehungsweise zufälliger Anteil von ζ_j bezeichnet. Die Dimensionen p_j und q_j werden durch den Rang der Strafmatrix K_j bestimmt. In der Regel besitzt diese Strafmatrix nicht vollen Rang, so dass ein nichttrivialer Nullraum $\text{Ke}(K_j) \subset \mathbb{R}^{r_j}$ existiert, für dessen Dimension $\dim(\text{Ke}(K_j)) = p_j$ gilt.

Die gewünschten Eigenschaften der Reparametrisierung lassen sich auch als Bedingungen an die $r_j \times p_j$ beziehungsweise $r_j \times q_j$-dimensionalen Matrizen $\tilde{X}_j$ und $\tilde{Z}_j$ formulieren. Man erhält so die Forderung, dass $\tilde{X}_j$ und $\tilde{Z}_j$ die folgenden Eigenschaften erfüllen sollen:

- $\tilde{X}_j' K_j \tilde{X}_j = 0$, so dass β_j nicht penalisiert wird, und
- $\tilde{Z}_j' K_j \tilde{Z}_j = I_{q_j}$, so dass die Elemente von b_j als unabhängig aufgefasst werden können.

Um entsprechende Matrizen $\tilde{X}_j$ und $\tilde{Z}_j$ zu erhalten, bestimmt man zunächst eine Basis von $\text{Ke}(K_j)$ und setzt $\tilde{X}$ dann aus den Basisvektoren von $\text{Ke}(K_j)$ zusammen. Zusätzlich zerlegt man K_j als $K_j = L_j L_j'$, wobei L_j eine $r_j \times q_j$-dimensionale Matrix mit vollem Spaltenrang sein soll, die $L_j' \tilde{X}_j = 0$ erfüllt. Dann setzt man $\tilde{Z}_j = L_j (L_j' L_j)^{-1}$.

Die Matrix L_j lässt sich dabei prinzipiell aus der Spektralzerlegung der Penalisierungsmatrix K_j bestimmen. Die Matrix K_j ist symmetrisch und positiv semidefinit und lässt sich damit zerlegen zu $K_j = \Gamma_j \Omega_j \Gamma_j'$ wobei Γ_j die orthogonale Matrix der Eigenvektoren von K_j ist, das heißt es gilt $\Gamma_j' \Gamma_j = I$, und $\Omega_j = \text{diag}(\omega_{j1}, \ldots, \omega_{jr_j})$ die Diagonalmatrix der absteigend geordneten Eigenwerte. Für die Eigenwerte gilt $\omega_{jl} \geq 0$, da K_j positiv semidefinit ist. Genauer gilt wegen $\text{rg}(K_j) = r_j - p_j = q_j$, dass p_j Eigenwerte gleich null sind, während q_j Eigenwerte positiv sind. L_j besteht nun aus den mit $\sqrt{\omega_{jl}}$ skalierten Eigenvektoren, deren Eigenwerte positiv sind, das heißt L_j kann gebildet werden aus $\widetilde{L}_j = \Gamma_j \Omega_j^{\frac{1}{2}}$ indem man aus $\widetilde{L}_j$ die letzten p_j Spalten streicht. Für die so definierte Matrix L_j gilt dann offensichtlich die gewünschte Eigenschaft $K_j = L_j L_j'$ und darüberhinaus $L_j' L_j = \widetilde{\Omega}_j$, wobei $\widetilde{\Omega}_j$ die Diagonalmatrix der positiven Eigenwerte bezeichnet.

Häufig lassen sich noch weitere Zerlegungen von K_j finden, die insbesondere aus numerischen Gesichtspunkten günstiger sind, da sie die aufwändige Bestimmung von Eigenwerten und Eigenvektoren vermeiden. Einige solche Möglichkeiten werden in den später behandelten Beispielen vorgestellt werden.

Aus der speziellen Wahl von $\tilde{X}_j$ und $\tilde{Z}_j$ folgen

- $L_j'\tilde{X}_j = \tilde{X}_j'L_j = 0$ (weil $\tilde{X}_j$ die Basisvektoren zu $\mathrm{Ke}(K_j)$ enthält) und damit wie gewünscht $\tilde{X}_j'K_j\tilde{X}_j = 0$,
- $\tilde{Z}_j'\tilde{X}_j = 0$, das heißt, die Matrizen $\tilde{Z}_j$ und $\tilde{X}_j$ sind orthogonal, und
- $\alpha_j\zeta_j'K_j\zeta_j = b_j'\Lambda_j^{-1}b_j$ mit der $q_j \times q_j$-Diagonalmatrix $\Lambda_j = \tau_j I_{q_j}$ und dem inversen Glättungsparameter $\tau_j = \frac{1}{\alpha_j}$.

Die dritte Eigenschaft erhält man leicht durch Einsetzen von $\zeta_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ und unter Verwendung der Definition von $\tilde{Z}_j$:

$$\begin{aligned}
\alpha_j\zeta_j'K_j\zeta_j &= \alpha_j(\tilde{X}_j\beta_j + \tilde{Z}_jb)_j'K_j(\tilde{X}_j\beta_j + \tilde{Z}_jb_j) \\
&= \alpha_jb_j'\tilde{Z}_j'K_j\tilde{Z}_jb_j \\
&= \alpha_jb_j'\underbrace{(L_j'L_j)^{-1}L_j'L_jL_j'L_j(L_j'L_j)^{-1}}_{=I}b_j \\
&= \alpha_jb_j'b_j \\
&= b_j'\Lambda_j^{-1}b_j.
\end{aligned}$$

Für die einzelnen Funktionen f_j, $j = 1, \ldots, s, spat$ erhält man so die Darstellung

$$f_j = B_j\tilde{X}_j\beta_j + B_j\tilde{Z}_jb_j = X_j\beta_j + Z_jb_j$$

mit $X_j = B_j\tilde{X}_j$ und $Z_j = B_j\tilde{Z}_j$ und für die Penalisierungen

$$-\frac{1}{2}\alpha_j\zeta_j'K_j\zeta_j = -\frac{1}{2}b_j'\Lambda_j^{-1}b_j.$$

Wie man sieht, besitzt die Penalisierung nun die gleiche Form, die man unter der Annahme

$$b_j \sim \mathrm{N}(0, \Lambda_j)$$

erhalten würde, das heißt, die Elemente von b_j können, wie gewünscht, als unabhängige und identisch verteilte zufällige Effekte betrachtet werden. Die Schätzung der Funktionen f_j wird damit zurückgeführt auf die Schätzung der Parameter β_j, b_j und τ_j, die sich alle mit Hilfe der in Kapitel 2 vorgestellten Verfahren

bestimmen lassen. Insbesondere lässt sich der inverse Glättungsparameter als Varianzparameter auffassen und über den Restricted-Maximum-Likelihood-Ansatz schätzen.

Um das gesamte Modell als generalisiertes lineares gemischtes Modell darstellen zu können, fasst man die Designmatrizen aller, nach der Reparametrisierung im Modell vorhandenen fixen Anteile in der Matrix

$$X = (X_{par}, X_1, \ldots, X_s, X_{spat})$$

und alle Designmatrizen der als zufällig betrachteten Anteile in der Matrix

$$Z = (Z_1, \ldots, Z_s, Z_{spat}, Z_{ran})$$

zusammen. Damit ergibt sich für den Prädiktor aus (3.7)

$$\eta = X\beta + Zb \tag{3.8}$$

mit den Parametervektoren $\beta = (\beta'_{par}, \beta'_1, \ldots, \beta'_s, \beta'_{spat})'$ und $b = (b'_1, \ldots, b'_s, b'_{spat}, b'_{ran})'$. Für die einzelnen zufälligen Anteile gelten die Annahmen

$$b_j \sim \mathrm{N}(0, \Lambda_j),\ j = 1, \ldots, s \qquad b_{spat} \sim \mathrm{N}(0, \Lambda_{spat})$$

mit den Kovarianzmatrizen

$$\Lambda_j = \tau_j I_{q_j},\ j = 1, \ldots, s \qquad \Lambda_{spat} = \tau_{spat} I_{q_{spat}}$$

und den inversen Glättungsparametern $\tau_j = \frac{1}{\alpha_j}$, $j = 1, \ldots, s$ und $\tau_{spat} = \frac{1}{\alpha_{spat}}$. Fasst man diese Kovarianzmatrizen noch zusammen in der Matrix

$$Q(\nu) = \mathrm{blockdiag}(\Lambda_1, \ldots, \Lambda_s, \Lambda_{spat}, Q_{ran}(\nu_{ran})),$$

mit dem Vektor der Varianzparameter $\nu = (\tau_1, \ldots, \tau_s, \tau_{spat}, \nu'_{ran})'$, so erhält man insgesamt ein generalisiertes lineares Modell mit linearem Prädiktor (3.8) und der Annahme

$$b \sim \mathrm{N}(0, Q(\nu)).$$

Man hat also durch die Reparametrisierung das gesamte generalisierte geoadditive gemischte Modell auf ein generalisiertes lineares gemischtes Modell zurückgeführt und kann die Schätzung der einzelnen Modellkomponenten, wie in Kapitel 2 beschrieben, durchführen.

Für eine Reihe von Beispielen soll nun die Reparametrisierung detaillierter dargestellt werden.

P-Splines

Die Penalisierung von P-Splines beruht auf Differenzen k_j-ter Ordnung und ergibt die Strafmatrix $K_j = D_j{}'D_j$, wobei mit D_j die entsprechende Differenzenmatrix bezeichnet wird. Durch k_j-te Differenzen nicht penalisiert wird ein Polynom vom Grad $k_j - 1$ in den äquidistanten Knoten $\xi_{j1}, \ldots, \xi_{jr_j}$, das heißt, die Matrix K_j besitzt einen $p_j = (k_j - 1)$-dimensionalen Kern. Eine Basis dieses Kerns erhält man durch die Spalten der Matrix

$$\tilde{X}_j = \begin{pmatrix} 1 & \xi_{j1} & \ldots & \xi_{j1}^{k_j-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \xi_{jr_j} & \ldots & \xi_{jr_j}^{k_j-1} \end{pmatrix}.$$

Außerdem gilt mit dieser Definition offensichtlich $D_j\tilde{X}_j = 0$ und man erhält als alternativen Zerlegungsfaktor L_j die Matrix D_j'. Damit ergibt sich $\tilde{Z}_j$ als $\tilde{Z}_j = D_j'(D_jD_j')^{-1}$.

Aus der Zerlegung der Einheit durch P-Splines folgt nun $B_j\mathbf{1}_{p_j} = \mathbf{1}_n$, wobei $\mathbf{1}_p$ den p-dimensionalen Einsvektor bezeichnet. Die erste Spalte des Produkts $B_j\tilde{X}_j$ besteht also nur aus Einsen. Darum besitzt die Designmatrix X nicht mehr vollen Spaltenrang, da jedes der Produkte $B_j\tilde{X}_j$, $j = 1, \ldots, s$ eine nur aus Einsen bestehende Spalte enthält.

Wie in den folgenden Beispielen gezeigt wird, erhält man diese Aussage nicht nur für P-Splines, sondern auch für Markov-Zufallsfelder und für Glättungssplines. Man beachte, dass sich hierin das in Kapitel 3.1 beschriebene Identifizierbarkeitsproblem wiederspiegelt. Für jede der reparametrisierten Funktionen f_1 bis f_s und f_{spat} erhält man einen Intercept, über den das Niveau der Funktion festgelegt wird. Um die Schätzbarkeit des Modells zu gewährleisten genügt eine leichte Modifikation der Definition der Matrizen $\tilde{X}_j$, wie sie in Algorithmus 5 in Kapitel 3.4 zugrunde gelegt werden wird. Dazu streicht man aus diesen Matrizen die enthaltene Einsspalte. Man beachte, dass dann wieder ein Intercept in der Designmatrix X_{par} enthalten sein muss.

Für die modifizierte Version von $\tilde{X}_j$ erhält man bei P-Splines beispielsweise

$$\tilde{X}_j = \begin{pmatrix} \xi_{j1} & \ldots & \xi_{j1}^{k_j-1} \\ \vdots & \ddots & \vdots \\ \xi_{jr_j} & \ldots & \xi_{jr_j}^{k_j-1} \end{pmatrix}.$$

Markov-Zufallsfeld

Bei Modellierung des räumlichen Effekts durch ein Markov-Zufallsfeld werden durch die Strafmatrix K_{spat} Abweichungen von einem globalen Mittelwert penalisiert. Man erhält also als Basis des eindimensionalen Nullraums $\mathrm{Ke}(K_{spat})$ den r_{spat}-dimensionalen Vektor $(1,\ldots,1)'$ und damit

$$\tilde{X}_{spat} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Da die Matrix B_{spat} aufgrund ihrer Definition als Pseudodesignmatrix in jeder Spalte aus genau einer Eins und $r_{spat} - 1$ Nullen besteht, folgt sofort $X_{spat} = B_{spat}\tilde{X}_{spat} = \mathbf{1}_n$. Um die Identifizierbarkeit des Gesamtmodells zu gewährleisten, wird also die vollständige Matrix X_{spat} aus X gestrichen.

Die Matrix $\tilde{Z}_{spat}$ muss wie oben beschrieben über die Spektralzerlegung der Strafmatrix K_{spat} bestimmt werden.

Glättungssplines

Die Penalisierung beruht bei Glättungssplines auf der zweiten Ableitung der Funktion f_j. Dementsprechend penalisiert die Strafmatrix K_j in diesem Fall nur Abweichungen von einer Geraden. Als Basis des zweidimensionalen Nullraums $\mathrm{Ke}(K_j)$ erhält man damit die Vektoren der Matrix

$$\tilde{X}_j = \begin{pmatrix} 1 & x_{j(1)} \\ \vdots & \vdots \\ 1 & x_{j(r_j)} \end{pmatrix},$$

wobei $x_{j(1)},\ldots,x_{j(r_j)}$ wieder die r_j verschiedenen, geordneten Ausprägungen der j-ten Kovariable bezeichnen.

Wie bei Markov-Zufallsfeldern ergibt sich aus der Definition von B_j als Pseudodesignmatrix, dass die erste Spalte von $X_j = B_j\tilde{X}_j$ nur aus Einsen besteht. Die Modifikation zur Sicherung der Identifizierbarkeit führt zur Definition von $\tilde{X}_j$ als

$$\tilde{X}_j = \begin{pmatrix} x_{j(1)} \\ \vdots \\ x_{j(r_j)} \end{pmatrix}.$$

Die Matrix $\tilde{Z}_j$ lässt sich aus der Zerlegung der Matrix K_j zu $K_j = E_j' C_j^{-1} E_j$ mit den auf Seite 54 definierten Matrizen E_j und C_j bestimmen. Man erhält $L_j = E_j' C_j^{-\frac{1}{2}}$ mit der symmetrischen Wurzel $C_j^{-\frac{1}{2}}$. Die notwendige Eigenschaft $L_j' \tilde{X}_j = 0$ folgt aus der Definition von E_j über die Differenzen $h_l = x_{j(l+1)} - x_{j(l)}$. Für die zwei Elemente der l-ten Zeile des Produkts $E_j \tilde{X}_j$ gilt nämlich

$$\frac{1}{h_l} - \left(\frac{1}{h_l} - \frac{1}{h_{l+1}} \right) + \frac{1}{h_{l+1}} = 0$$

und

$$\begin{aligned}
&\frac{x_{j(l)}}{h_l} - \left(\frac{x_{j(l+1)}}{h_l} + \frac{x_{j(l+1)}}{h_{l+1}} \right) + \frac{x_{j(l+2)}}{h_{l+1}} \\
&= \frac{h_{l+1}(x_{j(l)} - x_{j(l+1)}) + h_l(x_{j(l+2)} - x_{j(l+1)})}{h_l h_{l+1}} \\
&= \frac{h_l h_{l+1} - h_l h_{l+1}}{h_l h_{l+1}} \\
&= 0.
\end{aligned}$$

3.4 Schätzung über generalisierte lineare gemischte Modelle

Mit Hilfe der im vorigen Abschnitt vorgestellten Reparametrisierung ist die Schätzung generalisierter geoadditiver gemischter Modelle verhältnismäßig einfach möglich. Zunächst werden die Matrizen X und Z aus den ursprünglichen Daten erzeugt, dann lässt sich Algorithmus 2 aus Kapitel 2 anwenden. Man erhält damit Schätzer $\hat{\beta}$ und $\hat{b}$ sowie $\hat{\vartheta}$, wobei wieder der Vektor aller Varianzparameter gegeben ist durch $\vartheta = (\phi, \nu')'$. Aus diesen Schätzern lassen sich dann alle interessierenden Größen ableiten.

Das genaue Vorgehen zur Schätzung mit Hilfe von P-Splines und Markov-Zufallsfeldern soll im nächsten Algorithmus zusammengefasst werden:

Algorithmus 5 (Schätzung im generalisierten geoadditiven gemischten Modell)

(i) Bilde die Matrizen

$$\begin{aligned}
\tilde{X}_j &= \begin{pmatrix} \xi_{j1} & \dots & \xi_{j1}^{k_j-1} \\ \vdots & \ddots & \vdots \\ \xi_{jr_j} & \dots & \xi_{jr_j}^{k_j-1} \end{pmatrix}, \\
\tilde{Z}_j &= D_j'(D_j D_j')^{-1}, \qquad j = 1, \dots, s, \\
\tilde{Z}_{spat} &= L_{spat}(L_{spat}' L_{spat})^{-1},
\end{aligned}$$

wobei $\xi_{j1}, \dots, \xi_{jr_j}$ die r_j Knoten des j-ten P-Splines mit Penalisierungsmatrix D_j der Ordnung k_j und Dimension $(r_j - k_j) \times r_j$ bezeichne und L_{spat} wie beschrieben aus der Spektralzerlegung der Strafmatrix K_{spat} stammt.

(ii) Bilde die Designmatrizen

$$\begin{aligned}
X &= (X_{par}, B_1\tilde{X}_1, \dots, B_s\tilde{X}_s) \\
Z &= (B_1\tilde{Z}_1, \dots, B_s\tilde{Z}_s, B_{spat}Z_{spat}, Z_{ran}),
\end{aligned}$$

und verwende Algorithmus 2 zur Schätzung von

$$\begin{aligned}
\beta &= (\beta_{par}', \beta_1', \dots, \beta_s')', \\
b &= (b_1', \dots, b_s', b_{spat}', b_{ran}')' \qquad \text{und} \\
\vartheta &= (\phi, \tau_1, \dots, \tau_s, \tau_{spat}, \nu_{ran}')'.
\end{aligned}$$

(iii) Bestimme aus den Schätzern $\hat{\beta}$ und $\hat{b}$ die Funktionsschätzungen

$$\begin{aligned}
\hat{f}_j &= B_j(\tilde{X}_j\hat{\beta}_j + Z_j\hat{b}_j) \qquad j = 1, \dots, s \\
\hat{\zeta}_{spat} &= \tilde{Z}_{spat}\hat{b}_{spat}.
\end{aligned}$$

Die Schätzer $\hat{\beta}_{par}$ und $\hat{b}_{ran}$ sind direkt aus $\hat{\beta}$ und $\hat{b}$ erhältlich.

Der beschriebene Algorithmus führt nicht automatisch zu zentrierten Funktionsschätzungen, da die Restriktionen an das Niveau der einzelnen Funktionen durch das Streichen der Einsspalten aus den jeweiligen Designmatrizen $\tilde{X}_j$ zustande kamen. Häufig ist man jedoch an zentrierten Schätzungen interessiert, weil sich dann der Verlauf der einzelnen Funktionen als Abweichung vom mittleren Niveau interpretieren lässt. Daher kann es sinnvoll sein, die aus Algorithmus 5

erhaltenen Schätzungen nachträglich geeignet zu zentrieren. Man beachte, dass dabei auch $\hat{\beta}_0$ verändert werden muss, damit die Summe der Effekte erhalten bleibt.

Die Reparametrisierung des generalisierten geoadditiven Modells hat nicht nur eine einfache Möglichkeit zur Bestimmung der Glättungsparameter und damit zur Schätzung des gesamten Modells zur Folge, sondern es resultieren auch eine Reihe interessanter Eigenschaften.

Zunächst besitzen die als P-Splines modellierten Funktionen, Markov-Zufallsfelder und auch Glättungssplines über die Darstellung als Summe eines fixen und eines zufälligen Teils eine bayesianische Interpretation. Nimmt man für den fixen Parameter β_j eine flache Priori-Verteilung und für den zufälligen Parameter b_j die $\mathrm{N}(0, \Lambda_j)$-Verteilung als Priori an, so lässt sich das Produkt beider Priori-Verteilungen als Priori des gesamten Parametervektors und damit für f_j auffassen. Die entsprechenden Schätzer lassen sich dann wie in Kapitel 2 als Posteriori-Modus-Schätzer beziehungsweise als empirische Bayes-Schätzer interpretieren.

Bereits bei der Einführung des Penalisierungskonzepts für P-Splines fiel die Ähnlichkeit zwischen der penalisierten Likelihood für ζ_j und der Log-Posteriori (2.12) in einem Modell mit zufälligen Effekten auf. Prinzipiell ließe sich für die verschiedenen Modellkomponenten ζ_j, $j = 1, \ldots, s, spat$ auch der Penalisierungsterm $-\frac{1}{2}\alpha_j \zeta_j' K_j \zeta_j$ bereits als logarithmierte Priori-Verteilung interpretieren, wenn man uneigentliche Verteilungen als Prioris zulässt. Diese Interpretation wurde für Markov-Zufallsfelder bereits bei der Herleitung in Kapitel 3.1.2 verwendet, sie lässt sich aber auch auf P-Splines und Glättungssplines übertragen. Die Betrachtung nach der Reparametrisierung besitzt nun den Vorteil, eine explizitere Darstellung der Priori-Verteilung durch einen uneigentlichen Anteil für den Parameter β_j und einen eigentlichen Anteil für den Parameter b_j zu ermöglichen.

Im linearen gemischten Modell, das heißt für normalverteilten Response, sind, wie in Kapitel 2.1.2 gezeigt, die Schätzer für β und b bei gegebenen Varianzparametern bester linearer unverzerrter Schätzer beziehungsweise Prädiktor. Über die Reparametrisierung lassen sich auch für P-Splines, Glättungssplines und Markov-Zufallsfelder die Schätzer für ζ_j beziehungsweise für f_j bei gegebenen Hyperparametern als beste lineare unverzerrte Prädiktoren betrachten. Für Glättungssplines wurde diese Eigenschaft zuerst von Speed (1991) festgestellt. Die Optima-

litätseigenschaften bester linearer unverzerrter Schätzer beziehungsweise Prädiktor beinhalten insbesondere die Unverzerrtheit der Schätzer im in Kapitel 2.1.2 angegebenen Sinn. Damit lässt sich auch die Schätzung für ζ_j beziehungsweise für f_j in gewisser Weise als unverzerrt auffassen. Man beachte jedoch, dass dies nicht im herkömmlichen Sinn gilt, das heißt, es gilt nicht $\mathbb{E}_{\zeta_j}(\hat{\zeta}_j) = \zeta_j$, sondern $\mathbb{E}(\hat{\zeta}_j) = \mathbb{E}(\zeta_j) = \mathbb{E}(\tilde{X}_j\beta_j + \tilde{Z}_j b_j) = \tilde{X}_j\beta_j$.

Ein weiterer Vorteil der Darstellung des generalisierten geoadditiven gemischten Modells als Modell mit zufälligen Effekten besteht in der Möglichkeit, Quasi-Likelihood-Modelle (vergleiche Kapitel 2.5) zu schätzen. Die Darstellung als generalisiertes lineares gemischtes Modell erlaubt es also, auch in generalisierten geoadditiven gemischten Modellen dem Problem der Überdispersion geeignet zu begegnen. Die Annahme eines Quasi-Likelihood-Modells ändert dabei nichts an der Modellierung und der Interpretation der nonparametrischen Effekte oder des räumlichen Effekts, sondern resultiert lediglich in einer leichten Modifikation der Arbeitsgewichte. Zusätzlich ist jedoch wieder zu beachten, dass es sich bei den zur Schätzung verwendeten Größen Score-Funktion und Fisher-Information nun nicht mehr um die Ableitung einer Log-Likelihood beziehungsweise den negativen Erwartungswert der zweiten Ableitung einer Log-Likelihood handelt. Für den Response gilt also insbesondere nicht mehr die Annahme, dass seine Verteilung aus einer Exponentialfamilie stammt.

Im Rahmen der Simulationsstudien in Kapitel 5 wird auch untersucht werden, wie gut eine in Wahrheit lineare Funktion bei einer Modellierung über P-Splines tatsächlich als Gerade erkannt wird. Da in der Simulation stets P-Splines mit Differenzen der Ordnung $k = 2$ verwendet werden, entspräche dies der Schätzung des zugehörigen inversen Glättungsparameters als $\hat{\tau} = 0$. Obwohl die auf Seite 22 besprochene Parametrisierung der Varianzparameter ohne Restriktionen auch eine Schätzung sehr kleiner Varianzparameter erlauben sollte, ergaben sich in einer großen Zahl von Modellen Konvergenzprobleme. In Abbildung 3.8 ist die Restricted-Log-Likelihood $l^*(\tau)$ für zwei Realisationen eines Modells mit einem auf einer linearen Funktion basierenden Prädiktor visualisiert. Während bei der in Abbildung 3.8 (a) wiedergegebenen Realisation keine Konvergenzprobleme zu beobachten waren, ergaben sich diese für die Realisation in Abbildung 3.8 (b). Der Unterschied der beiden Restricted-Log-Likelihoods ist offensichtlich: Während die Restricted-Log-Likelihood in Abbildung 3.8 (a) ein globales Maximum im Inneren

des Parameterraums besitzt, liegt für Restricted-Log-Likelihood in Abbildung 3.8 (b) das Maximum auf dem Rand des Parameterraums. In diesem Fall konvergiert der Schätzer $\hat{\tau}$ zunächst bis nahe an den Rand des Parameterraums und springt dann wieder zurück zu einem weiter vom Rand entfernten Wert.

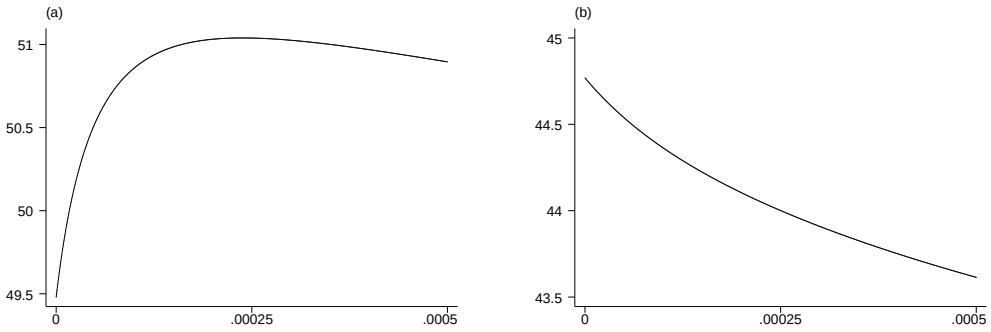

Abbildung 3.8: Restricted-Log-Likelihood $l^*(\tau)$ für zwei Modelle, deren Prädiktor auf einer linearen Funktion beruht. In Abbildung (a) ist eine Schätzung möglich, in Abbildung (b) kommt es zu Konvergenzproblemen.

Um diesem Problem in einer datengesteuerten Weise zu begegnen, wurden Varianzparameter, deren zugehörige zufällige Effekte im Verhältnis zum Gesamteffekt sehr klein sind, von der weiteren Schätzung ausgeschlossen. Konkret wird in jeder Iteration für sämtliche Varianzparameter mit Ausnahme von ϕ das Kriterium

$$c(\vartheta_j) = \frac{||b_j||}{||\eta||} \tag{3.9}$$

bestimmt. Unter b_j ist dabei der zu ϑ_j korrespondierende zufällige Effekt zu verstehen. Dabei kann es sich sowohl um einen durch die Reparametrisierung eines P-Splines oder eines Markov-Zufallsfeldes entstehenden zufälligen Effekt, als auch um einen ursprünglich im Modell enthaltenen zufälligen Effekt handeln. In b_j werden dann jeweils alle zufälligen Effekte zusammengefasst, die den gleichen Varianzparameter besitzen.

Unterschreitet das Kriterium einen kleinen Wert (zum Beispiel 0.001), so wird der zugehörige Varianzparameter beim aktuellen Schätzwert fixiert. Ein kleiner Wert des obigen Kriteriums korrespondiert dabei mit einer kleinen Schätzung des zugehörigen Varianzparameters. Da man den Betrag des Varianzparameters aber relativ zu den gegebenen Daten betrachten muss, wurde das obige, datenabhängige Kriterium gewählt. Durch diese Modifikation von Algorithmus 5 konnten die

auf dem beschriebenen Problem beruhenden Konvergenzprobleme in allen Fällen behoben werden. Dennoch blieben bei einigen simulierten Datensätzen in Kapitel 5.2 Konvergenzprobleme bestehen, deren konkrete Ursache jedoch nicht bestimmt werden konnte.

3.5 Konfidenzbänder

In der Regel interessiert man sich nicht nur für Punktschätzer der Modellparameter, sondern auch für einen Sicherheitsbereich, der die Unsicherheit der Schätzung berücksichtigt. Für parametrisch modellierte Effekte interessiert man sich also für Konfidenzintervalle. Ein Analogon für nonparametrisch modellierte Funktionen beziehungsweise räumliche Funktionen sind punktweise Konfidenzbänder, die für jeden beobachteten Designpunkt ein Konfidenzintervall zu einem bestimmten Sicherheitsgrad darstellen. Dabei ist zu beachten, dass die globale Überdeckungswahrscheinlichkeit jedoch sehr viel niedriger sein kann, als die Überdeckungswahrscheinlichkeit, die bei Betrachtung eines einzelnen Designpunkts zugrunde gelegt wurde. Mit Hilfe der Bonferroni-Ungleichung lässt sich eine Abschätzung auch für die globale Überdeckungswahrscheinlichkeit bestimmen.

Im generalisierten linearen Modell erhält man unter Regularitätsbedingungen, dass der Maximum-Likelihood-Schätzer $\hat{\beta}$ asymptotisch normalverteilt ist mit der inversen Fisher-Information als Kovarianzmatrix. Basierend auf dieser asymptotischen Verteilung lassen sich dann Konfidenzintervalle definieren. Es erscheint plausibel, dieses Ergebnis unter geeigneten Modifikationen auch in dem durch die Reparametrisierung entstandenen generalisierten linearen gemischten Modell zu verwenden.

Die Fisher-Scoring-Gleichungen zur iterativen Bestimmung von β und b in generalisierten linearen gemischten Modellen sind gegeben durch (vergleiche Kapitel 2.3.3):

$$\begin{pmatrix} X'W(\eta)X & X'W(\eta)Z \\ Z'W(\eta)X & Z'W(\eta)Z + Q(\nu)^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'W(\eta)\tilde{y}(\eta) \\ Z'W(\eta)\tilde{y}(\eta) \end{pmatrix}.$$

Bedingt auf die zufälligen Effekte erhält man für die Kovarianzmatrix von $\tilde{y}(\eta)$:

$$\mathrm{Var}(\tilde{y}(\eta)|b)) = W(\eta)^{-1}.$$

Definiert man zusätzlich die Matrizen

$$H = \begin{pmatrix} X'W(\eta)X & X'W(\eta)Z \\ Z'W(\eta)X & Z'W(\eta)Z + Q(\nu)^{-1} \end{pmatrix}$$

und

$$H_1 = \begin{pmatrix} X'W(\eta)X & X'W(\eta)Z \\ Z'W(\eta)X & Z'W(\eta)Z \end{pmatrix},$$

so ergibt sich nach Konvergenz aus den Fisher-Scoring-Gleichungen für die Kovarianzmatrix von $\hat{\beta}$ und $\hat{b}$:

$$\operatorname{Var}\begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = H^{-1}H_1H^{-1}.$$

Dabei wurden sowohl β als auch b vorübergehend als feste Parameter betrachtet. Dies lässt sich beispielsweise aus der in Kapitel 3.3 beschriebenen Reparametrisierung erklären: Üblicherweise werden die Schätzer in der nonparametrischen Regression frequentistisch, das heißt als feste und unbekannte Parameter betrachtet. Daher mag es sinnvoll erscheinen, dies auch nach der Reparametrisierung beizubehalten, obwohl die entsprechenden ‚zufälligen' Anteile b_j die Form zufälliger Effekte besitzen. Auch für die ursprünglichen zufälligen Effekte ist es möglich, diesen Standpunkt einzunehmen, wenn sie nämlich als hoch unstrukturierte Funktion etwa zur Modellierung lokaler Variationen in räumlichen Ansätzen verwendet werden.

Dennoch mag eine Betrachtung von b als zufällig natürlicher erscheinen. Man erhält in diesem Fall (vergleiche Lin & Zhang (1999)) für die Kovarianzmatrix von $\hat{\beta}$ und $\hat{b}$:

$$\operatorname{Var}\begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = H^{-1}.$$

Die beiden verschiedenen Formen sollen im Folgenden als frequentistische beziehungsweise bayesianische Versionen bezeichnet werden, da ihnen die entsprechenden Betrachtungen des Parameters b als fest beziehungsweise als zufällig zugrunde liegen.

Für die Schätzer $\hat{\beta}_{par}$ und $\hat{b}_{ran}$ sind die Kovarianzmatrizen nun unmittelbar aus den entsprechenden Blöcken von $H^{-1}H_1H^{-1}$ in der frequentistischen Version beziehungsweise H^{-1} in der bayesianischen Version ablesbar. Für die Parameter $\hat{\zeta}_j$, $j = 1, \ldots, s$ erhält man

$$\operatorname{Var}(\hat{\zeta}_j) = (\tilde{X}_j, \tilde{Z}_j) \operatorname{Var}\begin{pmatrix} \hat{\beta}_j \\ \hat{b}_j \end{pmatrix} (\tilde{X}_j, \tilde{Z}_j)'$$

und für $\hat{f}_j$

$$\operatorname{Var}(\hat{f}_j) = B_j \operatorname{Var}(\hat{\zeta}_j) B_j' = (X_j, Z_j) \operatorname{Var} \begin{pmatrix} \hat{\beta}_j \\ \hat{b}_j \end{pmatrix} (X_j, Z_j)',$$

wobei sich $\operatorname{Var} \begin{pmatrix} \hat{\beta}_j \\ \hat{b}_j \end{pmatrix}$ wieder aus den entsprechenden Blöcken von $H^{-1} H_1 H^{-1}$ beziehungsweise H^{-1} ablesen lässt. Für $\hat{\zeta}_{spat}$ ergibt sich

$$\operatorname{Var}(\hat{\zeta}_{spat}) = \tilde{Z}_{spat} \operatorname{Var}(\hat{b}_{spat}) \tilde{Z}_{spat}'.$$

Zur Konstruktion von Konfidenzintervallen beziehungsweise Konfidenzbändern nimmt man nun im Allgemeinen zusätzlich an, dass die asymptotische Verteilung von $(\hat{\beta}', \hat{b}')'$ eine Normalverteilung ist. Obwohl keine expliziten Beweise für diese Annahme existieren, erscheint sie in Analogie zum generalisierten linearen Modell zumindest plausibel. Basierend auf dieser Annahme lassen sich dann mit Hilfe der Quantile der Normalverteilung und den jeweiligen Standardabweichungen, die als Wurzeln der Diagonalelemente der Kovarianzmatrizen erhältlich sind, Konfidenzintervalle definieren. Man beachte, dass die bayesianischen Versionen der Konfidenzintervalle stets weiter sind als die entsprechenden frequentistischen Konfidenzintervalle.

4 LQ-Tests im linearen gemischten Modell

Wie in Kapitel 3.3 gezeigt wurde, lassen sich generalisierte geoadditive gemischte Modelle durch eine geeignete Reparametrisierung in generalisierte lineare gemischte Modelle überführen. Der inverse Glättungsparameter $\tau = \frac{1}{\alpha}$ eines Effekts wird dabei zum Varianzparameter der Verteilung der entsprechenden zufälligen Effekte. Von besonderem Interesse ist häufig der Fall $\tau = 0$ beziehungsweise $\alpha = \infty$. Bei P-Splines erhält man so beispielsweise eine Modellierung als Polynom vom Grad $k - 1$, für Markov-Zufallsfelder entspricht $\tau = 0$ dem Fall, dass kein räumlicher Effekt vorhanden ist.

Insbesondere die Frage, ob die Abhängigkeit zwischen einer Kovariablen und der abhängigen Variablen angemessen durch eine Gerade wiedergegeben werden kann, oder ob eine nonparametrische Modellierung adäquater wäre, ist häufig von Interesse. Eine Reihe von Möglichkeiten, die Modellanpassung eines linearen Modells mit der Modellanpassung einer nonparametrischen Modellierung zu vergleichen, liefern Tests, die in Analogie zu den F-Tests des linearen Modells konstruiert wurden. Man vergleiche etwa Azzalini & Bowman (1993), Cantoni & Hastie (2002) oder auch Hastie & Tibshirani (1990) Kapitel 3.9 zur Beschreibung solcher Tests. Hier soll jedoch ein anderer Ansatz behandelt werden, in dem der Varianzparameter τ über einen Likelihood-Quotienten-Test formal auf den Wert 0 getestet werden kann. Dieser Test kann sowohl basierend auf der Log-Likelihood als auch basierend auf der Restricted-Log-Likelihood durchgeführt werden. Die zugrunde liegende Theorie wird in Crainiceanu, Ruppert & Vogelsang (2002) und Crainiceanu & Ruppert (2002) dargestellt. Für P-Splines vom Grad 0, das heißt für Random-Walk-Modelle, findet man entsprechende Aussagen in Kuo (1999).

Leider lassen sich für die interessierenden Tests bisher keine allgemeinen Verteilungsaussagen machen, wie sie etwa für Likelihood-Quotienten-Tests der Regressionsparameter im generalisierten linearen Modell möglich sind. Vielmehr lassen sich nur Aussagen im Normalverteilungsfall und für Modelle mit nur einer Varianzkomponente herleiten. Da die asymptotischen Verteilungsaussagen stark vom speziellen Untersuchungsdesign abhängen, erscheint es auch schwierig, allgemeinere Aussagen zu erhalten.

Zunächst soll nun die Theorie eines Likelihood-Quotienten-Tests im linearen gemischten Modell vorgestellt werden, mit dem eine Varianzkomponente auf den

Wert 0 getestet werden kann. Besondere Beachtung findet dabei die Frage, mit welcher Wahrscheinlichkeit man Maxima der Likelihood-Quotienten im Punkt Null erhält. Diese Wahrscheinlichkeiten entsprechen nämlich den Punktmassen der Verteilungen der entsprechenden Likelihood-Quotienten-Teststatistiken im Punkt Null. Hierzu werden in Abschnitt 4.1.1 eine Reihe von Ergebnissen hergeleitet, die bereits für endlichen Stichprobenumfang gültig sind. Verteilungsaussagen sind dagegen in der Regel nur asymptotisch erhältlich und werden in Abschnitt 4.1.2 betrachtet. Anschließend werden drei Spezialfälle des linearen gemischten Modells behandelt und Verteilungsaussagen für die Teststatistiken in diesen Beispielen hergeleitet. Im Einzelnen sind dies ein einfaches, balanciertes ANOVA-Modell mit als zufällig angenommenen Gruppeneffekten sowie P-Splines und Markov-Zufallsfelder in ihren Repräsentationen als lineare gemischte Modelle.

4.1 Verteilung der Likelihood-Quotienten-Teststatistiken

Im Folgenden soll von einem linearen gemischten Modell der Form

$$y = X\beta + Zb + \varepsilon$$

ausgegangen werden, wie es in Kapitel 2.1 beschrieben wurde. Für den Parameter der zufälligen Effekte soll die Annahme

$$b \sim N(0, \tau I_q)$$

gelten, das heißt, in der Notation von Kapitel 2.1 gilt $Q(\nu) = \tau I_q$ und $\nu = \tau$. Im Unterschied zu Kapitel 2.1 wird also stets von unabhängigen und identisch verteilten zufälligen Effekten ausgegangen. Für den Fehler ε soll wieder die Annahme

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

erfüllt sein.

Definiert man γ als das Signal-Rauschen-Verhältnis $\gamma = \frac{\tau}{\sigma^2}$, so ergibt sich mit $V_\gamma = I_n + \gamma ZZ'$

$$\mathrm{Var}(y) = \sigma^2 V_\gamma.$$

Man beachte dabei die von Kapitel 2 abweichende Definition der Matrix V.

Betrachtet werden soll nun das Testproblem

$$H_0 : \tau = 0 \text{ versus } H_1 : \tau > 0$$

beziehungsweise äquivalent dazu

$$H_0 : \gamma = 0 \text{ versus } H_1 : \gamma > 0. \tag{4.1}$$

In einer exakteren Formulierung lässt sich dieses Testproblem auch schreiben als

$$H_0 : \gamma \in \Omega_0 \text{ versus } H_1 : \gamma \in \Omega_1$$

mit den Mengen $\Omega_0 = \{0\}$ und $\Omega_1 = (0, \infty)$. Zusätzlich bezeichne $\Omega = \Omega_0 \cup \Omega_1 = [0, \infty)$ den Parameterraum von γ.

Zur Durchführung des Tests (4.1) werden die Likelihood-Quotienten-Teststatistiken verwendet, die basierend auf der Log-Likelihood

$$l(\beta, \gamma, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log(|V_\gamma|) - \frac{1}{2\sigma^2}(y - X\beta)'V_\gamma^{-1}(y - X\beta)$$

beziehungsweise der Restricted-Log-Likelihood

$$\begin{aligned} l^*(\beta, \gamma, \sigma^2) &= -\frac{n-p-1}{2}\log(\sigma^2) - \frac{1}{2}\log(|V_\gamma|) - \frac{1}{2}\log(|X'V_\gamma^{-1}X|) \\ &\quad -\frac{1}{2\sigma^2}(y - X\beta)'V_\gamma^{-1}(y - X\beta) \end{aligned}$$

definiert werden können. In Abhängigkeit vom Signal-Rauschen-Verhältnis γ lassen sich Maximum-Likelihood- und Restricted-Maximum-Likelihood-Schätzer für β und σ^2 bestimmen:

$$\begin{aligned} \hat{\beta}(\gamma) &= (X'V_\gamma^{-1}X)^{-1}X'V_\gamma^{-1}y, \\ \hat{\sigma}^2_{ML}(\gamma) &= \frac{1}{n}(y - X\hat{\beta}(\gamma))'V_\gamma^{-1}(y - X\hat{\beta}(\gamma)), \\ \hat{\sigma}^2_{REML}(\gamma) &= \frac{1}{n-p-1}(y - X\hat{\beta}(\gamma))'V_\gamma^{-1}(y - X\hat{\beta}(\gamma)). \end{aligned}$$

Um die Verteilungen der Likelihood-Quotienten-Teststatistiken von β und σ^2 unabhängig zu machen, setzt man diese Schätzer in die Log-Likelihood beziehungsweise die Restricted-Log-Likelihood ein und erhält so die Profile-Log-Likelihood

$$l(\gamma) = -\frac{1}{2}\log(|V_\gamma|) - \frac{n}{2}\log(y'R_\gamma V_\gamma^{-1}R_\gamma y)$$

und die Profile-Restricted-Log-Likelihood

$$l^*(\gamma) = -\frac{1}{2}\log(|V_\gamma|) - \frac{1}{2}\log(|X'V_\gamma^{-1}X|) - \frac{n-p-1}{2}\log(y'R_\gamma V_\gamma^{-1}R_\gamma y)$$

mit der Residualmatrix

$$R_\gamma = I - X(X'V_\gamma^{-1}X)^{-1}X'V_\gamma^{-1}. \tag{4.2}$$

Der Likelihood-Quotient ist dann definiert als

$$LQ_n(\gamma) = 2l(\gamma) - 2l(0)$$

beziehungsweise basierend auf der Profile-Restricted-Log-Likelihood als

$$RLQ_n(\gamma) = 2l^*(\gamma) - 2l^*(0).$$

Dabei wird die Abhängigkeit vom Stichprobenumfang n explizit berücksichtigt, um im Folgenden besser zwischen Aussagen für endlichen Stichprobenumfang und asymptotischen Aussagen unterscheiden zu können. Bei $LQ_n(\gamma)$ und $RLQ_n(\gamma)$ handelt es sich um Zufallsgrößen im Funktionenraum $\mathcal{C}[0,\infty)$, dem Raum der stetigen Funktionen von $[0,\infty)$ nach $[0,\infty)$.

Die Likelihood-Quotienten-Teststatistiken erhält man nun als

$$LQ_n = \sup_{\gamma\geq 0} LQ_n(\gamma) = LQ_n(\hat{\gamma}_{ML})$$

und

$$RLQ_n = \sup_{\gamma\geq 0} RLQ_n(\gamma) = RLQ_n(\hat{\gamma}_{REML}).$$

Man beachte, dass der ML-Schätzer für γ nach dem Invarianzprinzip der Maximum-Likelihood-Schätzung (vergleiche Pruscha (2000) Seite 25) als Quotient der ML-Schätzer für τ und σ^2 erhältlich ist. Die gleiche Aussage gilt für den REML-Schätzer, weil dieser auf der Likelihood der Fehlerkontraste beruht. Bei LQ_n und RLQ_n handelt es sich nun wieder um reellwertige Zufallsvariablen.

Um eine Entscheidung im Testproblem (4.1) treffen zu können, benötigt man Verteilungsaussagen für die Teststatistiken LQ_n und RLQ_n unter H_0. In der Regel sind diese für endlichen Stichprobenumfang nur schwer zu erhalten, so dass man sich für die asymptotische Verteilung der Teststatistiken interessiert. Aus der Standardtheorie für Likelihood-Quotienten-Tests (vergleiche etwa Cox & Hinkley

(1974) Kapitel 9) ergäbe sich unter H_0 für beide Teststatistiken die χ^2_1-Verteilung als asymptotische Verteilung. Zu den Voraussetzungen, unter denen die Standardtheorie anwendbar ist, gehört aber, dass der Test basierend auf unabhängigen, identisch verteilten Zufallsvariablen $y_1, \ldots, y_n$ durchgeführt wird und dass der Parameter γ unter H_0 nicht auf dem Rand des Parameterraums liegt. Dies ist im vorliegenden Testproblem für $\gamma = 0$ und $\Omega = [0, \infty)$ aber der Fall.

Durch Self & Liang (1987) und Liang & Self (1996) wurde die Verteilung der Likelihood-Quotienten-Teststatistiken für den Fall hergeleitet, dass sich der Parameter unter H_0 auf dem Rand des Parameterraums befindet. Weiterhin ist jedoch Voraussetzung, dass der Test basierend auf unabhängigen, identisch verteilten Zufallvariablen $y_1, \ldots, y_n$ durchgeführt wird. Für Testproblem (4.1) würde man unter H_0 basierend auf der Theorie von Self & Liang erhalten, dass für die Verteilungen von LQ_n und RLQ_n

$$LQ_n \xrightarrow{\mathcal{D}} 0.5\chi^2_0 + 0.5\chi^2_1 \text{ und } RLQ_n \xrightarrow{\mathcal{D}} 0.5\chi^2_0 + 0.5\chi^2_1$$

bei $n \to \infty$ gilt. Die asymptotische Verteilung beider Likelihood-Quotienten-Teststatistiken wäre also die $(0.5, 0.5)$-Mischung zweier χ^2-Verteilungen mit null und einem Freiheitsgrad. Die χ^2_0-Verteilung entspricht dabei der Einpunktverteilung im Punkt Null, die asymptotische Verteilung besitzt also eine positive Wahrscheinlichkeitsmasse im Punkt Null.

Im linearen gemischten Modell ist die Voraussetzung unabhängiger Beobachtungen $y_1, \ldots, y_n$, zumindest unter der Alternative H_1, in der Regel nicht erfüllt, weil durch den zufälligen Effekt b Korrelationen zwischen den Beobachtungen entstehen. Außerdem sind die Beobachtungen selbst unter H_0 meist nicht identisch verteilt, sondern unterscheiden sich durch ihren Erwartungswert $\mathbb{E}(y_i) = x_i'\beta$. Dennoch lässt sich die Theorie von Self & Liang in bestimmten Spezialfällen auf Longitudinaldaten anwenden, wie Stram & Lee (1994, 1995) für die Likelihood-Quotienten-Teststatistik LQ_n und Morell (1998) für die Restricted-Likelihood-Quotienten-Teststatistik RLQ_n gezeigt haben. Die Grundlage hierfür liegt in der Möglichkeit, den Vektor aller Beobachtungen in unabhängige, identisch verteilte Subvektoren zu zerlegen, die jeweils die Beobachtungen einer Gruppe enthalten. Dazu nimmt man an, dass die Designmatrizen X_i der verschiedenen Gruppen identisch sind und betrachtet nun die Gruppen als Beobachtungen. Unter H_0 erhält man dann unabhängige und identisch verteilte Subvektoren, so dass sich

die Theorie aus Self & Liang (1987) anwenden lässt. Man beachte, dass gleiche Designmatrizen X_i insbesondere einschließen, dass alle Gruppen die gleiche Zahl an Beobachtungen aufweisen und dass nun die Zahl der Gruppen als Stichprobenumfang betrachtet wird. Man geht also in der asymptotischen Betrachtung davon aus, dass die Zahl der Gruppen gegen unendlich geht.

Häufig ist jedoch die Zahl der Gruppen in der Analyse von Longitudinaldaten relativ gering, so dass die Anwendbarkeit der Theorie von Self & Liang auch unter den angegebenen Einschränkungen fraglich erscheint. Pinheiro & Bates (2000) (Kapitel 2.4.1) haben die Verteilungen von LQ_n und RLQ_n für einen solchen Fall mit nur 11 Gruppen per Simulation untersucht und erhebliche Differenzen zwischen der theoretischen, asymptotischen Verteilung und der simulierten Verteilung gefunden. Dabei ergab sich, dass eine andere Wahl der Mischungsgewichte als (0.5,0.5) in einigen Fällen eine wesentlich bessere Anpassung an die simulierte Verteilung lieferte. Diese Beobachtungen lassen sich mit Hilfe der im Folgenden vorgestellten Ergebnisse erklären. Man vergleiche hierzu auch Kapitel 4.2, in dem genauer auf ein einfaches Modell für Longitudinaldaten eingegangen wird.

Im allgemeinen linearen gemischten Modell ist eine Zerlegung des Vektors y in unabhängige Subvektoren zumindest unter der Alternative in der Regel überhaupt nicht möglich. Beispielsweise besitzt die Kovarianzmatrix von y in einem linearen gemischten Modell, das durch die Reparametrisierung eines P-Splines entsteht, keine blockdiagonale Struktur, wie dies für Longitudinaldaten der Fall ist. Es müssen also alle Beobachtungen als korreliert beziehungsweise abhängig betrachtet werden. Zudem ist in einem solchen Modell auch unter der Nullhypothese die Annahme identisch verteilter y_i meist nicht erfüllt, da sich für die verschiedenen Beobachtungen unterschiedliche Kovariablenvektoren x_i ergeben.

Es sollen nun Ergebnisse für das allgemeine lineare gemischte Modell, beruhend auf Crainiceanu et al. (2002) und Crainiceanu & Ruppert (2002) hergeleitet, beziehungsweise vorgestellt werden. Diese Ergebnisse lassen sich folgendermaßen zusammenfassen: Die asymptotischen Verteilungen von LQ_n und RLQ_n sind von der Form

$$p_0\chi_0^2 + (1-p_0)G,$$

wobei für viele Fälle $p_0 > 0.5$ und $G \sim a\chi_1^2$ mit $a < 1$ gilt. Man erhält also wieder die Mischung zweier Verteilungen als asymptotische Verteilung. Die erste Verteilung ist die Einpunktverteilung im Punkt Null und korrespondiert

zur Wahrscheinlichkeit, tatsächlich $\hat{\tau} = 0$ beziehungsweise $\hat{\gamma} = 0$ zu schätzen. Der zweite Teil entspricht häufig einer skalierten χ^2_1-Verteilung. Aus $p_0 > 0.5$ und $a < 1$ folgt, dass die Verwendung von Quantilen der $(0.5, 0.5)$-Mischung der χ^2_0- und der χ^2_1-Verteilung zur Durchführung des Tests (4.1) zu konservativeren Testentscheidungen führt, als die Verwendung der exakten asymptotischen Verteilung.

4.1.1 Maxima der Likelihood-Quotienten im Punkt Null

Bevor die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken genauer bestimmt werden, sollen nun zunächst noch die Wahrscheinlichkeiten für Maxima des Likelihood-Quotienten $LQ_n(\gamma)$ und des Restricted-Likelihood-Quotienten $RLQ_n(\gamma)$ im Punkt Null untersucht werden. Diese Wahrscheinlichkeiten wurden für Random-Walk-Modelle bereits in Shephard & Harvey (1989) sowie Shephard (1993) betrachtet. Zu den allgemeineren Modellen, die hier behandelt werden, findet man weitere Details in Crainiceanu et al. (2002).

Betrachtet man globale Maxima, so entsprechen die interessierenden Wahrscheinlichkeiten zum einen den Punktmassen der Verteilungen von LQ_n und RLQ_n im Punkt Null und zum anderen den Wahrscheinlichkeiten $\hat{\gamma}_{ML} = 0$ beziehungsweise $\hat{\gamma}_{REML} = 0$ zu erhalten. Aufgrund der ersten Eigenschaft ist es so möglich, die von Pinheiro & Bates (2000) festgestellten Abweichungen von der (0.5,0.5)-Mischung zweier χ^2-Verteilungen nicht nur per Simulation, sondern auch theoretisch zu untersuchen. Außerdem sind Ergebnisse für diese Wahrscheinlichkeiten auch in Fällen erhältlich, in denen die später hergeleiteten asymptotischen Verteilungen von LQ_n und RLQ_n nicht bestimmbar sind. Insbesondere erhält man auch Ergebnisse unter der Alternative, während die Herleitung der asymptotischen Verteilungen nur unter der Nullhypothese möglich ist.

Leider lassen sich für globale Maxima im Allgemeinen keine Aussagen über die interessierenden Wahrscheinlichkeiten machen, so dass man stattdessen die Wahrscheinlichkeiten für lokale Maxima betrachtet, die eine obere Schranke für die Wahrscheinlichkeiten globaler Maxima liefern. Darüber hinaus bilden die Wahrscheinlichkeiten für lokale Maxima eine gute Approximation der Wahrscheinlichkeiten globaler Maxima, wie Crainiceanu et al. (2002) gezeigt haben.

Da diese Wahrscheinlichkeiten bereits für endlichen Stichprobenumfang erhältlich

sind, erlauben sie darüber hinaus in den Beispielen, die in den Abschnitten 4.2 bis 4.4 behandelt werden, eine Beurteilung der Frage, wie stark die Abweichungen zwischen der asymptotischen Verteilung und der Verteilung bei endlichem Stichprobenumfang ausfallen.

Aufgrund ihrer Definitionen kann man anstelle von Maxima der Likelihood-Quotienten äquivalent Maxima der Profile-Log-Likelihood $l(\gamma)$ beziehungsweise der Profile-Restricted-Log-Likelihood $l^*(\gamma)$ betrachten, wie dies in der folgenden Herleitung der Fall sein wird. Hinreichend und notwendig für ein lokales Maximum der Funktion $l(\gamma)$ in 0 ist, dass

$$\frac{\partial l(\gamma)}{\partial \gamma} \leq 0$$

für $\gamma = 0$ erfüllt ist, so dass sich für die interessierende Wahrscheinlichkeit eines lokalen Maximums von $l(\gamma)$ im Punkt Null der Ausdruck

$$\mathbb{P}_\gamma \left(\frac{\partial l(\gamma)}{\partial \gamma}\bigg|_{\gamma=0} \leq 0 \right)$$

ergibt. Man beachte dabei, dass γ in den folgenden Betrachtungen zwei verschiedene Bedeutungen besitzt. Einerseits bezeichnet γ das Argument der Funktion $l(\gamma)$ und andererseits den wahren Parameter γ. Dass die obige Ungleichung für $\gamma = 0$ gelten soll, bedeutet also, dass die Ableitung, ausgewertet am Funktionswert $\gamma = 0$, die Ungleichung erfüllen soll. Um die interessierenden Wahrscheinlichkeiten zu bestimmen, wird dabei der Parameter γ als wahrer Parameter zugrunde gelegt, was durch die Bezeichnung der Wahrscheinlichkeiten mit $\mathbb{P}_\gamma(\cdot)$ ausgedrückt werden soll.

Zunächst ist nun die Ableitung $\frac{\partial l(\gamma)}{\partial \gamma}$ herzuleiten. Da sich die Determinante einer Matrix als Produkt der Eigenwerte schreiben lässt (Toutenburg (2003), Seite 484), erhält man $|V_\gamma| = \prod_{i=1}^n (1+\gamma d_i)$, wobei d_i, $i = 1, \ldots, n$ die Eigenwerte der Matrix ZZ' bezeichnet. Dass sich die Eigenwerte von V_γ als $1 + \gamma d_i$ ergeben folgt dabei aus der Spektralzerlegung $ZZ' = UDU'$ mit deren Hilfe sich V_γ darstellen lässt als $V_\gamma = U(I_n + \gamma D)U'$, mit der Matrix der orthogonalen Eigenvektoren U und der Diagonalmatrix der Eigenwerte D.

Man erhält so

$$\frac{\partial}{\partial \gamma} \log(|V_\gamma|) = \sum_{i=1}^n \frac{d_i}{1 + \gamma d_i}.$$

Mit (A.18) aus Anhang A.3 gilt außerdem

$$\frac{\partial}{\partial\gamma}(y - X\hat{\beta}(\gamma))'V_\gamma^{-1}(y - X\hat{\beta}(\gamma)) = -(y - X\hat{\beta}(\gamma))'V_\gamma^{-1}ZZ'V_\gamma^{-1}(y - X\hat{\beta}(\gamma)),$$

so dass sich die folgenden Äquivalenzen ergeben:

$$\begin{aligned}\frac{\partial l(\gamma)}{\partial\gamma} \leq 0 \quad &\Leftrightarrow \quad n\frac{(y - X\hat{\beta}(\gamma))'V_\gamma^{-1}ZZ'V_\gamma^{-1}(y - X\hat{\beta}(\gamma))}{(y - X\hat{\beta}(\gamma))'V_\gamma^{-1}(y - X\hat{\beta}(\gamma))} - \sum_{i=1}^{n}\frac{d_i}{1+\gamma d_i} \leq 0 \\ &\Leftrightarrow \quad n\frac{y'R_\gamma'V_\gamma^{-1}ZZ'V_\gamma^{-1}R_\gamma y}{y'R_\gamma'V_\gamma^{-1}R_\gamma y} - \sum_{i=1}^{n}\frac{d_i}{1+\gamma d_i} \leq 0.\end{aligned}$$

Mit R_γ wird dabei wieder die in (4.2) definierte Residualmatrix bezeichnet.

Die Wahrscheinlichkeit eines lokalen Maximums des Likelihood-Quotienten $LQ_n(\gamma)$ im Punkt Null erhält man nun durch Einsetzen des Funktionswertes $\gamma = 0$:

$$\mathbb{P}_\gamma\left(\left.\frac{\partial l(\gamma)}{\partial\gamma}\right|_{\gamma=0} \leq 0\right) = \mathbb{P}_\gamma\left(\frac{y'R_0ZZ'R_0y}{y'R_0y} \leq \frac{1}{n}\sum_{i=1}^{n}d_i\right).$$

Dabei wurde ausgenutzt, dass $R_0 = I_n - X(X'X)^{-1}X'$ idempotent ist und $V_0 = I_n$ gilt. Zur weiteren Umformung verwendet man

$$\sum_{i=1}^{n}d_i = \operatorname{spur}(ZZ') = \operatorname{spur}(Z'Z).$$

Insbesondere die zweite Form ist numerisch vorteilhaft, weil die Matrix $Z'Z$ von einer wesentlich geringeren Dimension ist als ZZ'. Da y gemäß $\mathrm{N}(0, \sigma^2V_\gamma)$ verteilt ist, lässt sich nun die obige Wahrscheinlichkeit mit dem $\mathrm{N}(0, I_n)$-verteilten Zufallsvektor u und der symmetrischen Wurzel $V_\gamma^{\frac{1}{2}}$ umschreiben zu

$$\begin{aligned}\mathbb{P}_\gamma\left(\left.\frac{\partial l(\gamma)}{\partial\gamma}\right|_{\gamma=0} \leq 0\right) &= \mathbb{P}\left(\frac{u'V_\gamma^{\frac{1}{2}}R_0ZZ'R_0V_\gamma^{\frac{1}{2}}u}{u'V_\gamma^{\frac{1}{2}}R_0V_\gamma^{\frac{1}{2}}u} \leq \frac{1}{n}\operatorname{spur}(Z'Z)\right) \qquad (4.3)\\ &= \mathbb{P}(u'(A - \bar{d}B)u \leq 0), \qquad (4.4)\end{aligned}$$

wobei die Bezeichnungen $A = V_\gamma^{\frac{1}{2}}R_0ZZ'R_0V_\gamma^{\frac{1}{2}}$, $B = V_\gamma^{\frac{1}{2}}R_0V_\gamma^{\frac{1}{2}}$ und $\bar{d} = \frac{1}{n}\operatorname{spur}(Z'Z)$ benutzt wurden. Zusätzlich ist zu beachten, dass die Matrix B positiv semidefinit ist.

Die gesuchte Wahrscheinlichkeit wird also bestimmt durch die quadratische Form $u'(A - \bar{d}B)u$ in unabhängigen, standardnormalverteilten Zufallsvariablen $u \sim$

$\mathrm{N}(0, I_n)$. Die Verteilung solcher quadratischen Formen wird beispielsweise in Johnson & Kotz (1970) Kapitel 29 untersucht. Algorithmen zur exakten Berechnung von Wahrscheinlichkeiten der Form (4.4) findet man in Farebrother (1990) oder Davies (1980). Alternativ lassen sich diese Wahrscheinlichkeiten auch per Simulation bestimmen. Dazu bietet sich auch eine andere Darstellung an, die man über die Eigenwerte ψ_i der Matrix $A - \bar{d}B$ erhält:

$$\mathbb{P}_\gamma \left(\left. \frac{\partial l(\gamma)}{\partial \gamma} \right|_{\gamma=0} \leq 0 \right) = \mathbb{P} \left(\sum_{i=1}^{n} \psi_i v_i \right). \tag{4.5}$$

Dabei werden $v_1, \ldots, v_n$ als unabhängig und identisch χ^2_1-verteilt angenommen.

Gilt für den wahren Parameter $\gamma = 0$, so erhält man für die Matrizen A und B die einfacheren Formeln

$$\begin{aligned} A &= R_0 Z Z' R_0 \\ B &= R_0 = I_n - X(X'X)^{-1}X'. \end{aligned}$$

Sind zusätzlich die Designmatrizen X und Z orthogonal, das heißt $Z'X = 0$, so vereinfacht sich A weiter zu $A = ZZ'$.

In ähnlicher Weise lässt sich die Wahrscheinlichkeit für ein lokales Maximum des Restricted-Likelihood-Quotienten im Punkt Null bestimmen. Zunächst erhält man wieder die hinreichende und notwendige Bedingung

$$\left. \frac{\partial l^*(\gamma)}{\partial \gamma} \right|_{\gamma=0} \leq 0$$

für ein solches lokales Maximum. Zur Bestimmung der Ableitung $\frac{\partial l^*(\gamma)}{\partial \gamma}$ benutzt man

$$\frac{\partial}{\partial \gamma} \log(|V_\gamma|) + \frac{\partial}{\partial \gamma} \log(|X'V_\gamma X|) = \sum_{i=1}^{q} \frac{\omega_i}{1 + \gamma\omega_i},$$

wobei mit $\omega_1, \ldots, \omega_q$ die Eigenwerte der Matrix $Z'R_0Z$ bezeichnet werden (vergleiche Crainiceanu et al. (2002)) und erhält so, wieder unter Verwendung von (A.18), die folgende Äquivalenz:

$$\frac{\partial l^*(\gamma)}{\partial \gamma} \leq 0 \Leftrightarrow (n-p-1)\frac{y'R_\gamma' V_\gamma^{-1} Z Z' V_\gamma^{-1} R_\gamma y}{y'R_\gamma' V_\gamma^{-1} R_\gamma y} - \sum_{i=1}^{q} \frac{\omega_i}{1+\gamma\omega_i} \leq 0.$$

Durch Einsetzen des Funktionswertes $\gamma = 0$ bestimmt sich die Wahrscheinlichkeit eines lokalen Maximums des Restricted-Likelihood-Quotienten im Punkt Null als

$$\begin{aligned}
&\mathbb{P}_\gamma\left(\left.\frac{\partial l^*(\gamma)}{\partial\gamma}\right|_{\gamma=0} \leq 0\right) \\
&= \mathbb{P}_\gamma\left(\frac{y'R_0ZZ'R_0y}{y'R_0y} \leq \frac{1}{n-p-1}\sum_{i=1}^{q}\omega_i\right) \\
&= \mathbb{P}\left(\frac{u'V_\gamma^{\frac{1}{2}}R_0ZZ'R_0V_\gamma^{\frac{1}{2}}u}{u'V_\gamma^{\frac{1}{2}}R_0V_\gamma^{\frac{1}{2}}u} \leq \frac{1}{n-p-1}\,\mathrm{spur}(Z'R_0Z)\right) \qquad (4.6) \\
&= \mathbb{P}\left(u'\left(A - \frac{1}{n-p-1}\,\mathrm{spur}(Z'R_0Z)B\right)u \leq 0\right). \qquad (4.7)
\end{aligned}$$

Dabei bezeichnet u wieder einen $\mathrm{N}(0, I_n)$-verteilten Zufallsvektor. Die Auswertung der Wahrscheinlichkeit kann dann, wie zuvor beschrieben, entweder per Simulation oder über einen der angegebenen exakten Algorithmen erfolgen.

Man beachte, dass die Wahrscheinlichkeiten für lokale Maxima des Likelihood-Quotienten $LQ_n(\gamma)$ und des Restricted-Likelihood-Quotienten $RLQ_n(\gamma)$ im Punkt Null im Wesentlichen durch die gleichen Matrizen A und B bestimmt sind. Der einzige Unterschied besteht in der Tatsache, dass bei Betrachtung des Restricted-Likelihood-Quotienten $\bar{d}$ durch $\frac{1}{n-p-1}\,\mathrm{spur}(Z'R_0Z)$ ersetzt wird. Die Ähnlichkeit wird sogar noch ausgeprägter, wenn man von orthogonalen Designmatrizen des fixen und des zufälligen Effekts ausgeht. Dann vereinfacht sich nämlich $\frac{1}{n-p-1}\,\mathrm{spur}(Z'R_0Z)$ zu $\frac{n}{n-p-1}\bar{d}$. Insbesondere erhält man in diesem Fall für $n \to \infty$ die gleichen Wahrscheinlichkeiten für den Likelihood-Quotienten und den Restricted-Likelihood-Quotienten.

4.1.2 Asymptotische Ergebnisse

Nun sollen die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken unter H_0 vorgestellt werden. Diese hängen im Wesentlichen vom asymptotischen Verhalten der Eigenwerte $\kappa_{s,n}$, $s = 1, \ldots, q$ der Matrix $Z'R_0Z$ und der Eigenwerte $\pi_{s,n}$, $s = 1, \ldots, q$ der Matrix $Z'Z$ ab. Zunächst zeigen Crainiceanu & Ruppert (2002) und Kuo (1999), dass sich die Profile-Log-Likelihood und die Profile-Restricted-Log-Likelihood in Abhängigkeit von diesen Eigenwerten schrei-

ben lassen als

$$l(\gamma) \stackrel{\mathcal{D}}{=} -\frac{1}{2}\sum_{s=1}^{q}\log(1+\gamma\pi_{s,n}) - \frac{n}{2}\log\left[\sigma^2\left(\sum_{s=1}^{q}\frac{1}{1+\gamma\kappa_{s,n}}v_s + \sum_{s=q+1}^{n-p-1}v_s\right)\right]$$

und

$$l^*(\gamma) \stackrel{\mathcal{D}}{=} -\frac{1}{2}\sum_{s=1}^{q}\log(1+\gamma\kappa_{s,n}) - \frac{n-p-1}{2}\log\left[\sigma^2\left(\sum_{s=1}^{q}\frac{1}{1+\gamma\kappa_{s,n}}v_s + \sum_{s=q+1}^{n-p-1}v_s\right)\right],$$

wobei v_s, $s = 1, \ldots, n-p-1$ als unabhängig und identisch χ^2_1-verteilt angenommen werden. Dann erhält man das folgende, allgemeine Resultat (Crainiceanu & Ruppert 2002):

Es gelte H_0 aus (4.1). Existiert ein $a \geq 0$, so dass für die Eigenwerte $\kappa_{s,n}$, $s = 1, \ldots, q$ der Matrix $Z'P_0Z$ die Bedingung $\lim_{n\to\infty} n^{-a}\kappa_{s,n} = \kappa_s$ erfüllt ist, wobei mindestens ein κ_s positiv sein soll, so gilt für $n \to \infty$

$$RLQ_n(n^{-a}\gamma) \xrightarrow{\mathcal{D}} RLQ_\infty(\gamma) \tag{4.8}$$

mit

$$RLQ_\infty(\gamma) = \sum_{s=1}^{q}\frac{\gamma\kappa_s}{1+\gamma\kappa_s}v_s - \sum_{s=1}^{q}\log(1+\gamma\kappa_s). \tag{4.9}$$

Mit $\xrightarrow{\mathcal{D}}$ wird in diesem Fall die schwache Konvergenz bezeichnet, die eine Verallgemeinerung der Verteilungskonvergenz beispielsweise auf Zufallsvariablen in Funktionenräumen darstellt. Man beachte, dass es sich sowohl bei $RLQ_n(\gamma)$ als auch bei $RLQ_\infty(\gamma)$ um Zufallsgrößen im Raum der stetigen, positiven Funktionen $\mathcal{C}[0,\infty)$ handelt und die schwache Konvergenz also in diesem Funktionenraum erfolgt.

Erfüllen weiterhin die Eigenwerte $\pi_{s,n}$, $s = 1, \ldots, q$ der Matrix $Z'Z$ die Bedingung $\lim_{n\to\infty} n^{-a}\pi_{s,n} = \pi_s$, wobei wieder mindestens ein π_s positiv sein soll, so gilt für $n \to \infty$

$$LQ_n(n^{-a}\gamma) \xrightarrow{\mathcal{D}} LQ_\infty(\gamma) \tag{4.10}$$

mit

$$LQ_\infty(\gamma) = \sum_{s=1}^{q}\frac{\gamma\kappa_s}{1+\gamma\kappa_s}v_s - \sum_{s=1}^{q}\log(1+\gamma\pi_s). \tag{4.11}$$

Im Folgenden soll das präsentierte Ergebnis für den Restricted-Likelihood-Quotienten $RLQ_n(\gamma)$ plausibel gemacht, dabei aber auf mathematisch vollständige

Beweise verzichtet werden. Für den Likelihood-Quotienten $LQ_n(\gamma)$ lassen sich die obigen Aussagen dann durch analoge Überlegungen plausibel machen.

Aus der Darstellung der Restricted-Log-Likelihood mit Hilfe der Eigenwerte $\kappa_{s,n}$ ergibt sich für den Restricted-Likelihood-Quotienten

$$\begin{aligned} RLQ_n(\gamma) &= 2l^*(\gamma) - 2l^*(0) \\ &= -\sum_{s=1}^{q} \log(1+\gamma\kappa_{s,n}) - (n-p-1)\log\left(\frac{\displaystyle\sum_{s=1}^{q}\frac{1}{1+\gamma\kappa_{s,n}}v_s + \sum_{s=q+1}^{n-p-1} v_s}{\displaystyle\sum_{s=1}^{q} v_s + \sum_{s=q+1}^{n-p-1} v_s}\right) \\ &= -\sum_{s=1}^{q} \log(1+\gamma\kappa_{s,n}) - (n-p-1)\log\left(1 + \frac{\displaystyle -\sum_{s=1}^{q}\frac{\gamma\kappa_{s,n}}{1+\gamma\kappa_{s,n}}v_s}{\displaystyle\sum_{s=1}^{n-p-1} v_s}\right). \end{aligned}$$

Mit Hilfe dieser Darstellung kann nun die punktweise Konvergenz der Funktion $RLQ_n(n^{-a}\gamma)$ gegen $RLQ_\infty(\gamma)$ verhältnismäßig einfach gezeigt werden.

Für den ersten Term von $RLQ_n(n^{-a}\gamma)$ gilt offensichtlich

$$\lim_{n\to\infty} \sum_{s=1}^{q} \log(1+n^{-a}\gamma\kappa_{s,n}) = \sum_{s=1}^{q} \log(1+\gamma\kappa_s).$$

Zur Betrachtung des zweiten Terms entwickelt man zunächst $\log(1+x)$ in eine Taylorreihe um den Punkt Null, so dass sich

$$\log(1+x) = x + o_{pr}(x)$$

ergibt, wobei $o_{pr}(x)$ eine Zufallsvariable bezeichnet, für die

$$\lim_{x\to 0} \frac{o_{pr}(x)}{x} = 0 \ \ \mathbb{P}\text{–f.s.}$$

gilt. Damit erhält man aufgrund von

$$\lim_{n\to\infty} \frac{1}{n-p-1} \sum_{s=1}^{n-p-1} v_s = 1 \ \ \mathbb{P}\text{–f.s.}$$

und

$$\lim_{n\to\infty} \frac{-\sum_{s=1}^{q} \frac{n^{-a}\gamma\kappa_{s,n}}{1+n^{-a}\gamma\kappa_{s,n}} v_s}{\sum_{s=1}^{n-p-1} v_s} = 0 \ \mathbb{P}-\text{f.s.}$$

für den zweiten Term von $RLQ_n(n^{-a}\gamma)$

$$\lim_{n\to\infty} -(n-p-1)\log\left(1+\frac{-\sum_{s=1}^{\gamma} \frac{n^{-a}\gamma\kappa_{s,n}}{1+n^{-a}\gamma\kappa_{s,n}} v_s}{\sum_{s=1}^{n-p-1} v_s}\right) = \sum_{s=1}^{q} \frac{\gamma\kappa_s}{1+d\kappa_s} v_s \ \mathbb{P}-\text{f.s.}$$

Damit ist die $\mathbb{P}$-fast sichere, punktweise Konvergenz von $RLQ_n(n^{-a}\gamma)$ gegen $RLQ_\infty(\gamma)$ und damit auch die punktweise Verteilungskonvergenz gezeigt. Um die schwache Konvergenz im Funktionenraum $\mathcal{C}[0,\infty)$ zu erhalten, weist man nach, dass $RLQ_n(n^{-a}\gamma)$ für beliebiges $M > 0$ eine straffe Folge (tight sequence) in $\mathcal{C}[0,M]$ bildet. Dann folgt mit Satz 8.1 aus Billingsley (1968) die Behauptung. Auf den Nachweis dieser Eigenschaft soll hier aber verzichtet und stattdessen auf den Anhang in Crainiceanu & Ruppert (2002) verwiesen werden.

Für die asymptotischen Verteilungen der Teststatistiken LQ_n und RLQ_n erhält man nun unter den gleichen Bedingungen wie für (4.8) und (4.10)

$$LQ_n \xrightarrow{\mathcal{D}} \sup_{\gamma\geq 0} LQ_\infty(\gamma) = LQ_\infty, \tag{4.12}$$

$$RLQ_n \xrightarrow{\mathcal{D}} \sup_{\gamma\geq 0} RLQ_\infty(\gamma) = RLQ_\infty \tag{4.13}$$

für $n \to \infty$. Man betrachte wieder den Anhang in Crainiceanu & Ruppert (2002) für einen Beweis dieser Aussagen, der sich im wesentlichen auf die Anwendung des Continuous-Mapping-Theorems stützt.

Um diese Ergebnisse mit der asymptotischen Verteilung aus Self & Liang (1987) vergleichen zu können, interessiert man sich für die Punktmassen der asymptotischen Verteilungen im Punkt Null, das heißt für die Wahrscheinlichkeit $\mathbb{P}_0(LQ_\infty = 0)$ beziehungsweise $\mathbb{P}_0(RLQ_\infty = 0)$. Für diese Wahrscheinlichkeiten existieren aber keine einfachen Formeln, so dass man wieder, wie für endlichen

Stichprobenumfang, die Wahrscheinlichkeiten

$$\mathbb{P}_0\left(\frac{\partial}{\partial\gamma}LQ_\infty(\gamma)\Big|_{\gamma=0}\leq 0\right)$$

beziehungsweise

$$\mathbb{P}_0\left(\frac{\partial}{\partial\gamma}RLQ_\infty(\gamma)\Big|_{\gamma=0}\leq 0\right)$$

als Approximationen verwendet. Diese entsprechen den Wahrscheinlichkeiten für lokale Maxima der zufälligen Funktionen $LQ_\infty(\gamma)$ beziehungsweise $RLQ_\infty(\gamma)$ in $\gamma = 0$. Durch Bestimmung der Ableitungen erhält man die Formeln

$$\mathbb{P}_0\left(\frac{\partial}{\partial\gamma}LQ_\infty(\gamma)\Big|_{\gamma=0}\leq 0\right)=\mathbb{P}\left(\sum_{s=1}^{q}\kappa_s v_s\leq\sum_{s=1}^{q}\pi_s\right) \tag{4.14}$$

beziehungsweise

$$\mathbb{P}_0\left(\frac{\partial}{\partial\gamma}RLQ_\infty(\gamma)\Big|_{\gamma=0}\leq 0\right)=\mathbb{P}\left(\sum_{s=1}^{q}\kappa_s v_s\leq\sum_{s=1}^{q}\kappa_s\right), \tag{4.15}$$

wobei $v_1,\ldots,v_q$ wieder unabhängige, identisch χ^2_1-verteilte Zufallsvariablen bezeichnen. Die Bestimmung der Wahrscheinlichkeiten kann nun erneut durch Simulation oder exakt über einen der Algorithmen aus Farebrother (1990) oder Davies (1980) erfolgen. In den Abschnitten 4.2 bis 4.4 werden diese Wahrscheinlichkeiten für verschiedene Situationen berechnet. Dabei stellt sich heraus, dass die Wahrscheinlichkeitsmasse im Punkt Null für alle Beispiele zum Teil wesentlich größer als 0.5 ist, so dass die Verwendung der asymptotischen Verteilung aus Self & Liang (1987) zu falschen Ergebnissen geführt hätte.

Um die vollständige Gestalt der asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken bestimmen zu können, ist man in fast allen Fällen auf Simulationen angewiesen, da analytische Ergebnisse nur selten herzuleiten sind. Dazu benötigt man die asymptotischen Eigenwerte $\kappa_1,\ldots,\kappa_q$ und $\pi_1,\ldots,\pi_q$. Deren Bestimmung ist allerdings bisher nur in Fällen möglich, in denen die Designmatrizen X und Z eine einfache Struktur aufweisen und darüber hinaus häufig auch orthogonal sind. Möglicherweise lässt sich über eine Orthogonalisierung der Designmatrizen hier eine allgemeinere Anwendbarkeit erzielen (vergleiche Crainiceanu & Ruppert (2002) Abschnitt 5.3, allgemein zur Orthogonalisierung der Designmatrizen auch Wand (2002) und Nychka (2000)).

Sind die asymptotischen Eigenwerte erhältlich, so lassen sich die asymptotischen Verteilungen relativ einfach simulieren. Dazu wird zunächst eine Menge von Gitterpunkten $0 = d_1 < d_2 < \ldots < d_{max}$ aus $[0, \infty)$ bestimmt. Um eine Zufallszahl aus der Verteilung von LQ_∞ oder RLQ_∞ zu erhalten, zieht man dann q unabhängige, χ^2_1-verteilte Zufallszahlen und wertet basierend auf diesen Zufallszahlen $LQ_\infty(\gamma)$ beziehungsweise $RLQ_\infty(\gamma)$ an den Gitterpunkten aus. Das Maximum dieser Funktionswerte kann dann (bei geeigneter Wahl des Gitters) als Zufallszahl aus der Verteilung von LQ_∞ beziehungsweise RLQ_∞ aufgefasst werden.

Die Qualität der Zufallszahlen hängt dabei natürlich stark von der Wahl des Gitters ab. Zum einen sind die Schrittweiten der Gitterpunkte zu bestimmen und zum anderen der maximale Gitterpunkt d_{max}. Anstelle der Verwendung eines äquidistanten Gitters erscheint die Verwendung eines nahe bei Null dichteren Gitters sinnvoller. Zur Simulation in den folgenden Beispielen wurde daher zunächst ein äquidistantes Gitter $t_2 < \ldots < t_{max}$ gewählt, und dann das Gitter der möglichen Werte von γ als $d_1 = 0, d_2 = 10^{t_2}, \ldots, d_{max} = 10^{t_{max}}$ bestimmt. Wichtig ist es, in der Simulation zu kontrollieren, dass nicht zu häufig $LQ_\infty(d_{max})$ beziehungsweise $RLQ_\infty(d_{max})$ als Zufallszahlen verwendet werden und, falls dies der Fall ist, den Wert t_{max} entsprechend zu vergrößern.

Die Simulation der asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken soll nun noch in einem Algorithmus zusammengefasst werden:

Algorithmus 6 (Simulation der asymptotischen Verteilungen von LQ_n und RLQ_n unter H_0)

(i) Bestimme das äquidistante Gitter $t_2 < \ldots < t_{max}$, beispielsweise mit $t_2 = -3$, $t_{max} = 2$ und $max = 100$ und daraus $d_1 = 0, d_2 = 10^{t_2}, \ldots, d_{max} = 10^{t_{max}}$.

(ii) Wähle die Zahl der zu simulierenden Zufallszahlen K und setze $k = 1$.

(iii) Erzeuge q unabhängige, χ^2_1-verteilte Zufallszahlen $v_1, \ldots, v_q$ und werte die Funktionen

$$LQ_\infty(\gamma) = \sum_{s=1}^{q} \frac{\gamma\kappa_s}{1+\gamma\kappa_s} v_s - \sum_{s=1}^{q} \log(1+\gamma\pi_s)$$

und

$$RLQ_\infty(\gamma) = \sum_{s=1}^{q} \frac{\gamma\kappa_s}{1+\gamma\kappa_s} v_s - \sum_{s=1}^{q} \log(1+\gamma\kappa_s)$$

basierend auf diesen Zufallszahlen an den Werten d_1 bis d_{max} aus. Dann erhält man Realisationen aus den asymptotischen Verteilungen von LQ_n und RLQ_n als

$$LQ_\infty^{(k)} = \max\{LQ_\infty(d_1), \ldots, LQ_\infty(d_{max})\}$$

beziehungsweise

$$RLQ_\infty^{(k)} = \max\{RLQ_\infty(d_1), \ldots, RLQ_\infty(d_{max})\}.$$

(iv) Falls $k < K$, setze $k = k+1$ und gehe zurück zu (iii). Ansonsten erhält man durch $LQ_\infty^{(1)}, \ldots, LQ_\infty^{(K)}$ und $RLQ_\infty^{(1)}, \ldots, RLQ_\infty^{(K)}$ die Stichproben aus den asymptotischen Verteilungen von LQ_n und RLQ_n.

Aus den mit Hilfe von Algorithmus 6 erzeugten Zufallszahlen der asymptotischen Verteilungen von LQ_n und RLQ_n lassen sich nun Quantile dieser Verteilungen bestimmen. Dann kann der Test (4.1) über diese Quantile durchgeführt werden. Bezeichnet $c_{1-\alpha}$ das $(1-\alpha)$-Quantil der asymptotischen Verteilung von LQ_n, so verwirft man die Nullhypothese zugunsten der Alternative, wenn für die Realisation der Likelihood-Quotienten-Teststatistik $LQ_n > c_{1-\alpha}$ gilt und erhält so einen (asymptotischen) Test zum Signifikanzniveau α. Völlig analog wird der Test über die Restricted-Likelihood-Quotienten-Teststatistik durchgeführt.

Abschließend soll noch kurz ein Ergebnis zu allgemeineren Tests über den Likelihood-Quotienten vorgestellt werden (vergleiche Crainiceanu & Ruppert (2002)). Betrachtet man das Testproblem

$$H_0 : \gamma = 0, \beta_{i_1} = \beta_{i_1}^0, \ldots, \beta_{i_k} = \beta_{i_k}^0$$

versus (4.15)

$$H_1 : \gamma > 0 \text{ oder } \beta_{i_1} \neq \beta_{i_1}^0 \text{ oder} \ldots \text{oder } \beta_{i_k} \neq \beta_{i_k}^0$$

mit $\{i_1, \ldots, i_k\} \subset \{1, \ldots, p\}$ und $k \leq p$, dann gilt für die Likelihood-Quotienten-Teststatistik LQ_n^k zu diesem Test unter H_0 und den selben Bedingungen wie in (4.8) und (4.10)

$$LQ_n^k \xrightarrow{\mathcal{D}} LQ_\infty + \chi_k^2$$

bei $n \to \infty$. Der erste Teil der Verteilung korrespondiert zum Test des Varianzparameters, während die χ_k^2-Verteilung aus dem Test der Regressionskoeffizienten hervorgeht.

Man beachte, dass dieser Test nur über die auf der Log-Likelihood basierende Likelihood-Quotienten-Teststatistik durchgeführt werden kann. Die Restricted-Likelihood-Quotienten-Teststatistik lässt sich nämlich nur bei unter H_0 und H_1 gleicher Matrix X verwenden. Der Grund hierfür liegt in der Tatsache, dass die Restricted-Log-Likelihood $l^*(\beta, \gamma, \sigma^2)$ nicht invariant ist gegenüber eineindeutigen Transformationen des Parameters β. Während die Log-Likelihood nämlich von β nur über die Residuen $y - X\beta$ abhängt, erhält man in der Restricted-Log-Likelihood zusätzlich den Summanden $-\frac{1}{2}\log(|X'V_\gamma^{-1}X|)$, der explizit von X abhängt.

4.2 ANOVA

Nun soll als erstes Beispiel für die Anwendung der vorgestellten Tests ein balanciertes ANOVA-Modell mit als zufällig angenommenen Gruppeneffekten behandelt werden. Dies ist eines der wenigen Beispiele, in denen sich die asymptotischen Verteilungen der Teststatistiken auch analytisch bestimmen lassen.

Dazu betrachtet man das Modell

$$y_{ij} = \beta_0 + b_i + \varepsilon_{ij}$$

mit $i = 1, \ldots, N$ und $j = 1, \ldots, J$. Mit N wird dabei die Zahl der Gruppen und mit J die Zahl der Beobachtungen pro Gruppe bezeichnet. Insgesamt erhält man also $n = N \cdot J$ Beobachtungen. Für die gruppenspezifischen Effekte b_i nimmt man an, dass sie unabhängig und identisch $\mathrm{N}(0, \tau)$-verteilt und unabhängig von den Störgrößen ε_{ij} sind. Für diese soll wieder $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ gelten.

Fasst man die Beobachtungen einer Gruppe in dem Vektor y_i zusammen, so ergibt sich für jede einzelne Gruppe das Modell

$$y_i = X_i\beta_0 + Z_i b_i + \varepsilon_i$$

mit

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iJ} \end{pmatrix}, \; X_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \; Z_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \; \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iJ} \end{pmatrix}$$

für $i = 1, \ldots, N$. Insbesondere gilt $X_1 = \ldots = X_N = Z_1 \ldots = Z_N$, das heißt, $y_1, \ldots, y_N$ sind unabhängig und identisch verteilt gemäß $\mathrm{N}(\mathbf{1}_J\beta_0, \sigma^2 V_{\gamma,i})$ mit $V_{\gamma,i} = I_J + \gamma Z_i Z_i'$. Man beachte, dass auch $V_{\gamma,1} = \ldots = V_{\gamma,N}$ gilt.

Um die Notwendigkeit der Modellierung eines Gruppeneffekts über zufällige Effekte zu überprüfen, kann man nun den Test auf $\tau = 0$ beziehungsweise auf $\gamma = 0$ durchführen. Lässt man dazu die Zahl der Gruppen gegen unendlich gehen, während die Zahl der Beobachtungen pro Gruppe fest bleibt, so lassen sich die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken aus Self & Liang (1987) wie in Stram & Lee (1994) und Morell (1998) zur Durchführung des Tests verwenden. Man erhält dann sowohl für LQ_n als auch für RLQ_n die (0.5,0.5)-Mischung einer χ^2_0- und einer χ^2_1-Verteilung als asymptotische Verteilung. In vielen konkreten Datensituationen ist jedoch die Zahl der Gruppen verhältnismäßig klein, während der Gesamtstichprobenumfang n relativ groß ist. Formalisiert man dies zu der Fragestellung, welche asymptotische Verteilung man erhält, wenn die Zahl der Gruppen fest ist, während die Zahl der Beobachtungen pro Gruppe gegen unendlich geht, so greift die Theorie aus Self & Liang (1987) nicht mehr. Stattdessen muss die oben beschriebene, allgemeinere Theorie zugrunde gelegt werden.

Dazu formuliert man zunächst das gesamte Modell in Matrixnotation:

$$y = X\beta_0 + Zb + \varepsilon,$$

mit

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, Z = \begin{pmatrix} Z_1 & 0 & \dots & \dots & 0 \\ 0 & Z_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & Z_N \end{pmatrix},$$

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

Um die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken herzuleiten, benötigt man nun die asymptotischen Eigenwerte $\pi_1, \dots, \pi_N$ und $\kappa_1, \dots, \kappa_N$. Man erhält offensichtlich, dass alle Eigenwerte der Matrix $Z'Z$ gleich J sind, das heißt, es gilt $\pi_{s,n} = J$ für $s = 1, \dots, N$ und damit

$$\lim_{n\to\infty} n^{-1}\pi_{s,n} = \pi_s = \frac{1}{N} \text{ für } s = 1, \dots, N.$$

Für die Eigenwerte der Matrix $Z'R_0Z$ erhält man $\kappa_{s,n} = J$ für $s = 1, \ldots, N-1$ und $\kappa_{N,n} = 0$. Damit ergibt sich

$$\lim_{n\to\infty} n^{-1}\kappa_{s,n} = \kappa_s = \frac{1}{N} \text{ für } s = 1, \ldots, N-1 \text{ und } \lim_{n\to\infty} n^{-1}\kappa_{N,n} = \kappa_N = 0.$$

Mit Hilfe dieser einfachen Ausdrücke für die Eigenwerte, lassen sich nun die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken analytisch herleiten. Maximieren der allgemeinen Formeln in (4.9) und (4.11) liefert

$$LQ_\infty \stackrel{\mathcal{D}}{=} \left(X_{N-1} - N - N \log\left(\frac{X_{N-1}}{N}\right)\right) \mathbf{1}_{\{X_{N-1}>N\}} \tag{4.16}$$

und

$$RLQ_\infty \stackrel{\mathcal{D}}{=} \left(X_{N-1} - (N-1) - (N-1) \log\left(\frac{X_{N-1}}{N-1}\right)\right) \mathbf{1}_{\{X_{N-1}>N-1\}}, \tag{4.17}$$

wobei X_{N-1} eine χ^2_{N-1}-verteilte Zufallsgröße bezeichnet. Die Wahrscheinlichkeitsmasse der beiden Verteilungen im Punkt Null ist gegeben durch $\mathbb{P}(X_{N-1} > N-1)$ und $\mathbb{P}(X_{N-1} > N)$. Beide Wahrscheinlichkeiten sind größer als 0.5, konvergieren aber für $N \to \infty$ aufgrund des zentralen Grenzwertsatzes gegen 0.5. Dies ist im Einklang mit der Theorie von Self & Liang (1987), die als asymptotische Verteilung bei $N \to \infty$ eine (0.5,0.5)-Mischung von χ^2-Verteilungen mit null und einem Freiheitsgrad vorsieht.

	$N = 10$	$N = 20$	$N = 40$	$N = 80$	$N = 160$	$N = 320$
LQ_∞	0.6495	0.6054	0.5744	0.5526	0.5372	0.5263
RLQ_∞	0.5627	0.5432	0.5301	0.5212	0.5149	0.5105

Tabelle 4.1: Wahrscheinlichkeitsmassen im Punkt Null für die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistik LQ_n und der Restricted-Likelihood-Quotienten-Teststatistik RLQ_n bei verschiedenen Gruppenanzahlen N.

In Tabelle 4.1 sind die angegebenen Wahrscheinlichkeiten für verschiedene Gruppenanzahlen wiedergegeben. Wie man sieht, sind auch für eine relativ große Zahl von Gruppen die Wahrscheinlichkeiten noch deutlich größer als 0.5. Dabei ist die 0.5-Approximation für die asymptotische Verteilung der Likelihood-Quotienten-Teststatistik LQ_n wesentlich weiter vom wahren Wert entfernt, als für die Restricted-Likelihood-Quotienten-Teststatistik RLQ_n.

Zum Vergleich sind in Tabelle 4.2 noch einige ausgewählte Wahrscheinlichkeiten für lokale Maxima der Likelihood-Quotienten im Punkt Null bei endlichem

Stichprobenumfang und verschiedenen Gruppenanzahlen sowie Gruppengrößen angegeben. Die entsprechenden Werte wurden mit Hilfe von S-Plus-Funktionen bestimmt, die in Anhang C beschrieben werden. Man beachte, dass die Wahrscheinlichkeiten für lokale Maxima in Null nicht nur unter H_0, sondern auch für verschiedene Werte von γ unter H_1 erhältlich sind. Wie zu erwarten war, nehmen diese Wahrscheinlichkeiten immer stärker ab, je weiter man sich von der Nullhypothese entfernt. Außerdem erkennt man, dass sich bei steigender Gruppengröße J die Wahrscheinlichkeiten unter H_0 den asymptotischen Werten aus Tabelle 4.1 annähern.

$N = 10$								
	$LQ_n(\gamma)$				$RLQ_n(\gamma)$			
J	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
5	0.6227	0.5872	0.3302	0.0054	0.5442	0.5076	0.2634	0.0035
10	0.6384	0.5613	0.1700	0.0004	0.5531	0.4791	0.1270	0.0003
20	0.6431	0.4980	0.0523	0.0000	0.5564	0.4117	0.0376	0.0000
50	0.6467	0.3275	0.0042	0.0000	0.5603	0.2612	0.0027	0.0000
$N = 20$								
	$LQ_n(\gamma)$				$RLQ_n(\gamma)$			
J	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
5	0.5879	0.5299	0.1899	0.0001	0.5335	0.4775	0.1594	0.0000
10	0.5954	0.4844	0.0512	0.0000	0.5344	0.4261	0.0402	0.0000
20	0.6002	0.3872	0.0048	0.0000	0.5421	0.3350	0.0034	0.0000
50	0.6024	0.1812	0.0001	0.0000	0.5426	0.0146	0.0000	0.0000

Tabelle 4.2: Wahrscheinlichkeiten für lokale Maxima des Likelihood-Quotienten $LQ_n(\gamma)$ beziehungsweise des Restricted-Likelihood-Quotienten $RLQ_n(\gamma)$ im Punkt Null bei verschiedenen Gruppengrößen, Gruppenanzahlen und Werten von γ.

In Abbildung 4.1 sind die Verteilungen von LQ_∞ und RLQ_∞, das heißt die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistik LQ_n und der Restricted-Likelihood-Quotienten-Teststatistik RLQ_n, bedingt auf $LQ_\infty > 0$ beziehungsweise bedingt auf $RLQ_\infty > 0$ für verschiedene Gruppenanzahlen in Form von QQ-Plots wiedergegeben. Dabei werden die Quantile der χ^2_1-Verteilung, die sich aus der Theorie für unabhängige, identisch verteilte Beobachtungen nach Self & Liang (1987) für die bedingte Verteilung ergäbe, den Quantilen der auf $LQ_\infty > 0$ beziehungsweise $RLQ_\infty > 0$ bedingten Verteilungen aus (4.16) und (4.17) gegenübergestellt. Die Quantile der asymptotischen Verteilungen von LQ_n und RLQ_n wurden dabei aus Zufallsstichproben vom Umfang 100000 berechnet,

die mit Hilfe einer in Anhang C beschriebenen S-Plus-Funktion erzeugt wurden. Man beachte, dass unter der Bedingung $LQ_\infty > 0$ beziehungsweise $RLQ_\infty > 0$ nur noch ein Anteil von zwischen 50% und 65% der erzeugten Zufallszahlen übrig bleibt. Die restlichen Zufallszahlen entsprechen den Fällen, in denen der Likelihood-Quotient tatsächlich gleich 0 ist.

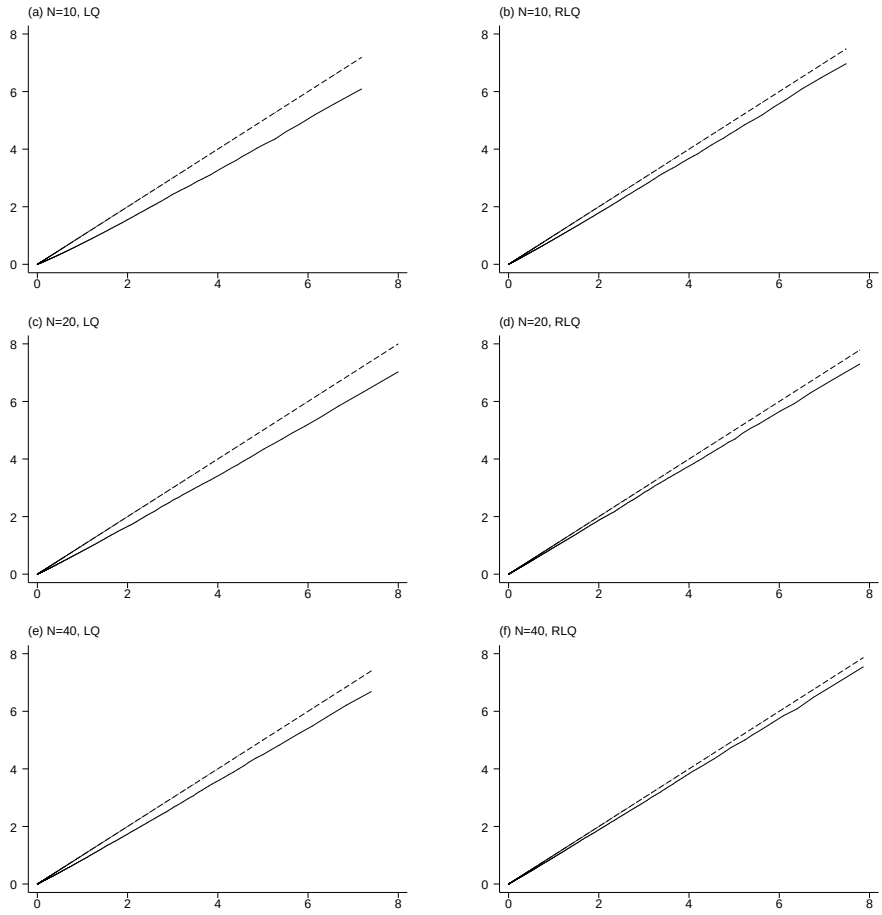

Abbildung 4.1: QQ-Plots der Verteilungen von LQ_∞ beziehungsweise RLQ_∞ bedingt auf $LQ_\infty > 0$ beziehungsweise $RLQ_\infty > 0$ und der χ^2_1-Verteilung bei verschiedenen Gruppenanzahlen N. Die Quantile der χ^2_1-Verteilung sind auf der horizontalen Achse abgetragen.

Wie man sieht, besitzen die asymptotischen Verteilungen jeweils die Form einer mit einem Faktor $a < 1$ skalierten χ^2_1-Verteilung. Mit steigender Gruppenzahl N nähert sich der Faktor a offenbar dem Wert 1, wobei analog zu den Wahrscheinlichkeiten in Tabelle 4.1 festzustellen ist, dass dieser Faktor für die asymptotische Restricted-Likelihood-Quotienten-Teststatistik RLQ_∞ für alle Gruppen-

zahlen wesentlich näher am Wert 1 liegt als für die asymptotische Likelihood-Quotienten-Teststatistik LQ_∞.

4.3 P-Splines

Wie in Kapitel 3.3.2 gezeigt wurde, ist die Modellierung einer nonparametrisch zu schätzenden Funktion über P-Splines äquivalent zu einem bestimmten linearen gemischten Modell. Es bezeichne nun wieder l den Grad des P-Splines, k die Ordnung der zur Penalisierung verwendeten Differenzen und m die Zahl der verwendeten Knoten. Im Folgenden werden nur P-Splines vom Grad $l = 0$ behandelt, weil diese eine besonders einfach strukturierte Designmatrix B aufweisen. Die Matrix B besteht dabei wieder aus den B-Spline-Basisfunktionen, ausgewertet an den beobachteten Designpunkten. Bei B-Splines vom Grad $l = 0$ nehmen diese entweder den Wert 1 oder 0 an. Man vergleiche hierzu auch die Definition von B-Splines vom Grad 0 in Kapitel 3.1.1.

Aufgrund der Einschränkungen, unter denen die vorgestellten Tests anwendbar sind, können nur Modelle betrachtet werden, in denen der Zusammenhang zwischen einer einzelnen Kovariablen x und der abhängigen Variable y nonparametrisch modelliert werden soll. Man geht also von dem Modell

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

aus. Nach der Reparametrisierung erhält man dann ein lineares gemischtes Modell mit den Designmatrizen $X = B\tilde{X}$ und $Z = B\tilde{Z}$, wobei $\tilde{X}$ und $\tilde{Z}$ wie in Kapitel 3.3.2 definiert sind. Man beachte, dass bei der nonparametrischen Modellierung nur einer Funktion keine Identifizierbarkeitsprobleme auftreten und daher hier die ursprüngliche Definition der Matrix $\tilde{X}$ verwendet werden kann.

Durch die Anwendung des Tests (4.1) erhält man nun die Möglichkeit, den Glättungsparameter α auf ∞ und damit den Zusammenhang zwischen x und y auf ein Polynom vom Grad $k - 1$ zu testen. Dies entspricht der Durchführung des Tests

$$H_0 : f(x) = \beta_0 + \beta_1 x + \ldots + \beta_{k-1} x^{k-1}$$

versus

$$H_1 : f(x) \neq \beta_0 + \beta_1 x + \ldots + \beta_{k-1} x^{k-1}.$$

Insbesondere ergibt sich für $k = 2$ ein Test auf Linearität und für $k = 1$ ein Test auf den Effekt von x. Diese beiden Testmöglichkeiten werden in einer Simulationsstudie in Kapitel 5.3 näher untersucht werden.

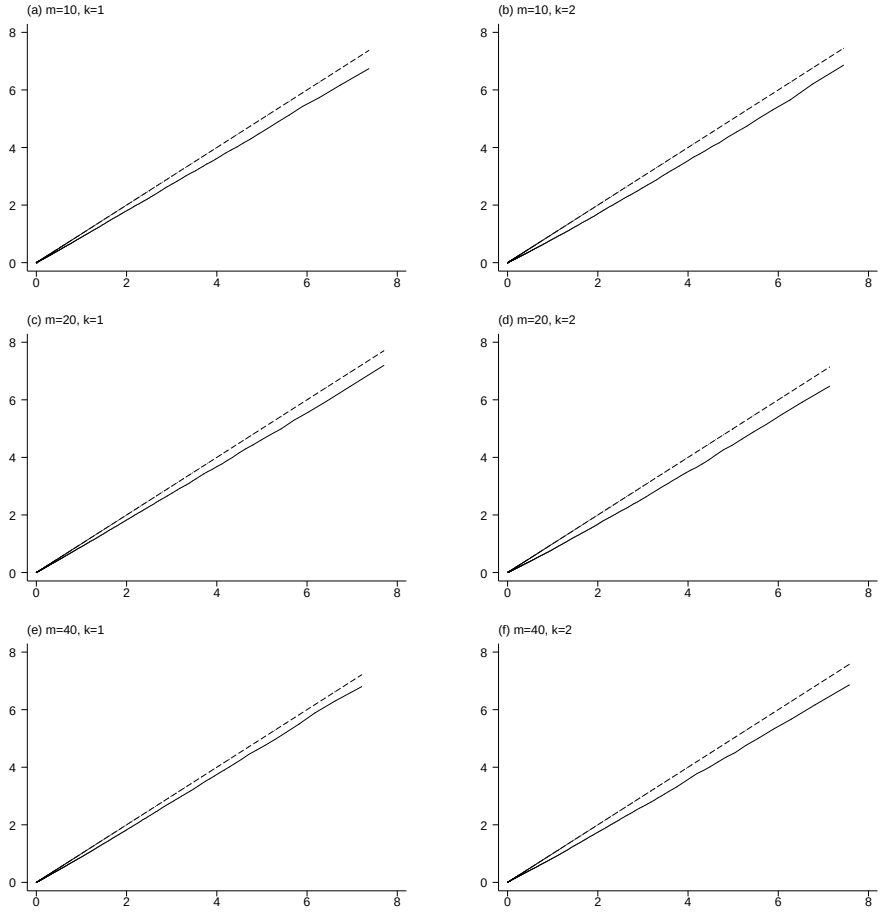

Abbildung 4.2: QQ-Plots der Verteilung von LQ_∞ bedingt auf $LQ_\infty > 0$ und der χ^2_1-Verteilung bei verschiedenen Knotenzahlen m und Differenzenordnungen k. Die Quantile der χ^2_1-Verteilung sind auf der horizontalen Achse abgetragen.

Zur Herleitung der asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken benötigt man die asymptotischen Eigenwerte der Matrizen $Z'R_0Z$ und $Z'Z$. Aufgrund der Konstruktion der Designmatrizen erhält man $Z'X = 0$, das heißt, die Designmatrizen sind orthogonal. Dies folgt aus der Tatsache, dass $B'B$ eine Diagonalmatrix ist und für $\tilde{X}$ und $\tilde{Z}$ stets $\tilde{X}'\tilde{Z} = 0$ gilt (vergleiche Kapitel 3.3.2). Damit vereinfacht sich $Z'R_0Z$ zu $Z'Z$ und es gilt $\kappa_{s,n} = \pi_{s,n}$. Insbesondere sind also die asymptotische Verteilung für die Likelihood-Quotienten-Teststatistik und die Restricted-Likelihood-Quotienten-Teststatistik identisch. Im Folgenden

werden nur die Bezeichnungen für die Likelihood-Quotienten-Teststatistik verwendet, die Ergebnisse gelten aber ebenso für den Restricted-Likelihood-Quotienten.

	unbedingt					
	$k=1$			$k=2$		
m	0.90	0.95	0.99	0.90	0.95	0.99
10	1.023	1.948	4.358	0.886	1.769	4.178
20	1.023	1.958	4.423	0.858	1.740	4.169
40	1.015	1.969	4.526	0.857	1.751	4.147
	bedingt auf LQ_∞					
	$k=1$			$k=2$		
m	0.90	0.95	0.99	0.90	0.95	0.99
10	2.443	3.475	6.047	2.337	3.393	6.044
20	2.484	3.541	6.115	2.301	3.362	5.979
40	2.503	3.596	6.310	2.378	3.409	5.982

Tabelle 4.3: Quantile der Verteilung von LQ_∞ bei verschiedenen Knotenzahlen m und Differenzenordnungen k. In der zweiten Tabelle sind die Quantile unter der Bedingung $LQ_\infty > 0$ bestimmt, in der ersten Tabelle ohne diese Bedingung.

Für die Matrix $Z'Z$ erhält man aus der Definition von Z

$$Z'Z = (DD')^{-1}DB'BD'(DD')^{-1},$$

wobei D die Differenzenmatrix des P-Splines bezeichnet. Aufgrund der einfachen Struktur der Matrix B lassen sich nun unter gewissen Annahmen die asymptotischen Eigenwerte bestimmen. Es gilt nämlich $B'B = N = \text{diag}(n_1, \ldots, n_r)$ mit $r = m+l-1 = m-1$. Das heißt, $B'B$ ist eine Diagonalmatrix, deren Einträge n_j gegeben sind durch die Anzahl der Beobachtungen, deren Kovariablenausprägungen x_i in das j-te, durch die Knoten des P-Splines gebildete Intervall $[\xi_j, \xi_{j+1})$ fallen. Nimmt man nun an, dass der Anteil der Beobachtungen in diesem Intervall für $n \to \infty$ gegen $h_j = \lim_{n\to\infty} \frac{n_j}{n}$ konvergiert, so erhält man die Eigenwerte $\pi_1, \ldots, \pi_q$ mit $q = r - k$ als Eigenwerte der Matrix

$$(DD')^{-1}D\tilde{N}D'(DD')^{-1},$$

mit $\tilde{N} = \text{diag}(h_1, \ldots, h_r)$. Damit kann nun die asymptotische Verteilung von LQ_n, wie in Algorithmus 6 beschrieben, simuliert werden.

Es wurden wieder 100000 Zufallszahlen aus der Verteilung von LQ_∞ erzeugt, wobei angenommen wurde, dass die Werte der Kovariablen asymptotisch über den

Wertebereich von x gleichverteilt sind. Für die asymptotischen Anteile h_j erhält man unter dieser Annahme $h_j = \frac{1}{m-1}$. Zur Durchführung der Simulation und auch für die weiteren Berechnungen wurden wieder S-Plus-Funktionen verwendet, die in Anhang C dokumentiert sind.

In den QQ-Plots in Abbildung 4.2 sind die Quantile der so simulierten Verteilung unter der Bedingung $LQ_\infty > 0$ gegen Quantile einer χ^2_1-Verteilung abgetragen. Die einzelnen Verteilungen unterscheiden sich dabei durch die Anzahl der Knoten m und die Ordnung der Differenzen k. Auch hier besitzen die asymptotischen Verteilungen bedingt auf $LQ_\infty > 0$ wieder die Form skalierter χ^2_1-Verteilungen.

In Tabelle 4.3 sind eine Reihe von Quantilen der asymptotischen Verteilungen verzeichnet, mit deren Hilfe die entsprechenden Tests durchgeführt werden können und die im Rahmen der Simulationsstudie in Kapitel 5.3 verwendet werden. Die Quantile der auf $LQ_\infty > 0$ bedingten Verteilung erlauben einen Vergleich mit den entsprechenden Quantilen der χ^2_1-Verteilung.

	$m = 10$	$m = 20$	$m = 40$
$k = 1$	0.6467	0.6511	0.6495
$k = 2$	0.6659	0.6695	0.6801

Tabelle 4.4: Wahrscheinlichkeitsmasse der Verteilung von LQ_∞ im Punkt Null.

	$m = 10$	$m = 20$	$m = 40$
$k = 1$	0.6563	0.6579	0.6575
$k = 2$	0.6754	0.6731	0.6803

Tabelle 4.5: Wahrscheinlichkeiten für lokale Maxima der Funktion $LQ_\infty(\gamma)$ im Punkt Null.

Zusätzlich sind in den Tabellen 4.4 und 4.5 Informationen über die Wahrscheinlichkeitsmasse der asymptotischen Verteilung der Likelihood-Quotienten-Teststatistik LQ_n im Punkt Null zusammengestellt. Tabelle 4.4 enthält diese Wahrscheinlichkeitsmasse für verschiedene Knotenzahlen und Differenzen der Ordnung $k = 1$ und $k = 2$. Die Wahrscheinlichkeiten basieren dabei auf der Simulation der Verteilung von LQ_∞, die auch den QQ-Plots zugrunde liegt. Tabelle 4.5 dagegen enthält die Wahrscheinlichkeiten für lokale Maxima der Funktion $LQ_\infty(\gamma)$ im Punkt Null für die gleichen Knotenzahlen und Differenzenordnungen. Wie man sieht, bilden die Wahrscheinlichkeiten in Tabelle 4.5 eine obere Schranke für die

$m = 10$								
	$k = 1$				$k = 2$			
n	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
50	0.6414	0.5424	0.2502	0.0269	0.6576	0.5271	0.2381	0.0476
100	0.6477	0.4861	0.1520	0.0063	0.6671	0.4597	0.1634	0.0229
200	0.6515	0.3907	0.0742	0.0012	0.6716	0.3711	0.1010	0.0088
500	0.6533	0.2524	0.0191	0.0000	0.6733	0.2469	0.0437	0.0021
$m = 20$								
	$k = 1$				$k = 2$			
n	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
50	0.6454	0.4773	0.1570	0.0144	0.6582	0.2500	0.0535	0.0067
100	0.6486	0.3875	0.0823	0.0026	0.6680	0.1733	0.0268	0.0017
200	0.6534	0.2806	0.0339	0.0002	0.6742	0.1097	0.0118	0.0004
500	0.6552	0.1512	0.0063	0.0000	0.6755	0.0505	0.0031	0.0001
$m = 40$								
	$k = 1$				$k = 2$			
n	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
50	0.6465	0.3822	0.0876	0.0076	0.6508	0.0608	0.0073	0.0014
100	0.6509	0.2773	0.0368	0.0010	0.6692	0.0309	0.0022	0.0002
200	0.6511	0.1793	0.0116	0.0000	0.6726	0.0137	0.0006	0.0000
500	0.6575	0.0767	0.0014	0.0000	0.6761	0.0042	0.0001	0.0000

Tabelle 4.6: Wahrscheinlichkeiten für lokale Maxima der Funktion $LQ_n(\gamma)$ im Punkt Null.

Werte in Tabelle 4.4. Die Unterschiede sind aber relativ gering, so dass sich die Wahrscheinlichkeitsmassen im Punkt Null relativ gut durch die Wahrscheinlichkeiten aus Tabelle 4.5 approximieren lassen.

Tabelle 4.6 und 4.7 schließlich beinhalten die Wahrscheinlichkeiten für lokale Maxima der Funktionen $LQ_n(\gamma)$ beziehungsweise $RLQ_n(\gamma)$ im Punkt Null, also bei endlichem Stichprobenumfang. Wie man sieht, erreichen die Wahrscheinlichkeiten unter H_0 relativ schnell Werte, die nahe an den asymptotischen Grenzwerten liegen. Zusätzlich ist es wieder möglich, auch Wahrscheinlichkeiten unter H_1 zu bestimmen.

$m = 10$								
	$k = 1$				$k = 2$			
n	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
50	0.6469	0.5537	0.2549	0.0272	0.6711	0.5380	0.2474	0.0501
100	0.6519	0.4880	0.1527	0.0069	0.6732	0.4637	0.1680	0.0238
200	0.6531	0.3936	0.0759	0.0010	0.6726	0.3721	0.1039	0.0094
500	0.6523	0.2486	0.0201	0.0000	0.6727	0.2474	0.0434	0.0021
$m = 20$								
	$k = 1$				$k = 2$			
n	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
50	0.6529	0.4843	0.1628	0.0150	0.6731	0.2558	0.0559	0.0071
100	0.6553	0.3911	0.0814	0.0025	0.6730	0.1796	0.0272	0.0018
200	0.6554	0.2858	0.0329	0.0003	0.6745	0.1112	0.0115	0.0006
500	0.6559	0.1511	0.0062	0.0000	0.6746	0.0499	0.0035	0.0001
$m = 40$								
	$k = 1$				$k = 2$			
n	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
50	0.6550	0.3932	0.0916	0.0084	0.6752	0.0626	0.0082	0.0017
100	0.6569	0.2841	0.0373	0.0010	0.6755	0.0320	0.0025	0.0003
200	0.6570	0.1800	0.0116	0.0001	0.6757	0.0144	0.0006	0.0000
500	0.6574	0.0774	0.0012	0.0000	0.6761	0.0044	0.0001	0.0000

Tabelle 4.7: Wahrscheinlichkeiten für lokale Maxima der Funktion $RLQ_n(\gamma)$ im Punkt Null.

4.4 Markov-Zufallsfelder

Wie P-Splines besitzen auch Markov-Zufallsfelder eine Repräsentation als lineares gemischtes Modell, wie in Abschnitt 3.3.2 gezeigt wurde. Wieder kann mit den vorgestellten Tests nur ein Modell analysiert werden, in dem lediglich der räumliche Effekt vorkommt. Man betrachtet also das Modell

$$y_i = f_{spat}(R_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Der Test des Parameters γ auf 0 entspricht dann einem Test des Glättungsparameters des Markov-Zufallsfelds auf ∞ und damit einem Test auf das Vorhandensein eines räumlichen Effekts.

Die Reparametrisierung des Markov-Zufallsfeldes zu einem linearen gemischten Modell liefert Designmatrizen $X = B\tilde{X}$ und $Z = B\tilde{Z}$. Man beachte dabei wieder,

dass hier kein Identifizierbarkeitsproblem vorliegt und also die ursprüngliche Definition von $\tilde{X}$ für Markov-Zufallsfelder verwendet werden kann. Wie für P-Splines vom Grad $l = 0$ erhält man orthogonale Designmatrizen X und Z, da $B'B$ eine Diagonalmatrix ist und $\tilde{X}$ und $\tilde{Z}$ nach Konstruktion orthogonal sind. Damit gilt $\kappa_{s,n} = \pi_{s,n}$ und die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistik LQ_n und der Restricted-Likelihood-Quotienten-Teststatistik RLQ_n sind wieder identisch.

Aus der Definition der Matrix Z in Kapitel 3.3.2 erhält man

$$\begin{aligned} Z'Z &= (L'L)^{-1}L'B'BL(L'L)^{-1} \\ &= \tilde{\Omega}^{-1}L'NL\tilde{\Omega}^{-1}, \end{aligned}$$

mit L als Zerlegungsfaktor der Strafmatrix K und $\tilde{\Omega}$ als Diagonalmatrix der positiven Eigenwerte von K. Die Matrix $N = B'B$ ergibt sich als Diagonalmatrix $N = \text{diag}(n_1, \ldots, n_r)$, wobei r die Anzahl der Regionen und n_j die Zahl der Beobachtungen in Region j bezeichnet. Zur Bestimmung der asymptotischen Eigenwerte $\pi_1, \ldots, \pi_q$ mit $q = r - 1$ nimmt man nun an, dass für $n \to \infty$ der Anteil der Beobachtungen in Region j gegen $h_j = \lim_{n\to\infty} \frac{n_j}{n}$ konvergiert. Dann erhält man die asymptotischen Eigenwerte $\pi_1, \ldots, \pi_q$ als Eigenwerte der Matrix

$$\tilde{\Omega}^{-1}L'\tilde{N}L\tilde{\Omega}^{-1}$$

mit $\tilde{N} = \text{diag}(h_1, \ldots, h_r)$.

Die Eigenwerte $\pi_1, \ldots, \pi_q$ hängen nun nicht nur von bestimmten Charakteristika wie der Zahl der Regionen und den h_j ab, sondern auch von der speziellen Wahl der Strafmatrix K, da die positiven Eigenwerte von K in der Berechnung von $\pi_1, \ldots, \pi_q$ verwendet werden. Für die folgenden Berechnungen wurde die Strafmatrix aus der Simulation in Abschnitt 5.2 zugrunde gelegt, die eine Nachbarschaftsstruktur der Kreise aus Bayern und Baden-Württemberg definiert. Zwei Regionen werden dabei als benachbart betrachtet, wenn sie gemeinsame Grenzen besitzen. Die Karte besitzt 124 Regionen und ist in Abbildung 5.12 in Abschnitt 5.2 wiedergegeben. Zusätzlich wird davon ausgegangen, dass $h_j = \frac{1}{124}$ für $j = 1, \ldots, 124$ gilt, das heißt, dass asymptotisch betrachtet in allen Regionen der gleiche Anteil an Beobachtungen vorhanden ist.

Die asymptotische Verteilung der Likelihood-Quotienten-Teststatistik LQ_n wurde wieder mit Hilfe der in Anhang C beschriebenen Funktionen per Simulation

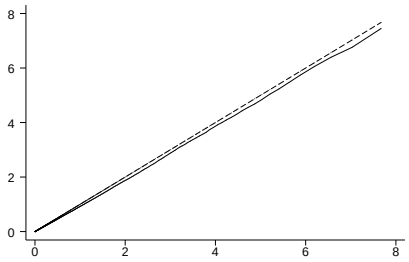

Abbildung 4.3: QQ-Plot der Verteilung von LQ_∞ bedingt auf $LQ_\infty > 0$ und der χ^2_1-Verteilung. Die Quantile der χ^2_1-Verteilung sind auf der horizontalen Achse abgetragen.

bestimmt und ist bedingt auf $LQ_\infty > 0$ in Abbildung 4.3 in Form eines QQ-Plots gegen die χ^2_1-Verteilung wiedergegeben. Auch diese Verteilung besitzt die Form einer mit einem Faktor $a < 1$ skalierten χ^2_1-Verteilung.

	$LQ_n(\gamma)$				$RLQ_n(\gamma)$			
n	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
124	0.5790	0.5481	0.3428	0.0277	0.5924	0.5599	0.3573	0.0288
248	0.5821	0.5196	0.1995	0.0010	0.5874	0.5208	0.2053	0.0012
496	0.5829	0.4859	0.1216	0.0000	0.5837	0.4891	0.1234	0.0001

Tabelle 4.8: Wahrscheinlichkeiten für lokale Maxima der Funktionen $LQ_n(\gamma)$ und $RLQ_n(\gamma)$ im Punkt Null.

Die Wahrscheinlichkeitsmasse der asymptotischen Verteilung im Punkt Null beträgt 0.5857 und ist damit größer als 0.5, die Wahrscheinlichkeit für ein lokales Maximum der Funktion $LQ_\infty(\gamma)$ im Punkt Null ist 0.5832. In Tabelle 4.8 sind außerdem Wahrscheinlichkeiten für lokale Maxima der Funktionen $LQ_n(\gamma)$ beziehungsweise $RLQ_n(\gamma)$ im Punkt Null für verschiedene Stichprobenumfänge und wahre Werte von γ zusammengestellt.

5 Simulationsstudien

Nachdem in Kapitel 3 ein Verfahren zur Schätzung von generalisierten geoadditiven gemischten Modellen vorgestellt wurde, stellt sich die Frage, wie gut dieses Verfahren die den Daten zugrunde liegenden Strukturen reproduzieren kann. Insbesondere stellt sich auch die Frage, inwiefern sich der vorgestellte Ansatz von anderen Verfahren unterscheidet, die ebenfalls die Schätzung von generalisierten geoadditiven gemischten Modellen oder Subklassen dieses Modells erlauben. Da sich in den komplexen Modellen aus Kapitel 3 bisher keine theoretischen Ergebnisse herleiten lassen, muss ein solcher Vergleich auf der Basis von Simulationsstudien erfolgen. Dazu soll in diesem Kapitel die Schätzung von zwei verschiedenen Modellklassen untersucht werden.

In einer ersten Simulationsstudie werden generalisierte additive Modelle behandelt. Zur Schätzung dieser Modelle existiert bereits eine relativ breite Software-Auswahl, die die automatische Wahl der Glättungsparameter beziehungsweise die automatische Optimierung der Knotenwahl erlauben. Im ersten Abschnitt werden daher Ergebnisse für eine ganze Reihe von Ansätzen zusammengefasst, wobei der Schwerpunkt aber auf dem in Kapitel 3 vorgestellten Verfahren liegen soll.

Betrachtet man den additiven Prädiktor eines generalisierten geoadditiven gemischten Modells, so ist die größere Komplexität dieses Modells im Vergleich zu reinen additiven Modellen offensichtlich. Dementsprechend existiert auch nur eine geringe Zahl von Programmen, die die Schätzung solcher Modelle erlauben. Im zweiten Abschnitt dieses Kapitel werden generalisierte geoadditive gemischte Modelle in ihrer vollen Komplexität mit Hilfe des Ansatzes aus Kapitel 3 und mit Hilfe eines vollen Bayes-Ansatzes behandelt.

Im letzten Abschnitt werden dann noch die Eigenschaften der Test-Möglichkeiten in der nonparametrischen Regression, die sich aus den in Kapitel 4 vorgestellten Likelihood-Quotienten-Tests ergeben, im Rahmen einer Simulationsstudie untersucht. Dazu wurden zum einen Modelle unter der Nullhypothese simuliert, um den Fehler erster Art abzuschätzen und zu überprüfen, ob die Tests das Signifikanzniveau einhalten. Zum anderen wurden Modelle unter der Alternative simuliert, um einen Eindruck von der Güte der Tests zu erhalten.

5.1 Generalisierte additive Modelle

5.1.1 Modell

Zur Simulation der Daten im generalisierten additiven Modell wurde der additive Prädiktor

$$\eta_i = c \cdot [\beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + f_5(x_{i5})] \tag{5.1}$$

verwendet. Der Verlauf der Funktionen f_1 bis f_5 ist in Abbildung 5.1 wiedergegeben. Der Faktor c dient der Skalierung des Modells und sichert, dass für alle simulierten Verteilungen vernünftige Schätzungen möglich sind. Bei binomialverteiltem Response ist dafür beispielsweise Voraussetzung, dass der Prädiktor im Intervall $[-3, 3]$ liegt, damit genügend Variation in den Daten vorhanden ist.

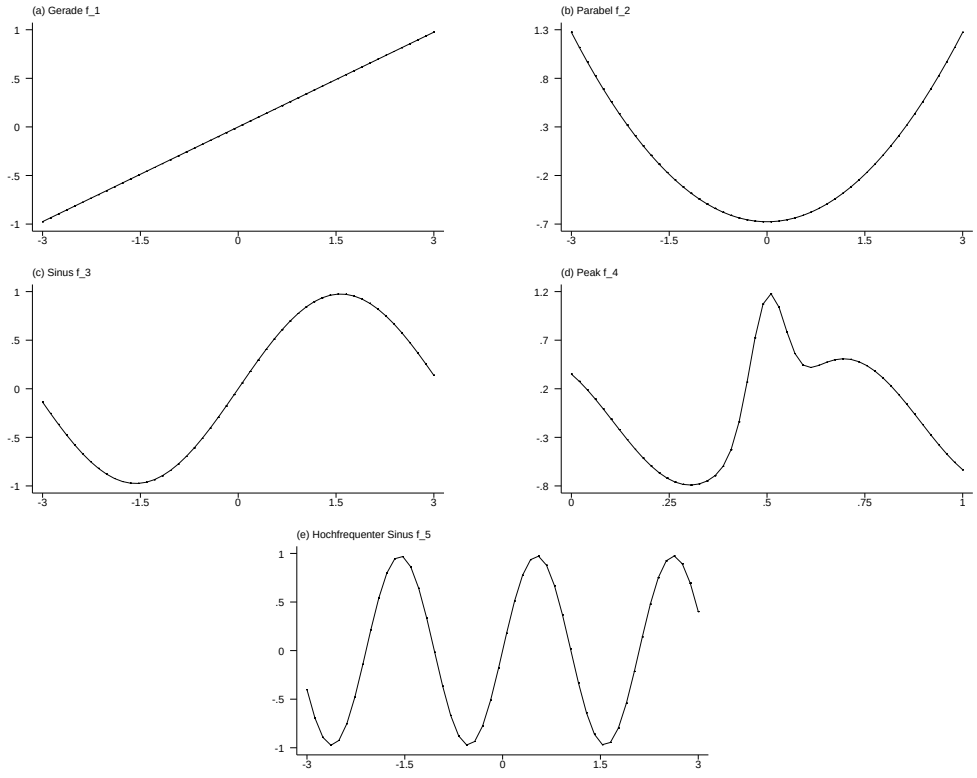

Abbildung 5.1: Die fünf in der Simulation zu generalisierten additiven Modellen verwendeten Funktionen.

Die fünf Funktionen stellen äußerst unterschiedliche Anforderungen an die Schätzverfahren, da sie sehr verschieden in der Ausprägung ihrer Nichtlinearität sind.

Während über die lineare Funktion f_1 getestet wird, wie gut ein zugrunde liegender, linearer Effekt auch bei nonparametrischer Modellierung als linear erkannt wird, weist f_5 eine starke Oszillation auf. An der Qualität der Schätzung von f_5 lässt sich also ablesen, wie flexibel die nonparametrische Schätzung ist. Eine Besonderheit stellt noch Funktion f_4 dar, die bei $x_4 = 0.5$ eine stark ausgeprägte Spitze besitzt. Von Interesse ist hier, wie gut diese Spitze in der Schätzung reproduziert wird beziehungsweise ob sie in vielen Fällen durch zu starke Glättung nicht in der Schätzung zu erkennen ist. Die Funktionen f_2 und f_3 sind jeweils Funktionen mittlerer Variabilität.

Simuliert wurden zunächst zwei Modelle unter Annahme der Normalverteilung, die sich im Signal-Rauschen-Verhältnis unterscheiden. Während das erste Modell für den Response eine Varianz von $\sigma^2 = 0.09$ zugrunde legt, also ein im Verhältnis zum Rauschen recht ausgeprägtes Signal, ist im zweiten Modell mit $\sigma^2 = 0.36$ eine deutlich stärkere Störung durch den Fehlerterm ε vorhanden. Zusätzlich wurden basierend auf dem additiven Prädiktor (5.1) bernoulliverteilte, also $\{0, 1\}$-wertige Zufallszahlen und poissonverteilte Zufallszahlen erzeugt. Der Stichprobenumfang beträgt in allen vier Modellen jeweils 500 Beobachtungen, die Zahl der Replikationen beträgt jeweils 250.

5.1.2 Verwendete Programme

Wie bereits erwähnt, existieren zur Schätzung generalisierter additiver Modelle bereits mehrere Programme, die eine automatische Schätzung ermöglichen. Im Folgenden sollen nun die Programme vorgestellt werden, die im Rahmen der Simulationsstudie verglichen wurden.

ggamm

Bei dem Programm `ggamm` handelt es sich um eine im Rahmen dieser Arbeit entstandene S-Plus-Implementation der in Kapitel 3 vorgestellten Verfahren. Zu einer genaueren Erläuterung der Verwendung und der Möglichkeiten dieser Implementation sei auf Anhang B verwiesen, der eine kurze Software-Beschreibung zu `ggamm` wiedergibt.

Zur Modellierung der glatten Funktionen mit `ggamm` wurden jeweils P-Splines vom Grad 3 mit 20 Knoten und auf zweiten Differenzen basierenden Penalisierungen verwendet.

BayesX

`BayesX` ist ein Programm zur bayesianischen, semiparametrischen Regressionsanalyse basierend auf Markov-Chain-Monte-Carlo-Verfahren. Möglich sind dabei insbesondere die Schätzung nonparametrischer Effekte metrischer Kovariablen, räumlicher Effekte und zufälliger Effekte. `BayesX` erlaubt also die Bestimmung der in Kapitel 3.1 vorgestellten Modelle in einem voll-bayesianischen Ansatz. Das heißt, nicht nur die Regressionsparameter werden als Zufallsvariablen mit geeigneten Priori-Verteilungen modelliert, sondern auch Varianz- beziehungsweise Glättungsparameter. Ähnlich wie bei der Reparametrisierung in Kapitel 3.3 werden die inversen Glättungsparameter dabei als unbekannte Varianzparameter interpretiert. Für unbekannte Varianzparameter wird in `BayesX` als Priori-Verteilung die inverse Gamma-Verteilung $\mathrm{IG}(a, b)$ mit Parametern $a > 0$ und $b > 0$ sowie der Dichte

$$p(x) \propto \frac{1}{x^{a+1}} \exp\left(-\frac{1}{bx}\right)$$

verwendet. Bei der inversen Gamma-Verteilung handelt es sich um die zur Normalverteilung konjugierte Verteilung zur Schätzung des Varianzparameters σ^2 (vergleiche Rüger (1999), Seite 207-210). Zur vollständigen Spezifizierung der Priori-Verteilung müssen noch die beiden Parameter a und b gewählt werden. Im Folgenden werden zwei Alternativen betrachtet: Zum einen die Möglichkeit, a und b klein zu wählen, was eine an Jeffreys Priori $p(x) \propto \frac{1}{x}$ angenäherte Priori-Verteilung liefert (vergleiche Rüger (1999) Seite 233-35). Konkret wird hier $a = b = 0.001$ verwendet. Zum anderen werden die Analysen mit den Standardeinstellungen aus `BayesX` ($a = 1$, $b = 0.001$) durchgeführt.

Zur Schätzung der nonparametrisch modellierten Funktionen wurden im Rahmen der Simulationsstudie P-Splines vom Grad 3 mit 20 Knoten und Differenzen der Ordnung $k = 2$ als Penalisierung verwendet. Für eine Einführung in das Programmpaket `BayesX` vergleiche man Brezger, Kneib & Lang (2002), zur zugrunde liegenden Theorie betrachte man Fahrmeir & Lang (2001a, 2001b). Speziell bayesianische P-Splines werden in Lang & Brezger (2002) behandelt.

BVCM

Das Programm `BVCM` bietet eine Möglichkeit der adaptiven Knotenwahl in generalisierten additiven Modellen basierend auf einem voll-bayesianischen Verfahren. Die Schätzung wird also nicht durch einen einzelnen Glättungsparameter gesteuert, sondern es werden sowohl Position als auch Zahl der Knoten optimiert. Im Rahmen der Simulationsstudie wurden dabei die Standardeinstellungen aus `BVCM` verwendet. Man vergleiche Biller (2000a) für eine detailliertere Beschreibung der genauen Vorgehensweise und Biller (2000b) für eine Beschreibung der `BVCM`-Software.

Bei der Analyse von Modellen mit poissonverteiltem Response mit `BVCM` wurden in vielen Fällen alle fünf Funktionen nahezu horizontal geschätzt, obwohl bei der Analyse mit anderen Programmen keine Probleme auftraten. Daher wurde die Verwendung von `BVCM` auf Modelle mit normalverteiltem Response und bernoulliverteiltem Response beschränkt.

mgcv

Bei dem Programmpaket `mgcv` handelt es sich um eine Reihe von Funktionen für die Statistik-Software R. Mit Hilfe von `mgcv` ist die automatische Bestimmung der Glättungsparameter mehrerer, über penalisierte Splines modellierter Funktionen möglich. Die Glättungsparameter werden dabei optimal bezüglich des generalisierten Kreuzvalidierungskriteriums $GCV(\alpha)$ gewählt. Man vergleiche Wood (2000) für Details über das entsprechende Vorgehen.

step.gam

Die S-Plus-Funktion `step.gam` erlaubt die bezüglich Akaikes Informationskriterium optimale Auswahl von Glättungsparametern $\alpha = (\alpha_1, \ldots, \alpha_s)'$ aus einem vorgegebenen s-dimensionalen Gitter möglicher Werte. Dabei werden nicht alle Möglichkeiten des Gitters berücksichtigt, was bei einer größeren Zahl von Funktionen schnell zu Rechenzeitproblemen führen würde. Stattdessen werden ausgehend von Startwerten $\alpha^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_s^{(0)})'$ iterativ neue Glättungsparameter $\alpha^{(k)}$ bestimmt, die näher an der optimalen Lösung sein sollen. Dazu werden in jeder Iteration circa $2 \cdot s$ Modelle über die S-Plus-Funktion `gam` geschätzt, in denen

jeweils ein Glättungsparameter auf dem Gitter um einen Wert nach oben oder unten versetzt wurde. Zusätzlich wird gespeichert, welche Modelle bereits berechnet wurden, so dass die Schätzung für diese Modelle nicht mehrmals durchgeführt werden muss und in jeder Iteration nicht exakt $2 \cdot s$ Modelle bestimmt werden. Aus den in einer Iteration resultierenden Modellen wird das Modell mit dem geringsten $AIC(\alpha)$ ausgewählt. Ist keine Verbesserung mehr möglich, so bricht das Programm ab und liefert die aktuellen Schätzungen. Durch die schrittweise Optimierung ausgehend von einem Startwert ist nicht garantiert, dass man sich am Ende des Iterationsprozesses tatsächlich an einem globalen Minimum von $AIC(\alpha)$ befindet, da jeweils nur in der Umgebung der aktuellen Werte nach einer optimaleren Lösung gesucht wird. Dadurch ist es möglich, dass die Funktion ein lokales Minimum als optimale Glättungsparameter bestimmt. Außerdem können natürlich nur Glättungsparameter aus den vorgegebenen Werten ausgewählt werden, so dass die Qualität der Schätzung auch von diesen Vorgaben abhängt.

Die einzelnen Funktionen wurden in der folgenden Simulationsstudie als P-Splines vom Grad 3 mit 20 Knoten und Differenzen der Ordnung 2 als Penalisierung modelliert. Als Liste möglicher Glättungsparameter wurde für jede der fünf Funktionen ausgehend von den äquidistanten Werten $t = (-2, -1.733, \ldots, 1.733, 2)'$ das aus 16 Werten bestehende Gitter $(10^{t_1}, \ldots, 10^{t_{16}})' = (0.01, 0.0185, \ldots, 54.117, 100)'$ verwendet. Zusätzlich wurde noch die Möglichkeit berücksichtigt, eine Funktion als linear einzustufen. Als Startwert wurde jeweils der mittlere Wert des Gitters benutzt.

PROC GAM

Innerhalb der SAS-Prozedur `PROC GAM` hat der Benutzer die Möglichkeit, die Glättungsparameter von Glättungssplines optimal bezüglich des generalisierten Kreuzvalidierungskriteriums wählen zu lassen. Leider sind keine weiteren Informationen über das konkret zur Optimierung verwendete Verfahren erhältlich.

5.1.3 Ergebnisse

Zum Vergleich der Schätzungen der verschiedenen Verfahren wird der empirische mittlere quadratische Fehler (mean squared error, MSE) betrachtet, der durch

$$MSE(f_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} \left[\hat{f}_j(x_{(i)j}) - f_j(x_{(i)j}) \right]^2 \tag{5.2}$$

definiert ist, wobei mit n_j die Anzahl der verschiedenen Ausprägungen der j-ten Kovariable und mit $x_{(1)j}, \ldots, x_{(n_j)j}$ die geordneten, verschiedenen Ausprägungen der j-ten Kovariable bezeichnet werden. In den Abbildungen 5.2 bis 5.5 sind die aus den verschiedenen Schätzverfahren resultierenden logarithmierten MSEs für die vier verschiedenen Verteilungen des Response als Boxplots wiedergegeben. Die MSEs wurden dabei zusätzlich logarithmiert, da auf der logarithmischen Skala Unterschiede deutlicher zu erkennen sind als bei direkter Verwendung der MSEs. In den Abbildungen ist jeweils der Median der logarithmierten MSEs, die sich aus den Schätzungen mit Hilfe der Funktion `ggamm` ergaben, durch eine Linie markiert. Damit ist ein unmittelbarer Vergleich des in dieser Arbeit vorgestellten Ansatzes mit den übrigen Schätzverfahren möglich. Man beachte noch, dass sich die Skalen der einzelnen Boxplots teilweise recht deutlich unterscheiden. Die Verwendung einer einheitlichen Skala würde aber den Vergleich erschweren, da in vielen Fällen die Unterschiede nicht mehr erkennbar wären.

Generell unterscheiden sich die verschiedenen Verfahren am stärksten in ihrer Fähigkeit die Linearität der Funktion f_1 zu erkennen. Wie man jeweils aus Grafik (a) der Abbildungen ablesen kann, lassen sich dabei zwei Gruppen von Verfahren unterscheiden. Während die voll-bayesianischen Ansätze, die in `BayesX` und `BVCM` implementiert sind, etwas größere MSEs liefern, sind diese für die übrigen Verfahren auf einem niedrigeren Niveau. Die Funktion `ggamm` liefert dabei gemeinsam mit `step.gam` die jeweils besten Ergebnisse. Es ist jedoch zu beachten, dass auch aus den Schätzungen mit den voll-bayesianischen Ansätzen meist deutlich zu erkennen ist, dass der Einfluss von x_1 auf den Response linear ist und sich auch für diese Verfahren grundsätzlich niedrige MSEs ergeben. Dies lässt sich auch aus einem Vergleich des Wertebereichs der mit (a) bezeichneten Grafiken mit den Wertebereichen der übrigen Boxplots in den Grafiken (b)-(e) erkennen.

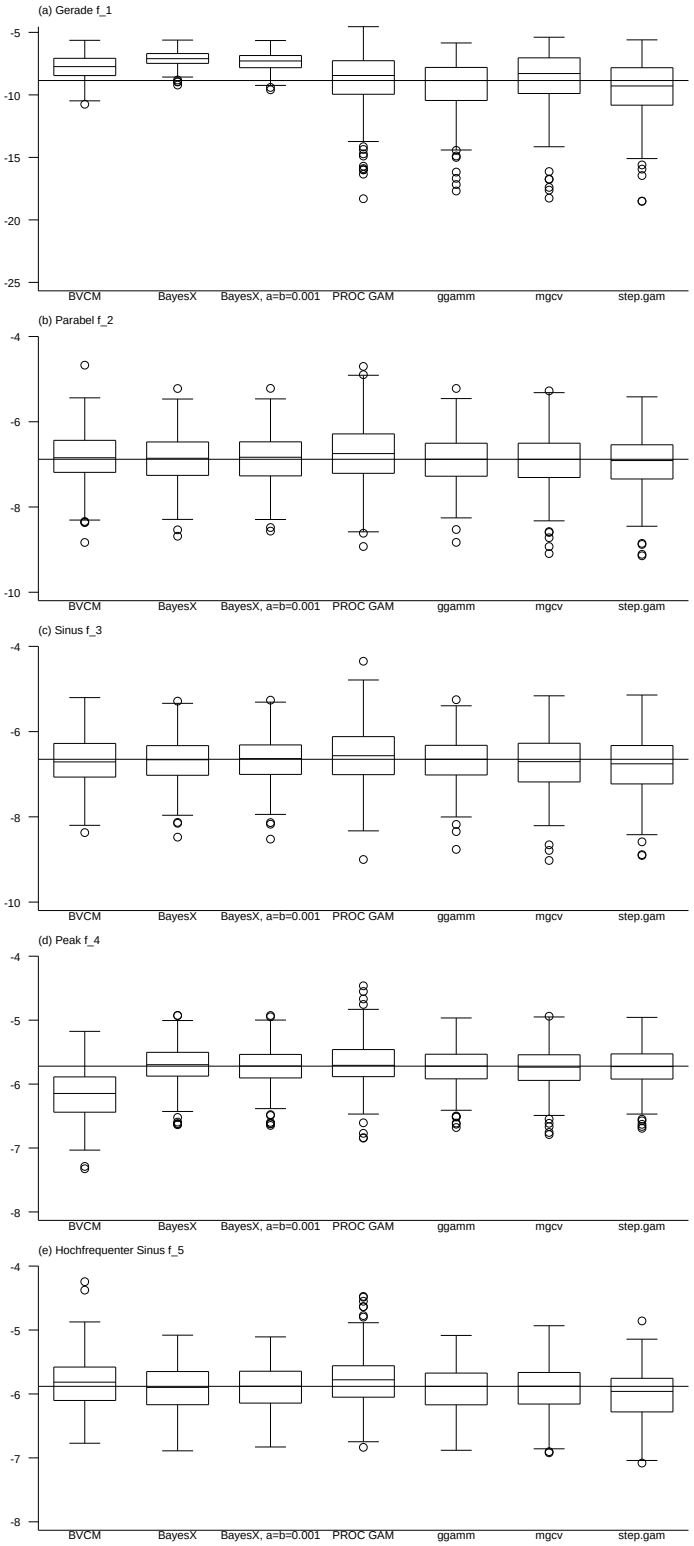

Abbildung 5.2: Normalverteilter Response: Boxplots der logarithmierten MSEs bei hohem Signal-Rauschen-Verhältnis ($\sigma^2 = 0.09$).

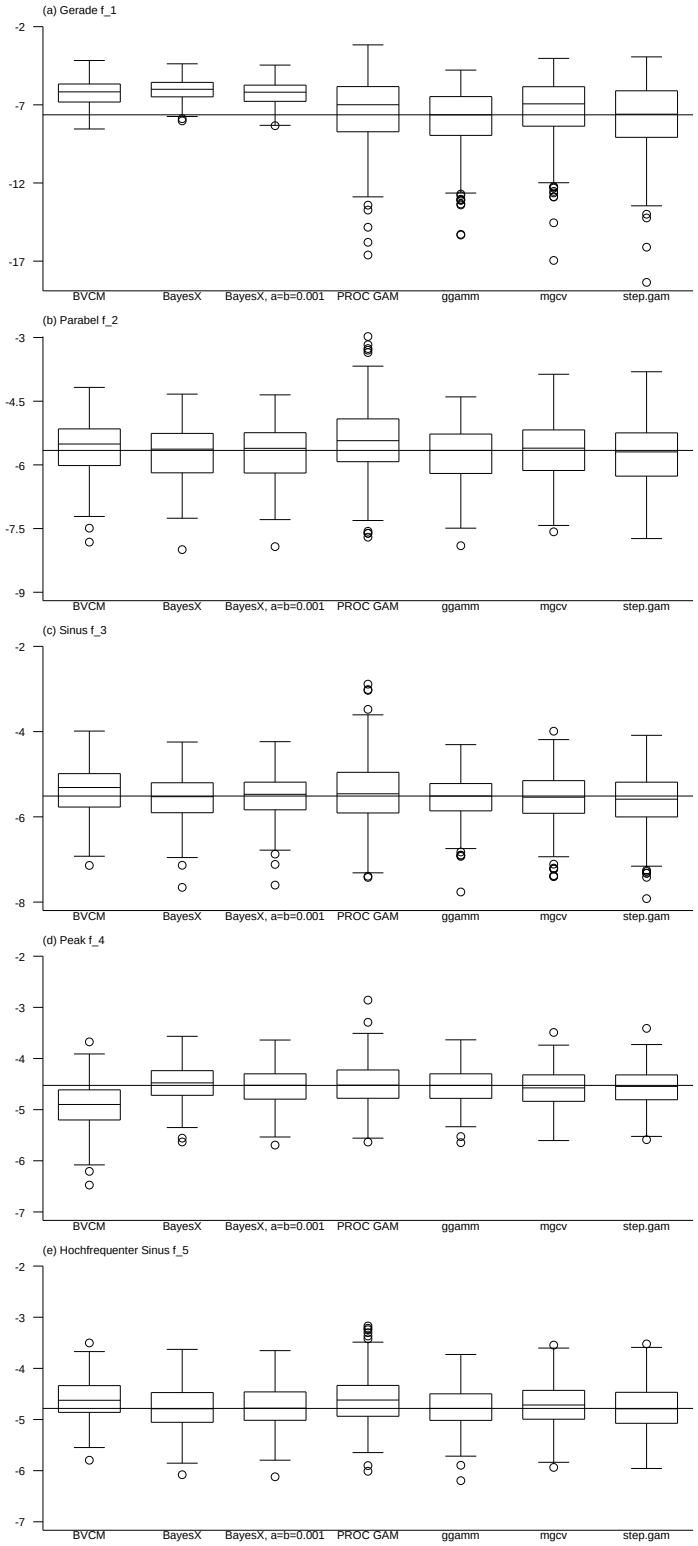

Abbildung 5.3: Normalverteilter Response: Boxplots der logarithmierten MSEs bei niedrigem Signal-Rauschen-Verhältnis ($\sigma^2 = 0.36$).

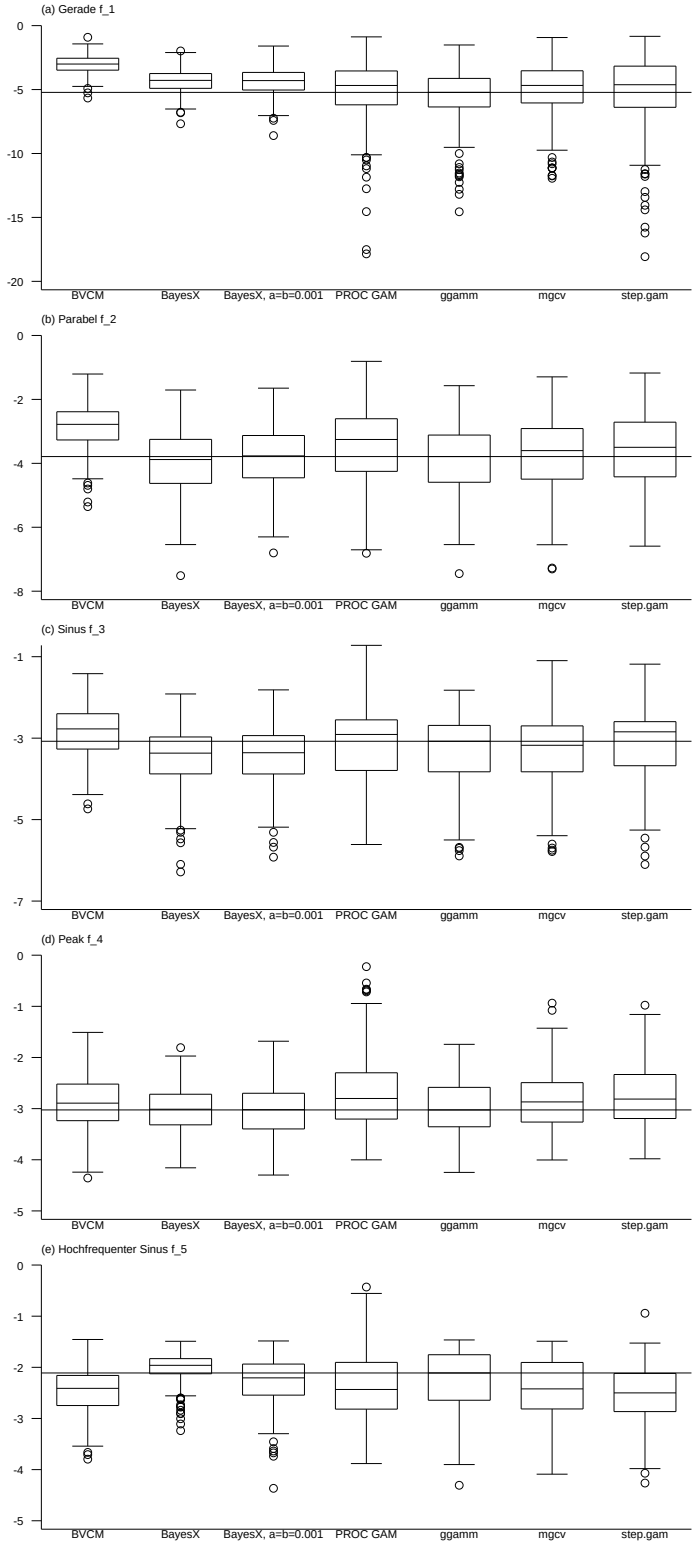

Abbildung 5.4: Bernoulliverteilter Response: Boxplots der logarithmierten MSEs.

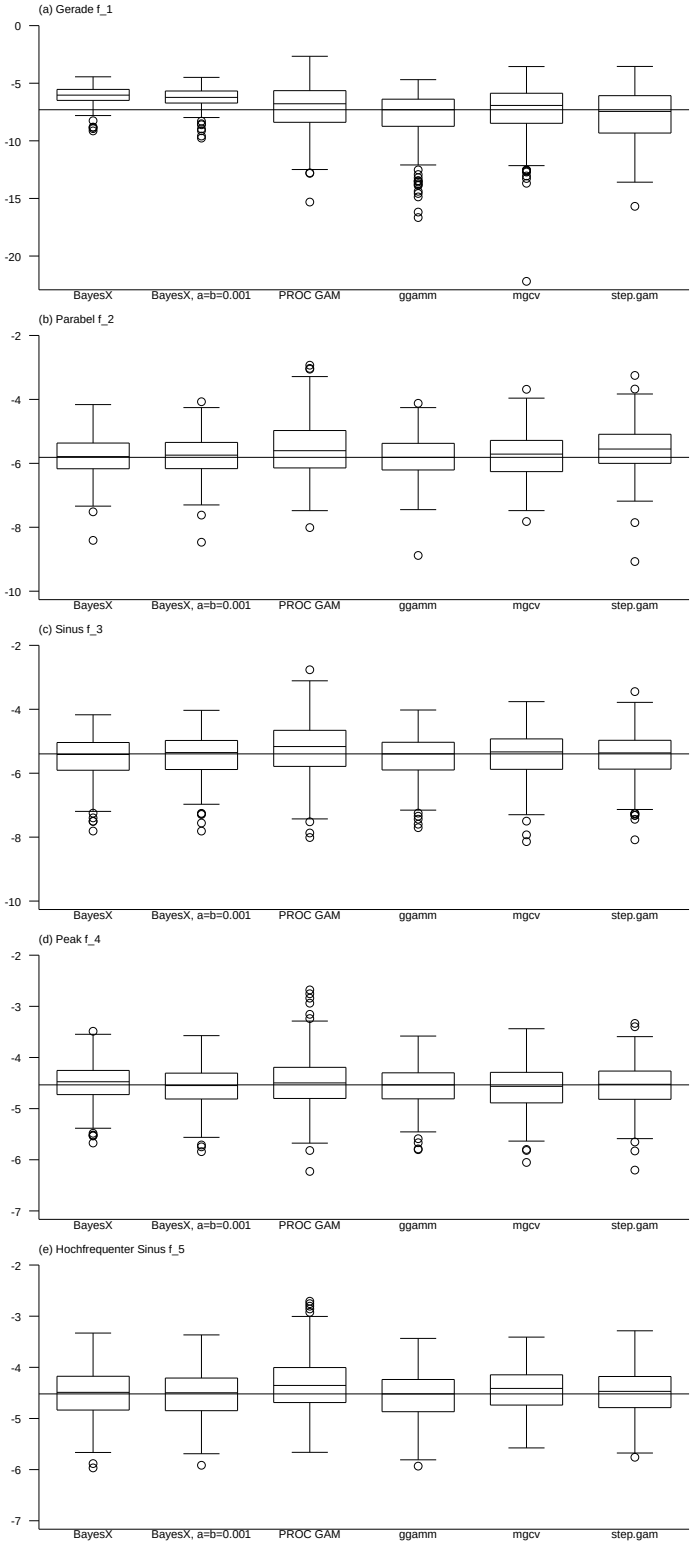

Abbildung 5.5: Poissonverteilter Response: Boxplots der logarithmierten MSEs.

Wie in Kapitel 3.4 beschrieben, wird bei Verwendung der Funktion `ggamm` die Schätzung sehr kleiner Varianzparameter, die in diesem Fall den inversen Glättungsparametern und damit bei Verwendung zweiter Differenzen als Penalisierung einer nahezu linearen Schätzung entsprechen, gestoppt, wenn das in (3.9) definierte Kriterium einen kleinen Wert unterschreitet. Im Rahmen der Simulationsstudie wurde hierfür der Wert 0.001 verwendet, was der Standardeinstellung aus `ggamm` entspricht. Für die Funktion f_1 wurde die Schätzung des inversen Glättungsparameters für die vier Response-Verteilungen zwischen 155 und 172 mal gestoppt, so dass man in diesen Fällen von einer tatsächlichen Schätzung als Gerade ausgehen kann. Man beachte auch, dass der entsprechende Anteil von 62% bis 68.8% relativ dicht an der in Kapitel 4.3 in Tabelle 4.4 berechneten Wahrscheinlichkeitsmasse der asymptotischen Verteilung des Restricted-Likelihood-Quotienten bei Verwendung zweiter Differenzen und 20 Knoten liegt. Obwohl also die in Kapitel 4 beschriebene Theorie nur bei normalverteiltem Response und einem Varianzparameter exakt anwendbar ist, ergibt sich auch in diesem allgemeineren Fall eine recht gute Übereinstimmung.

Für die übrigen Funktionen f_2 bis f_5 ergeben sich meist wesentlich geringere Unterschiede in der Schätzqualität. Auffällig ist beispielsweise noch, dass mit Hilfe von `BVCM` die Schätzung der Funktion f_4, zumindest im Normalverteilungsfall, wesentlich besser erfolgt als für alle übrigen Verfahren. Dies lässt sich aber leicht aus den unterschiedlichen Schätzansätzen erklären. Während durch `BVCM` sowohl Zahl als auch Positionen der verwendeten Knoten optimiert werden, es sich also um ein Verfahren zur adaptiven Knotenwahl handelt, basieren die übrigen Verfahren auf Penalisierungsansätzen und verwenden demzufolge eine feste Zahl äquidistanter Knoten. Während also bei der Schätzung mit `BVCM` mehr Knoten bei $x_4 = 0.5$ platziert werden können, wo die Funktion f_4 eine ausgeprägte Spitze besitzt, ist dies für die anderen Verfahren nicht möglich. Mit Hilfe von Verfahren zur adaptiven Knotenwahl lassen sich also offenbar starke lokale Schwankungen wesentlich besser erkennen als mit Penalisierungsansätzen. Überraschenderweise stimmt diese Aussage allerdings nicht mehr bei bernoulliverteiltem Response. Möglicherweise reicht hier die Information in den Daten nicht mehr aus, deutliche Hinweise auf die Spitze bei $x_4 = 0.5$ zu liefern.

Weiterhin lässt sich feststellen, dass sich die beiden Varianten der Hyperprioris des voll-bayesianischen Ansatzes mit `BayesX` in Bezug auf die resultierenden

MSEs kaum unterscheiden. In der Simulation zu generalisierten geoadditiven gemischten Modellen wird dies dagegen zumindest für einige Modellkomponenten recht deutlich der Fall sein.

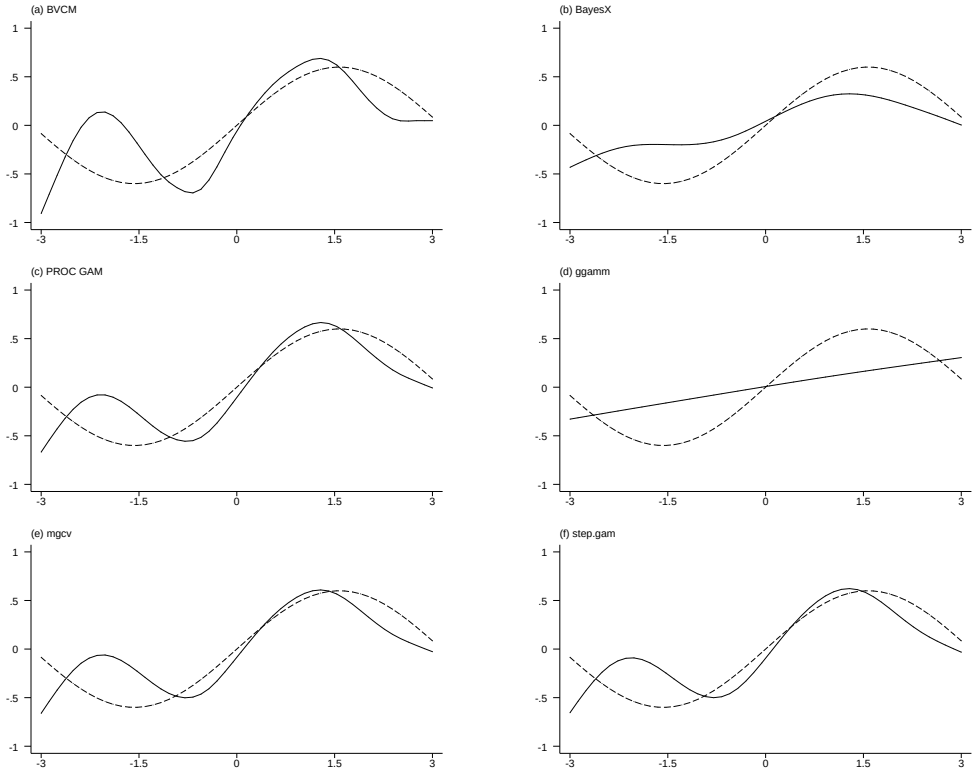

Abbildung 5.6: Bernoulliverteilter Response: Schätzungen der Funktion f_3 im 222. Modell für die verschiedenen Verfahren. Die wahre Funktion ist gestrichelt wiedergegeben.

Vergleicht man die mit Hilfe der Funktion `ggamm` erzielten Schätzungen mit den übrigen Verfahren, so liefert diese in fast allen Fällen zumindest Ergebnisse vergleichbarer Qualität. Besonders gut scheint die Schätzung mit `ggamm` für normalverteilten Response mit niedrigem Signal-Rauschen-Verhältnis und bei poissonverteiltem Response zu funktionieren. Die größten Probleme erhält man bei den Schätzungen der Funktionen f_3 und f_5 für bernoulliverteilten Response. Hier schneidet `ggamm` im Vergleich zu den übrigen Verfahren relativ schlecht ab. Diese Tatsache lässt sich erklären, wenn man berücksichtigt, dass f_3 in 59 Replikationen und f_5 in 48 Replikationen fälschlicherweise als linear geschätzt wurde. Für Funktion f_4 ist dies immerhin noch in 23 Replikationen der Fall. Die Schätzungen werden dabei als linear bezeichnet, wenn die Bestimmung der entsprechenden in-

versen Glättungsparameter bei einem kleinen Wert gestoppt wurde. Betrachtet man die übrigen Verteilungen des Response, so wurde in keinem Fall eine der Funktionen f_2 bis f_5 fälschlicherweise als linear eingestuft. Die häufige lineare Schätzung der Funktionen f_3 und f_5 besitzt nun nicht nur negative Auswirkungen auf den MSE, sondern wird sich auch bei Betrachtung des Bias der Schätzungen und der Überdeckungswahrscheinlichkeiten der Konfidenzbänder wieder bemerkbar machen.

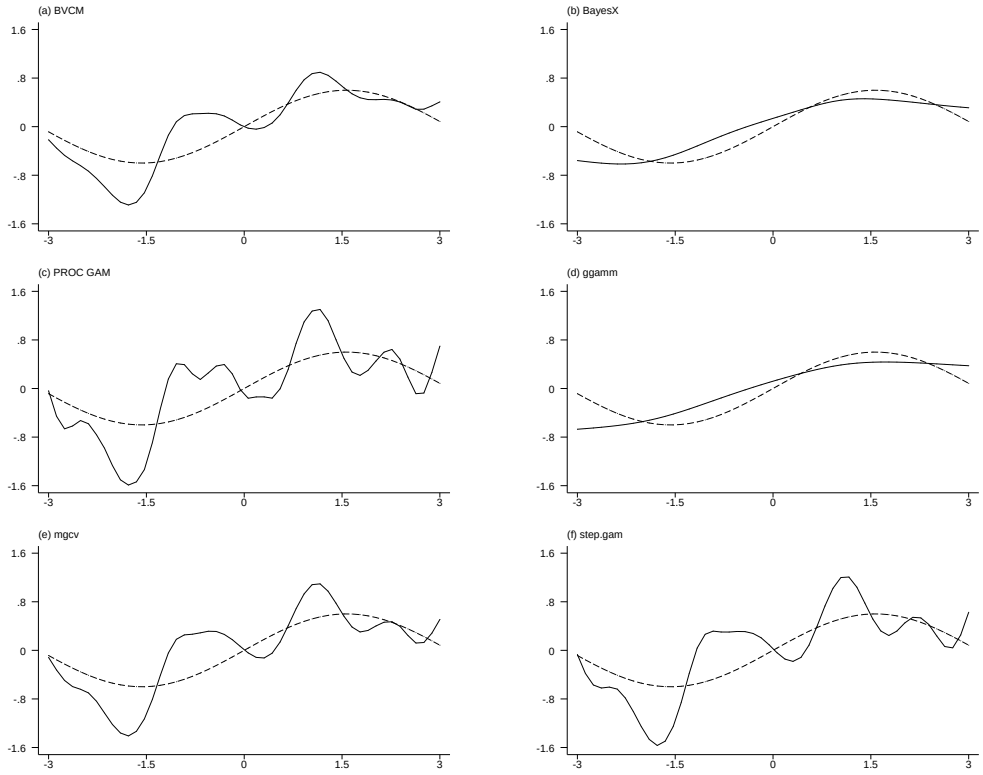

Abbildung 5.7: Bernoulliverteilter Response: Schätzungen der Funktion f_3 im 245. Modell für die verschiedenen Verfahren. Die wahre Funktion ist gestrichelt wiedergegeben.

Während also `ggamm` für bernoulliverteilten Response in einigen Fällen deutlich zu glatte Schätzungen liefert, fallen die Schätzungen der übrigen Verfahren, mit Ausnahme von `BayesX`, häufig deutlich zu rau aus. In keinem einzigen der betrachteten Fälle wiesen die Schätzungen der Funktion `ggamm` einen wesentlich raueren Verlauf als die der übrigen Verfahren auf. Obwohl einige Verfahren in Bezug auf den MSE also bessere Ergebnisse erzielen, mag eine glattere Schätzung dennoch als vorteilhafter empfunden werden. Insbesondere in der Analyse rea-

ler Datensätze zieht man in der Regel etwas glattere Schätzungen sehr rauen Schätzungen vor, weil diese sich wesentlich einfacher interpretieren lassen.

In den Abbildungen 5.6 und 5.7 sind beispielhaft die Schätzungen für f_3 mit den unterschiedlichen Verfahren aus zwei Replikationen visualisiert. Dabei wurde auf die Wiedergabe der Schätzungen, die sich bei Verwendung der alternativen Hyperparameter mit `BayesX` ergeben, verzichtet, weil die Unterschiede zu den Standardeinstellungen sehr gering ausfielen. In Abbildung 5.6 wird f_3 durch `ggamm` als Gerade geschätzt, während die übrigen Verfahren näher an der wahren Funktion gelegene Schätzungen ergeben. Damit erscheint zumindest plausibel, dass man für `ggamm` in diesem Fall einen deutlich größeren MSE zu erwarten hat. In Abbildung 5.7 sind dagegen die Schätzungen eines Modells wiedergegeben, in dem mit Ausnahme von `BayesX` und `ggamm` alle Verfahren deutlich zu raue Schätzungen liefern.

Auf der beiliegenden CD-Rom sind im Verzeichnis `simulation` weitere Grafiken zum unmittelbaren Vergleich der einzelnen Schätzungen enthalten. Für jede der 250 Replikationen einer Response-Verteilung und jede Funktion wurden die aus den sieben beziehungsweise sechs Verfahren resultierenden Schätzungen gegenüber gestellt und ermöglichen es so, sich einen direkten Eindruck von der Qualität und den Eigenschaften der einzelnen Verfahren zu machen.

Neben der unmittelbaren Betrachtung der MSEs über die Boxplots in den Abbildungen 5.2 bis 5.5 wurden in jeder Replikation für die MSEs Ränge unter den sechs beziehungsweise sieben Verfahren vergeben. Das heißt, für jedes geschätzte Modell wurden für jede Funktion die Ränge 1 bis 6 beziehungsweise 1 bis 7 auf die verschiedenen Verfahren nach dem jeweiligen MSE verteilt. In den Tabellen 5.1 bis 5.4 sind jeweils die Mittelwerte der resultierenden Ränge wiedergegeben. Diese ermöglichen einen Vergleich der relativen Schätzgenauigkeit, sollten aber nicht unabhängig von den tatsächlichen MSEs betrachtet werden. Niedrige mittlere Ränge müssen nämlich nicht zwangsläufig eine wesentlich bessere Schätzung bedeuten, sondern zeigen lediglich an, dass die Schätzung mit einem Verfahren in vielen Fällen besser als mit den übrigen Verfahren erfolgt. Wie groß der Unterschied in der Schätzqualität ist, wird jedoch nicht berücksichtigt. Dennoch bieten die Ränge eine interessante Zusatzinformation, da sie auf einem unmittelbaren Vergleich der Schätzungen eines Modells beruhen. Beispielsweise kann man aus Tabelle 5.1 ablesen, dass `BVCM` mit einem mittleren Rang von 1.316 für nahezu

	f_1	f_2	f_3	f_4	f_5	ar. Mittel
BVCM	4.568	4.440	3.936	1.316	4.672	3.786
BayesX	6.672	4.168	4.000	5.224	3.532	4.719
BayesX, $a = b = 0.001$	5.528	4.344	4.884	4.752	4.464	4.794
ggamm	2.304	3.524	3.960	4.364	3.928	3.616
mgcv	3.320	3.740	3.732	3.628	3.908	3.666
PROC GAM	3.372	4.360	4.416	4.496	5.404	4.410
step.gam	2.236	3.424	3.072	4.220	2.092	3.009

Tabelle 5.1: Normalverteilter Response: Mittlere Ränge der MSEs bei hohem Signal-Rauschen-Verhältnis ($\sigma^2 = 0.09$).

	f_1	f_2	f_3	f_4	f_5	ar. Mittel
BVCM	5.112	5.000	5.584	1.348	5.824	4.574
BayesX	6.324	4.020	3.288	5.644	2.864	4.428
BayesX, $a = b = 0.001$	4.96	4.196	4.520	4.384	3.760	4.364
ggamm	2.204	3.240	3.564	4.500	3.084	3.318
mgcv	3.172	3.928	3.756	3.620	4.376	3.770
PROC GAM	3.532	4.728	4.160	4.348	5.244	4.402
step.gam	2.696	2.888	3.128	4.156	2.848	3.143

Tabelle 5.2: Normalverteilter Response: Mittlere Ränge der MSEs bei niedrigem Signal-Rauschen-Verhältnis ($\sigma^2 = 0.36$).

	f_1	f_2	f_3	f_4	f_5	ar. Mittel
BayesX	5.364	3.300	2.968	4.468	3.440	3.908
BayesX, $a = b = 0.001$	4.424	3.748	4.012	3.300	3.240	3.745
ggamm	2.428	2.720	3.260	3.648	2.412	2.894
mgcv	2.828	2.944	3.548	2.548	4.152	3.204
PROC GAM	3.328	3.704	4.148	3.456	4.624	3.852
step.gam	2.628	4.584	3.064	3.580	3.132	3.398

Tabelle 5.3: Bernoulliverteilter Response: Mittlere Ränge der MSEs.

	f_1	f_2	f_3	f_4	f_5	ar. Mittel
BVCM	6.820	6.176	5.524	4.264	3.952	5.347
BayesX	4.576	2.440	3.296	3.548	5.504	3.873
BayesX, $a = b = 0.001$	4.484	3.776	3.100	3.328	4.036	3.745
ggamm	2.376	2.98	4.044	3.416	4.436	3.450
mgcv	3.108	3.996	3.528	3.856	3.384	3.574
PROC GAM	3.372	4.484	4.252	4.688	3.696	4.098
step.gam	3.264	4.148	4.256	4.900	2.992	3.912

Tabelle 5.4: Poissonverteilter Response: Mittlere Ränge der MSEs.

alle Fälle die beste Schätzung der Funktion f_4 lieferte. Für die Gerade f_1 erkennt man ebenfalls wieder, dass die jeweils besten Schätzungen für normalverteilten und poissonverteilten Response durch `ggamm` und `step.gam`, für bernoulliverteilten Response durch `ggamm` erfolgen.

In der letzten Spalte der Tabellen ist zusätzlich der Mittelwert der fünf Spalten der mittleren Ränge angegeben. Dies ermöglicht theoretisch die Auswahl eines ‚optimalen' Verfahrens. Dabei ist jedoch zu beachten, dass dies nur für das hier untersuchte Modell mit den fünf Funktionen f_1 bis f_5 gelten kann. Je nach untersuchter Fragestellung kann dann auch ein weniger ‚optimales' Verfahren vorzuziehen sein. Beispielsweise erhielte man bei alleiniger Betrachtung des arithmetischen Mittels die Funktion `ggamm` als optimales Verfahren bei bernoulliverteiltem Response. Diese Einschätzung beruht aber im wesentlichen darauf, dass `ggamm` die Funktion f_1 am besten schätzt, während für alle übrigen Funktionen die mittleren Ränge sehr dicht zusammen liegen. Bei Betrachtung rauerer Funktionen hätte man also sicher ein anderes Ergebnis für das optimale Verfahren erhalten.

Zusätzlich zum Vergleich der MSEs sollen nun für die mit Hilfe der Funktion `ggamm` erfolgten Schätzungen noch einige weitere Betrachtungen angestellt werden. Dazu sind zunächst in den Abbildungen 5.8 bis 5.11 die mittleren Funktionsschätzungen für die verschiedenen Response-Verteilungen zusammen mit den wahren Funktionen wiedergegeben. Aus diesen Abbildungen lässt sich also auch der Bias der entsprechenden Schätzungen ablesen. Wie man sieht, erfolgt die Schätzung für normalverteilten Response sowohl bei hohem als auch bei niedrigem Signal-Rauschen-Verhältnis nahezu unverzerrt. Lediglich bei Funktion f_4 wird die Spitze bei $x_4 = 0.5$ leicht unterschätzt, wobei die Verzerrung für niedriges

Signal-Rauschen-Verhältnis etwas größer ausfällt.

Für bernoulliverteilten Response erhält man recht deutliche Verzerrungen. Während diese für die lineare Funktion f_1 und die Parabel f_2 noch relativ gering sind, fallen die mittleren Schätzungen für die übrigen Funktionen deutlich zu glatt aus. Man beachte hierbei wieder, dass diese Funktionen für bernoulliverteilten Response in relativ vielen Fällen als linear geschätzt werden und somit ein großer Teil der Verzerrung aus dieser Tatsache resultieren dürfte.

Für poissonverteilten Response erhält man wieder nahezu unverzerrte Schätzungen. Die einzige Ausnahme bildet Funktion f_4, für die die mittlere Funktionsschätzung zu glatt ausfällt. Die Spitze bei $x_4 = 0.5$ wird also in vielen Schätzungen nicht ausreichend wiedergegeben.

Neben Punktschätzungen der Funktionen interessieren in der Regel auch Konfidenzbänder, die einen Sicherheitsbereich für die Funktionsschätzungen wiedergeben. Tabelle 5.5 enthält die mittleren Überdeckungswahrscheinlichkeiten der Konfidenzbänder zu einem punktweisen Sicherheitsgrad von 95%. Man vergleiche hierzu Kapitel 3.5. Unterschieden werden frequentistische und bayesianische Konfidenzbänder.

Verteilung		f_1	f_2	f_3	f_4	f_5
Normal	frequentistisch	0.939	0.957	0.948	0.929	0.966
$\sigma^2 = 0.09$	bayesianisch	0.959	0.98	0.974	0.965	0.990
Normal	frequentistisch	0.940	0.946	0.949	0.894	0.961
$\sigma^2 = 0.36$	bayesianisch	0.966	0.975	0.974	0.932	0.984
Bernoulli	frequentistisch	0.940	0.904	0.692	0.795	0.636
	bayesianisch	0.968	0.939	0.737	0.836	0.678
Poisson	frequentistisch	0.927	0.942	0.950	0.868	0.947
	bayesianisch	0.948	0.971	0.974	0.910	0.976

Tabelle 5.5: Mittlere Überdeckungswahrscheinlichkeiten (nominaler Sicherheitsgrad: 95%).

Offenbar halten die frequentistischen Konfidenzbänder im Normalverteilungsfall den vorgegebenen Sicherheitsgrad im Mittel recht gut ein, während die bayesianischen Gegenstücke eher zu konservativ sind. Eine Ausnahme bildet Funktion f_4. Hier liegen die bayesianischen Konfidenzintervalle näher am vorgegebenen Sicherheitsgrad. Für bernoulliverteilten Response erhält man dagegen für beide Varianten teilweise deutliche Abweichungen. Während diese für die Funktionen f_1 und

f_2 noch relativ akzeptabel ausfallen und für die bayesianischen Konfidenzbänder sogar praktisch nicht vorhanden sind, fallen sie für die übrigen Funktionen gravierender aus. Dies lässt sich wieder aus der Tatsache erklären, dass diese Funktionen in relativ vielen Fällen fälschlicherweise als linear geschätzt werden. In diesen Fällen erhält man natürlich auch extrem fehlerhafte Konfidenzbänder. Generell erscheint bei bernoulliverteiltem Response die Verwendung der bayesianischen Konfidenzbänder sinnvoller.

Für poissonverteilten Response erhält man wieder dichter am vorgegebenen Sicherheitsgrad liegende mittlere Überdeckungswahrscheinlichkeiten. Auch hier scheinen aber die bayesianischen Konfidenzbänder die besseren Eigenschaften zu besitzen. Die Probleme bei Funktion f_4 resultieren weitgehend aus der zu starken Glättung der Funktion, die auch aus den mittleren Funktionsschätzungen ersichtlich war.

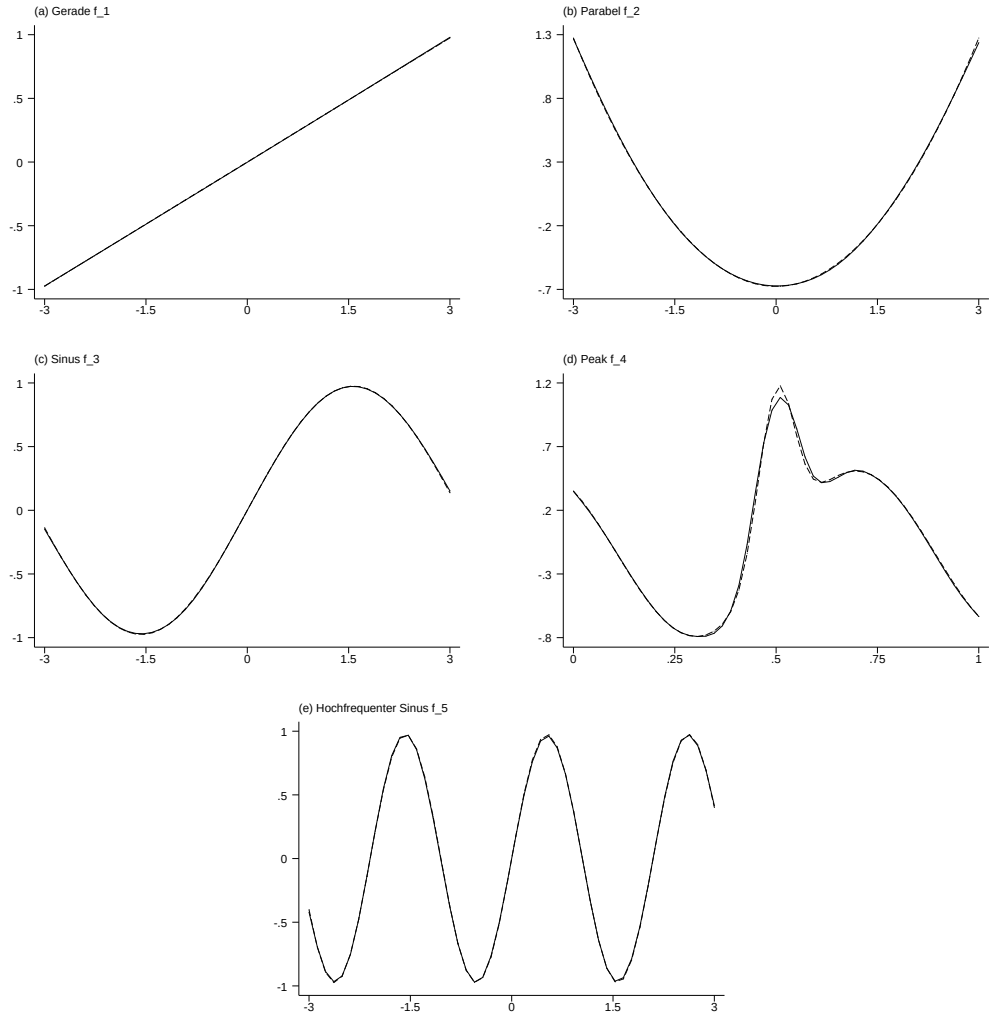

Abbildung 5.8: Normalverteilter Response: Mittlere Funktionsschätzungen bei hohem Signal-Rauschen-Verhältnis ($\sigma^2 = 0.09$). Die wahren Funktionen sind gestrichelt wiedergegeben.

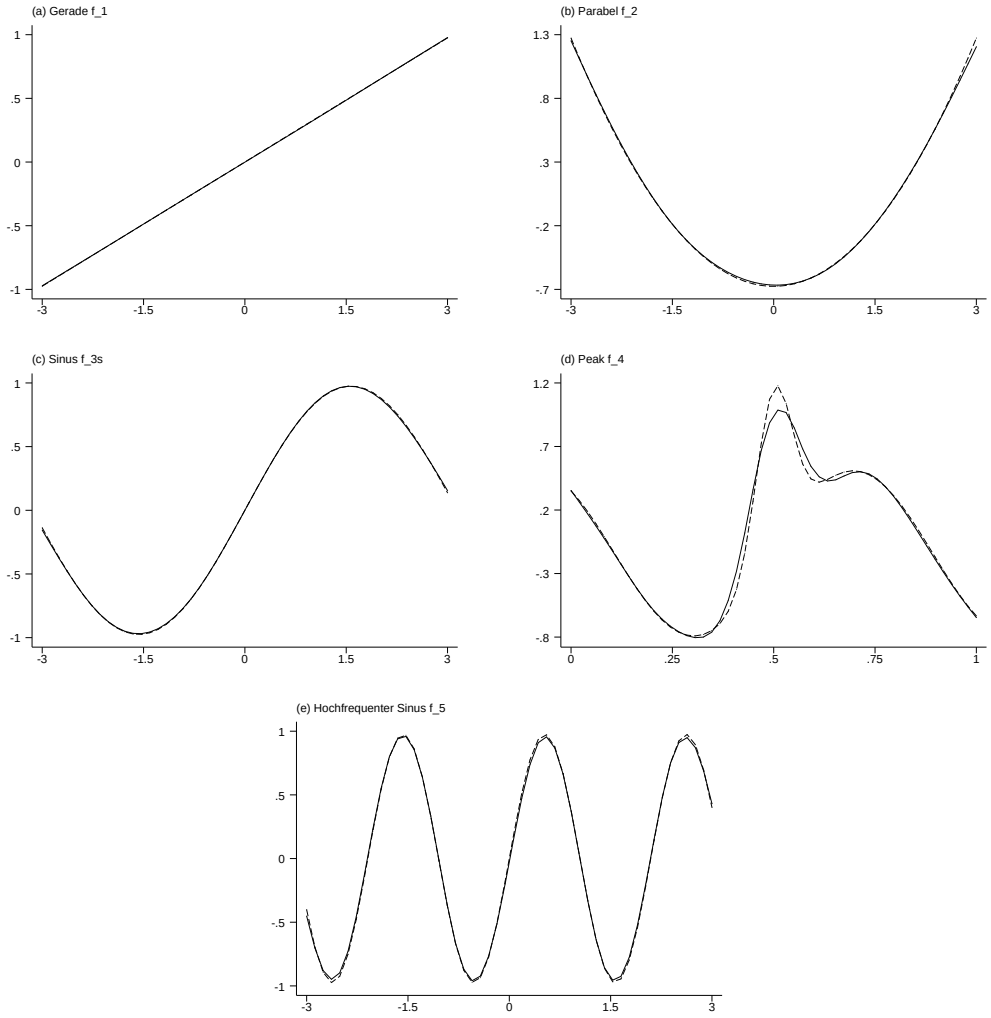

Abbildung 5.9: Normalverteilter Response: Mittlere Funktionsschätzungen bei niedrigem Signal-Rauschen-Verhältnis ($\sigma^2 = 0.36$). Die wahren Funktionen sind gestrichelt wiedergegeben.

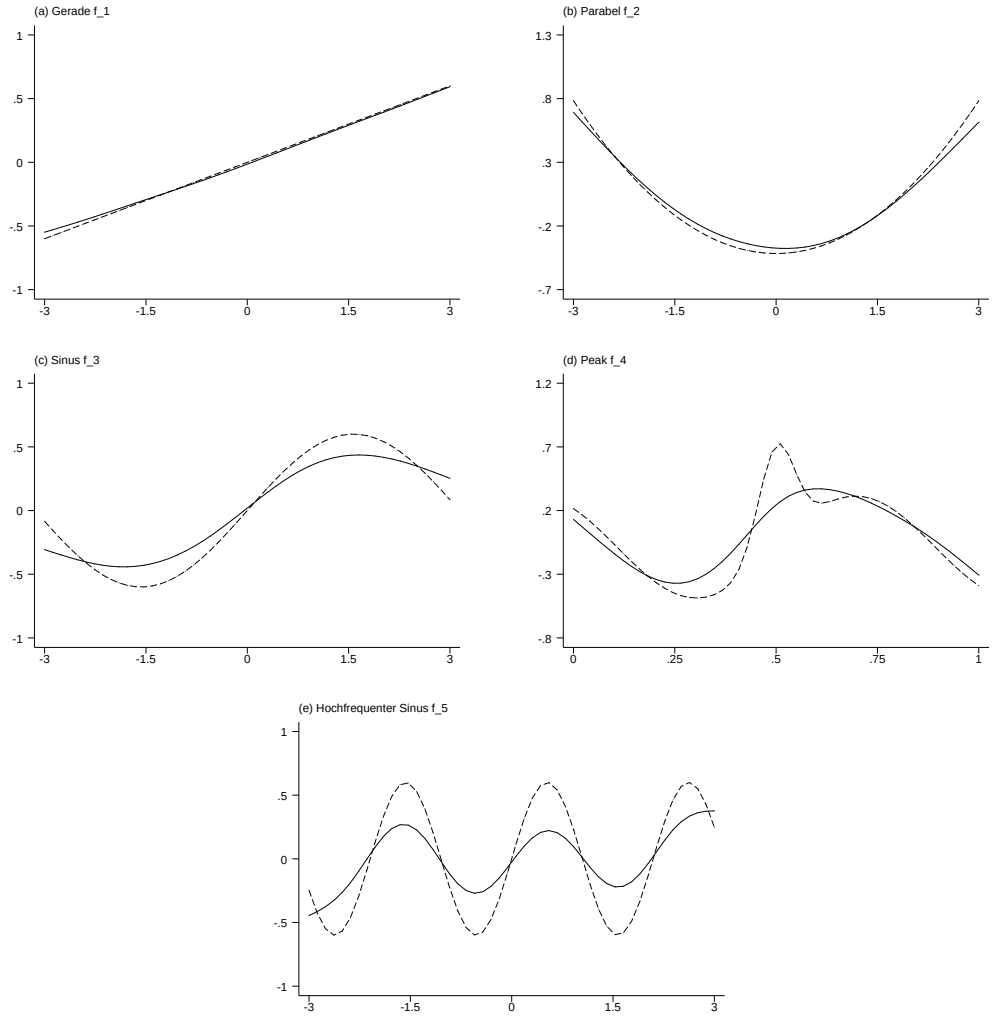

Abbildung 5.10: Bernoulliverteilter Response: Mittlere Funktionsschätzungen. Die wahren Funktionen sind gestrichelt wiedergegeben.

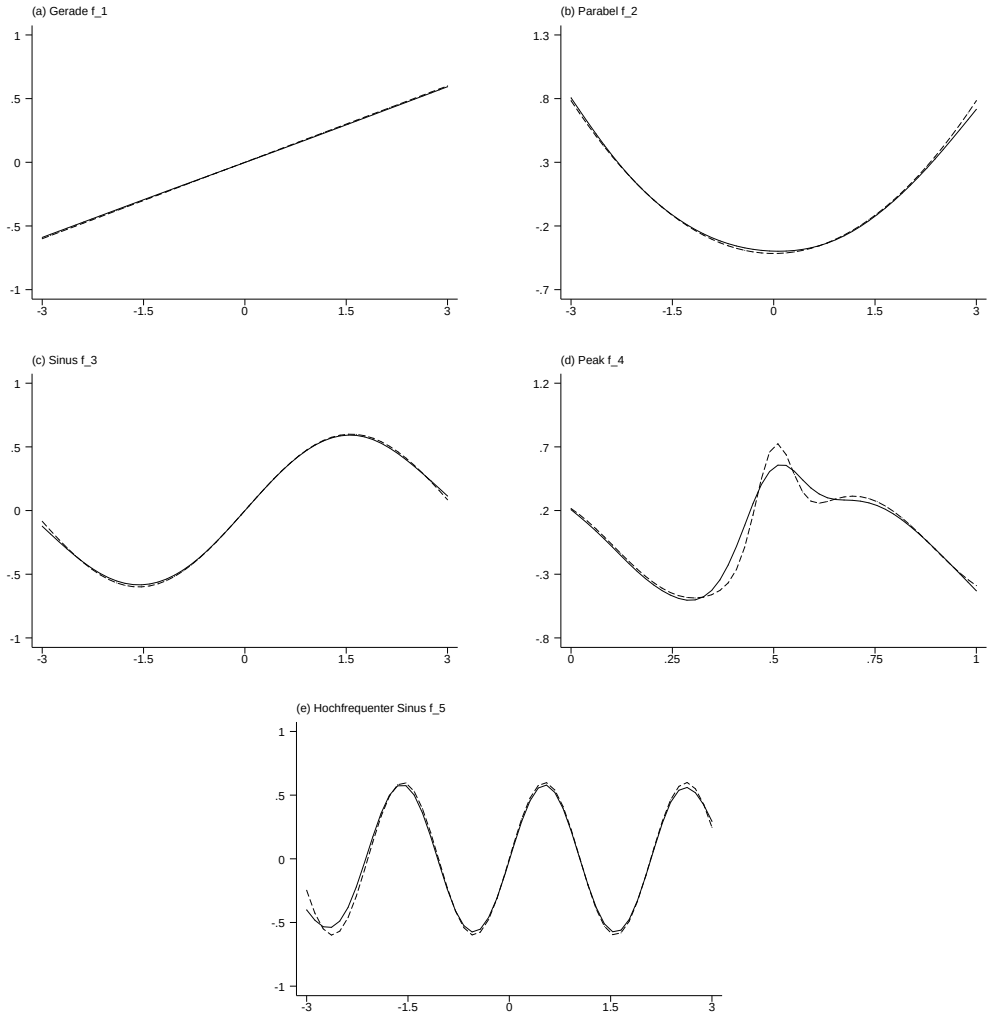

Abbildung 5.11: Poissonverteilter Response: Mittlere Funktionsschätzungen. Die wahren Funktionen sind gestrichelt wiedergegeben.

5.2 Generalisierte geoadditive gemischte Modelle

5.2.1 Modell

Um die Qualität der beschriebenen Schätzverfahren in generalisierten geoadditiven gemischten Modellen zu beurteilen, wurden Longitudinaldaten erzeugt, wobei jede der 24 vorhandenen Gruppen aus 31 Beobachtungen besteht, so dass man als Gesamtstichprobenumfang 744 Beobachtungen erhält. Dem simulierten Modell liegt dann der additive Prädiktor

$$\eta_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + f_1(x_{ij3}) + f_{spat}(R_{ij}) + b_{i0} + x_{ij1}b_{i1} + x_{ij2}b_{i2}$$

mit $i = 1, \ldots, 24$ und $j = 1, \ldots, 31$ zugrunde. Mit f_1 wird dabei die in Abbildung 5.12 (a) wiedergegebene glatte Funktion bezeichnet. Die Ausprägungen der Kovariablen x_{ij3} stammen aus einem äquidistanten Gitter von 186 Werten aus dem Intervall $[-3, 3]$. Jeder Wert des Gitters wurde insgesamt vier mal zufällig den Beobachtungen zugewiesen, woraus sich wieder eine Gesamtzahl von 744 Beobachtungen ergibt.

Die räumliche Funktion f_{spat} ist in Abbildung 5.12 (b) wiedergegeben. Bei den Regionen R_{ij} handelt es sich um die 124 Kreise Bayerns und Baden-Württembergs, die jeweils sechs mal zufällig auf die Beobachtungen aufgeteilt wurden.

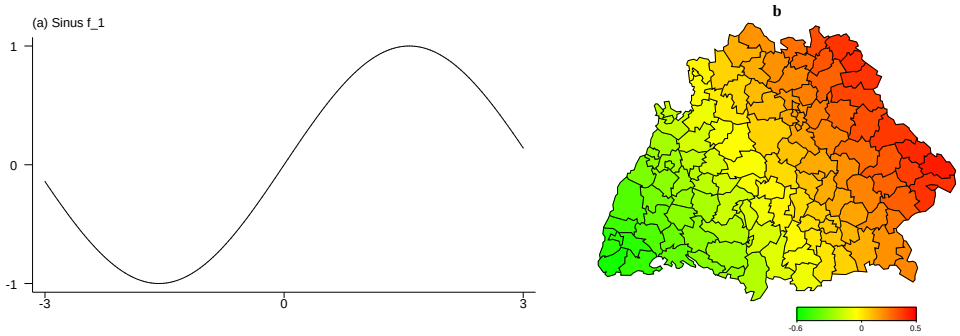

Abbildung 5.12: Wahre Funktionen der Simulation zu generalisierten geoadditiven gemischten Modellen.

Die zufälligen Effekte b_{i0}, b_{i1} und b_{i2}, $i = 1, \ldots, 24$ wurden als unabhängig und identisch verteilt gemäß den folgenden Annahmen realisiert:

$$b_{i0} \sim \mathrm{N}(0, 0.25),\ b_{i1} \sim \mathrm{N}(0, 0.25),\ b_{i2} \sim \mathrm{N}(0, 0.36).$$

Durch b_{i0} erhält man einen zufälligen Intercept, während b_{i1} und b_{i2} gruppenspezifische, zufällige Steigungen zu den Kovariablen x_{ij1} und x_{ij2} sind. Zusätzlich zu diesen zufälligen Effekten sind auch die fixen Haupteffekte durch β_1 und β_2 im Modell berücksichtigt. Für beide wurde der Wert $\beta_1 = \beta_2 = 0.25$ in der Simulation zugrunde gelegt. Die Kovariablen x_{ij1} und x_{ij2} stammen jeweils aus einem aus 186 Werten bestehenden, äquidistanten Gitter aus dem Intervall $[-1, 1]$, wobei wieder jeder Wert zufällig vier mal auf die Beobachtungen aufgeteilt wurde.

Untersucht wurden vier verschiedene Verteilungen der abhängigen Variablen: Normalverteilter Response mit Varianz $\sigma^2 = 0.25$, bernoulliverteilter Response, binomialverteilter Response mit jeweils drei bernoulliverteilten Versuchswiederholungen sowie poissonverteilter Response. Der binomialverteilte Response wurde zusätzlich zum bernoulliverteilten Response betrachtet, um abschätzen zu können, wie schnell die, für bernoulliverteilten Response verhältnismäßig große Verzerrung geringer wird, wenn mehr Information in den Daten vorhanden ist. Für jede Verteilung wurden wieder jeweils 250 Replikationen geschätzt.

Zur Analyse generalisierter geoadditiver gemischter Modelle stehen nur zwei der vorgestellten Programme zur Verfügung, nämlich `BayesX` und `ggamm`. Wie zuvor werden auch hier für `BayesX` zwei verschiedene Spezifikationen der Priori-Verteilungen der Varianzparameter betrachtet. Die glatte Funktion f_1 wird sowohl mit Hilfe der Funktion `ggamm` als auch mit `BayesX` als P-Spline vom Grad 3 mit 20 Knoten und Differenzen der Ordnung $k = 2$ als Penalisierung geschätzt. Für die räumliche Funktion wird ein Markov-Zufallsfeld angenommen, wobei zwei Regionen als benachbart betrachtet werden, wenn sie gemeinsame Grenzen besitzen.

5.2.2 Ergebnisse

Zunächst ist zu beachten, dass es bei einem Teil der Schätzungen mit Hilfe der Funktion `ggamm` zu Konvergenzproblemen kam, die sich aus Tabelle 5.6 ablesen lassen. Diese gibt für die verschiedenen Verteilungen des Response einige Kennziffern zu den zur Schätzung benötigten Iterationen wieder. Aus der letzten Spalte ist dabei abzulesen, in wie vielen Fällen keine Konvergenz erzielt werden konnte. Eine genauere Betrachtung ergab, dass meist lediglich ein Glättungsparameter nicht eindeutig geschätzt werden konnte und dann zwischen zwei relativ ähnlichen

Werten wechselte ohne zu konvergieren. Für die übrigen Glättungs- und Varianzparameter konvergierten die Schätzungen dagegen in der Regel bereits nach einer relativ geringen Zahl von Iterationen gegen einen festen Grenzwert. Ein direkter Vergleich der Schätzungen mit und ohne Konvergenz ergab nur geringe Unterschiede. Daher wurden in den Fällen, in denen keine Konvergenz erzielt werden konnte, die Ergebnisse der letzten Iteration als Schätzungen verwendet.

Verteilung	Min.	25%-Qu.	ar. Mit.	Med.	75%-Qu.	Max.	keine Konv.
Normal	6	13	140.56	36.5	400	400	68
Bernoulli	9	13.25	125	26	227.75	400	61
Binomial	9	16	149.67	30.5	400	400	76
Poisson	11	19	119.13	31	157.5	400	53

Tabelle 5.6: Minimum, 25%-Quantil, arithmetisches Mittel, Median, 75%-Quantil und Maximum der zur Schätzung mit ggamm benötigten Iterationen. In der letzten Spalte ist angegeben, wie oft keine Konvergenz erzielt werden konnte.

Wie zu den Schätzungen der generalisierten additiven Modelle in Kapitel 5.1 befinden sich auch für die Schätzungen der generalisierten geoadditiven gemischten Modelle auf der beiliegenden CD-Rom Grafiken, in denen die Schätzungen der verschiedenen Verfahren für alle Replikationen visualisiert werden. Im Verzeichnis `simulation` sind für jede der vier Response-Verteilungen die Schätzungen aller 250 Replikationen in jeweils einer Datei gespeichert.

Um die unterschiedlichen Ansätze vergleichen zu können, wird wieder der mittlere quadratische Fehler betrachtet, der für f_1 und f_{spat} wie in (5.2) definiert ist. Für die zufälligen Effekte wurden ebenfalls MSEs berechnet und zwar nach der Formel

$$MSE(b_j) = \frac{1}{24}\sum_{i=1}^{24}(\hat{b}_{ij} - b_{ij})^2, \qquad j = 0, 1, 2.$$

In den Abbildungen 5.13 bis 5.16 sind die logarithmierten MSEs der einzelnen Modellkomponenten für die vier verschiedenen Verteilungen des Response als Boxplots wiedergegeben. Wie man sieht, ergeben sich für normalverteilten, binomialverteilten und poissonverteilten Response keine wesentlichen Unterschiede in Bezug auf den mittleren quadratischen Fehler. Für bernoulliverteilten Response gilt dies zwar für die Funktion f_1 und den zufälligen Intercept b_0, jedoch nicht für die räumliche Funktion f_{spat} und die zufälligen Steigungen b_1 und b_2. Diese werden am besten durch die Funktion `ggamm` geschätzt, während `BayesX` mit den Standardeinstellungen für die Priori-Verteilungen der Varianzparameter am schlechtesten

abschneidet. Durch die Verwendung der alternativen Priori-Verteilung mit Hyperparametern $a = b = 0.001$ ergibt sich zwar eine Verbesserung, die aber nicht ganz an die Schätzungen durch `ggamm` heranreicht.

Zusätzlich wurden wieder für jede der 250 Replikationen der einzelnen Modelle Ränge bezüglich der MSEs unter den drei Verfahren vergeben. In den Tabellen 5.7 bis 5.10 sind die Mittelwerte dieser Ränge wiedergegeben. Zur Interpretation sind dabei die gleichen Anmerkungen zu beachten, die in Kapitel 5.1 gemacht wurden. Aus der Betrachtung der mittleren Ränge ergeben sich nämlich teilweise wesentlich deutlichere Abgrenzungen zwischen den einzelnen Verfahren. So liefern offenbar die Standardeinstellungen von `BayesX` in nahezu allen Replikationen des Modells mit normalverteiltem Response die besten Schätzungen für den räumlichen Effekt. Betrachtet man dagegen die entsprechenden Boxplots, so erkennt man zwar auch hier die Standardeinstellungen von `BayesX` als bestes Verfahren, der Unterschied zu den beiden übrigen Verfahren fällt aber eher gering aus. Umgekehrt werden die Unterschiede, die bei bernoulliverteiltem Response bestehen, durch die mittleren Ränge weniger stark wiedergegeben als durch die entsprechenden Boxplots. Es empfiehlt sich also stets die beiden Betrachtungen zu kombinieren.

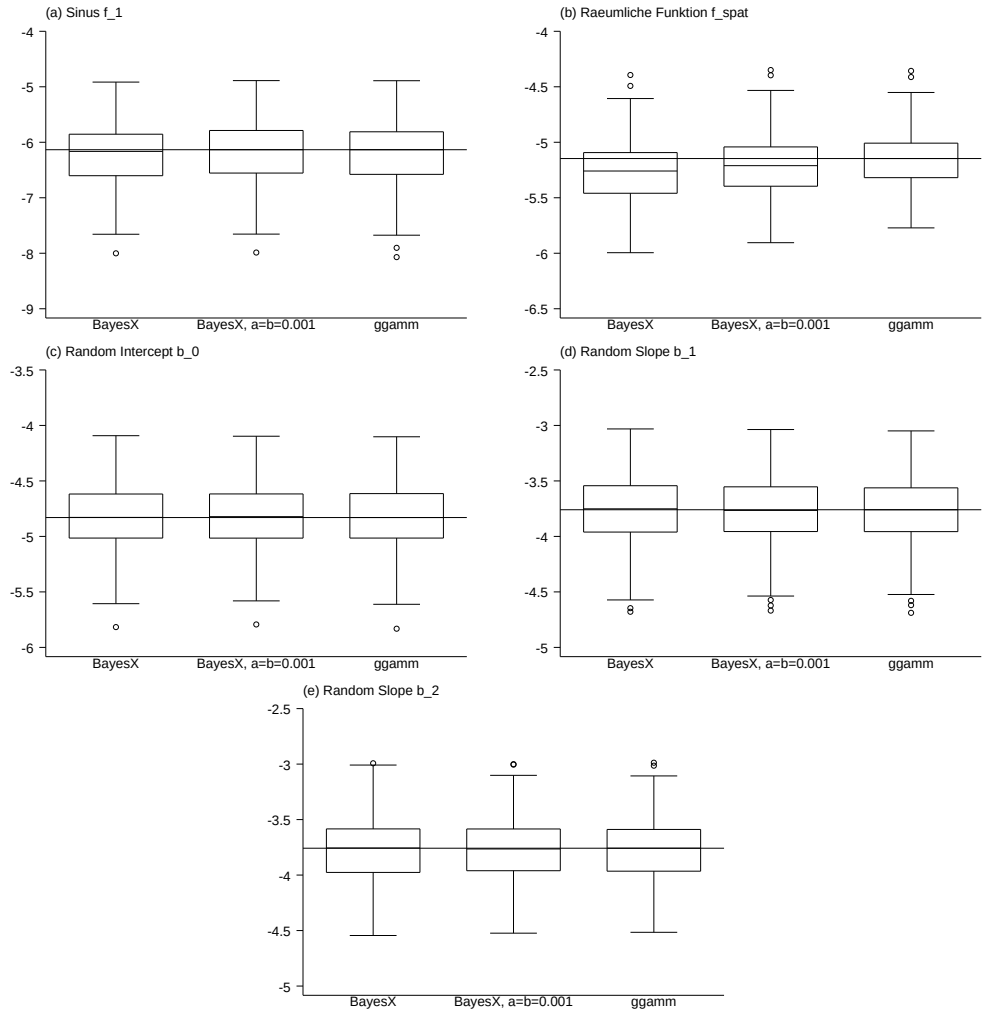

Abbildung 5.13: Normalverteilter Response: Boxplots der logarithmierten MSEs.

	f	f_{spat}	b_0	b_1	b_2	ar. Mittel
BayesX	1.456	1.100	1.892	2.204	1.864	1.703
BayesX, $a = b = 0.001$	2.548	2.552	2.052	1.916	2.052	2.224
ggamm	1.996	2.348	2.056	1.880	2.084	2.073

Tabelle 5.7: Normalverteilter Response: Mittlere Ränge der MSEs.

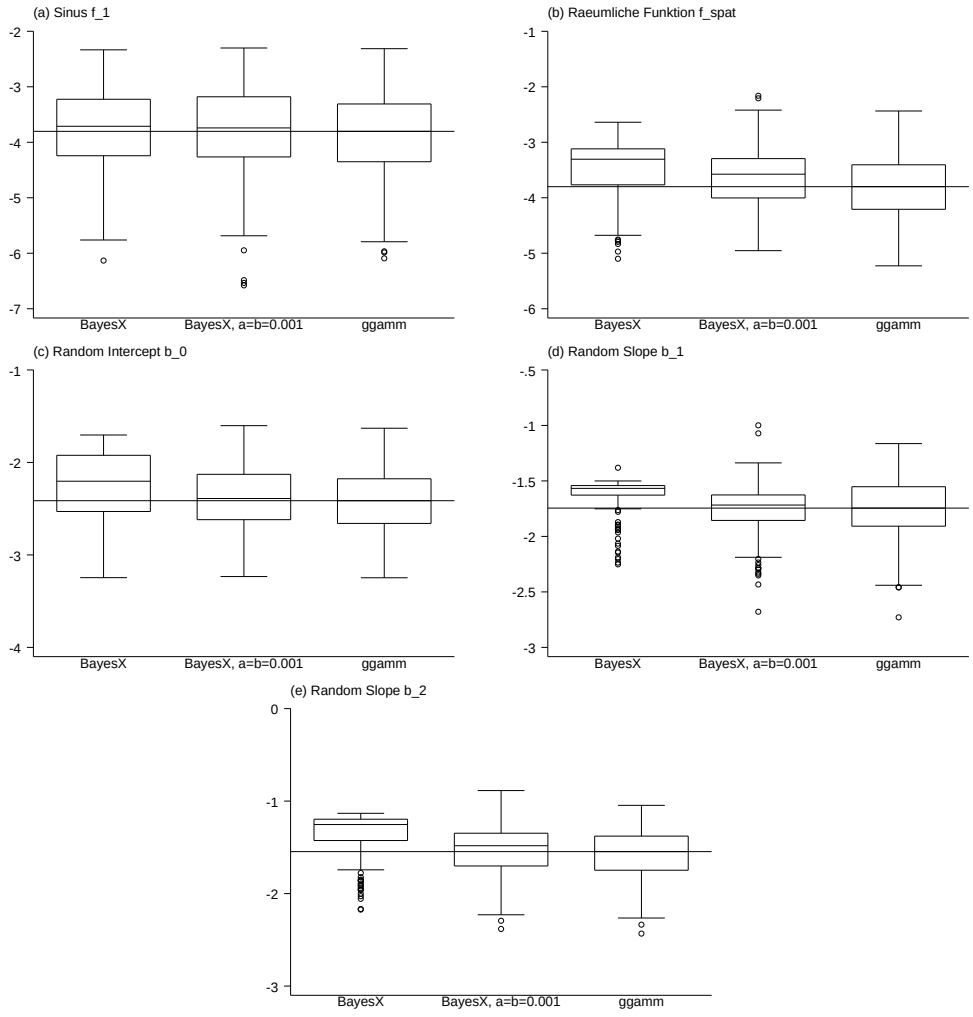

Abbildung 5.14: Bernoulliverteilter Response: Boxplots der logarithmierten MSEs.

	f	f_{spat}	b_0	b_1	b_2	ar. Mittel
BayesX	2.228	2.464	2.604	2.632	2.708	2.527
BayesX, $a = b = 0.001$	2.084	2.028	1.964	1.576	1.896	1.910
ggamm	1.688	1.508	1.432	1.792	1.396	1.563

Tabelle 5.8: Bernoulliverteilter Response: Mittlere Ränge der MSEs.

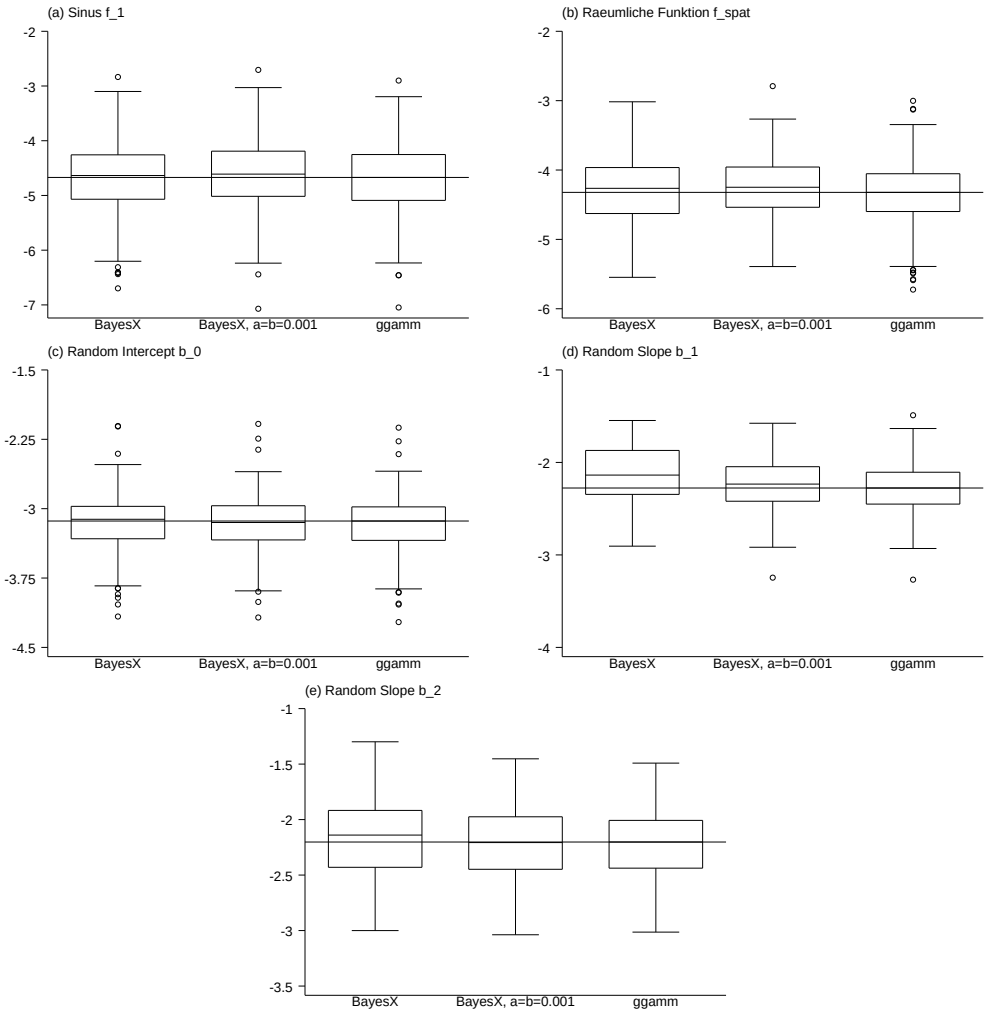

Abbildung 5.15: Binomialverteilter Response: Boxplots der logarithmierten MSEs.

	f	f_{spat}	b_0	b_1	b_2	ar. Mittel
BayesX	1.8	1.948	2.280	2.500	2.244	2.154
BayesX, $a = b = 0.001$	2.264	2.428	1.904	2.076	2.020	2.138
ggamm	1.936	1.624	1.816	1.424	1.736	1.707

Tabelle 5.9: Binomialverteilter Response: Mittlere Ränge der MSEs.

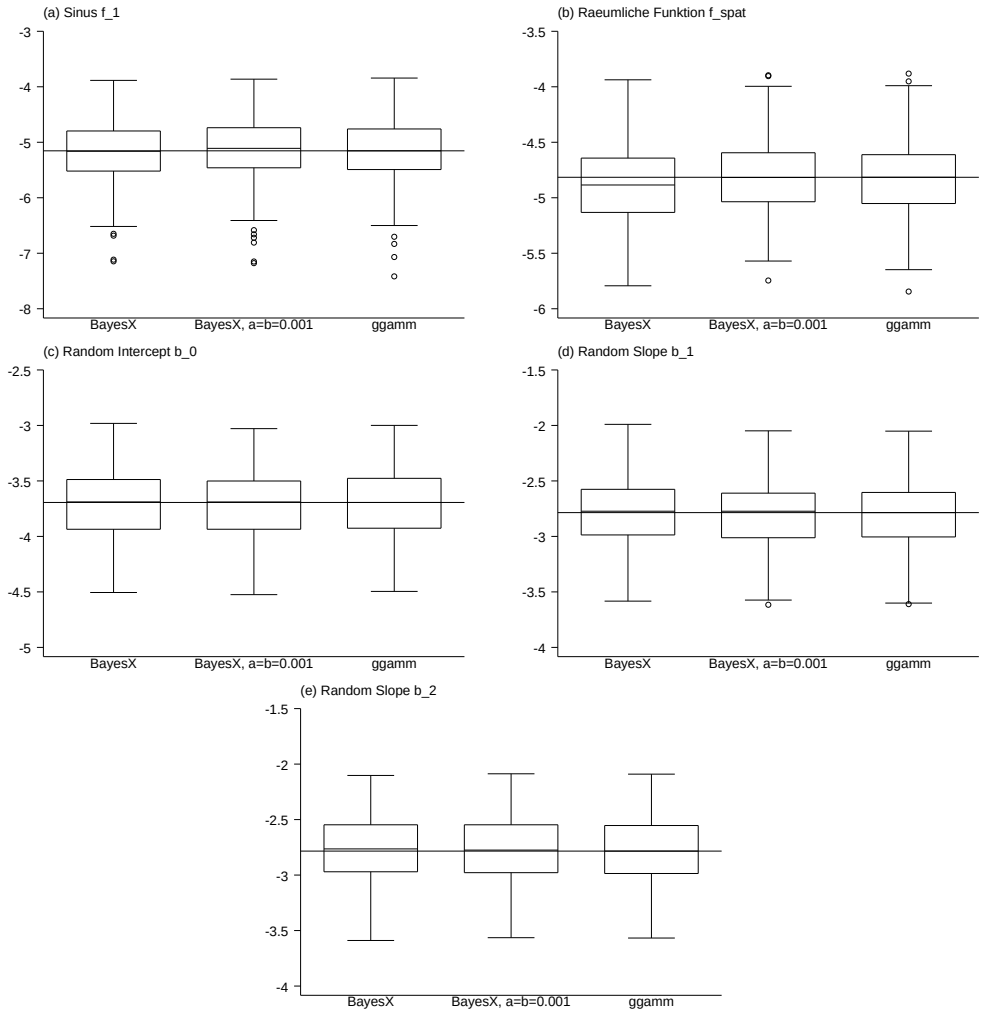

Abbildung 5.16: Poissonverteilter Response: Boxplots der logarithmierten MSEs.

	f	f_{spat}	b_0	b_1	b_2	ar. Mittel
BayesX	1.624	1.368	2.016	2.444	2.172	1.925
BayesX, $a = b = 0.001$	2.424	2.620	1.784	1.856	2.112	2.159
ggamm	1.952	2.012	2.200	1.700	1.716	1.916

Tabelle 5.10: Poissonverteilter Response: Mittlere Ränge der MSEs.

		emp. Wert	Bias			
			Normal	Bernoulli	Binomial	Poisson
BayesX	$\sigma^2_{b_0}$	0.196 (0.25)	0.009	-0.066	0.002	-0.001
	$\sigma^2_{b_1}$	0.226 (0.25)	0.001	-0.177	-0.070	-0.019
	$\sigma^2_{b_2}$	0.329 (0.36)	0.013	-0.215	-0.032	-0.004
BayesX	$\sigma^2_{b_0}$	0.196 (0.25)	0.030	0.024	0.039	0.026
	$\sigma^2_{b_1}$	0.226 (0.25)	0.028	-0.019	0.014	0.020
$(a = b = 0.001)$	$\sigma^2_{b_2}$	0.329 (0.36)	0.051	0.024	0.057	0.048
ggamm	$\sigma^2_{b_0}$	0.196 (0.25)	0.010	-0.014	0.003	-0.005
	$\sigma^2_{b_1}$	0.226 (0.25)	0.006	-0.047	-0.014	-0.006
	$\sigma^2_{b_2}$	0.329 (0.36)	0.017	-0.029	-0.003	0.007

Tabelle 5.11: Bias der Schätzungen der Varianzparameter. In der dritten Spalte ist jeweils der empirische Wert, in Klammern dahinter der theoretische Wert angegeben. Zur Berechnung des Bias wurden die empirischen Werte zugrunde gelegt.

In generalisierten geoadditiven gemischten Modellen sind nicht nur Regressionsparameter, sondern auch die Varianzparameter der zufälligen Effekte zu schätzen. Für diese Schätzungen ist sowohl für `ggamm` als auch für die beiden Varianten der Hyperparameter mit `BayesX` der jeweilige Bias in Tabelle 5.11 wiedergegeben. Dabei wurden zur Bestimmung der Verzerrungen nicht die theoretischen Werte der Varianzparameter verwendet, sondern die empirischen Varianzen, die sich aus den Realisationen $b_{1j}, \ldots, b_{24j}$, $j = 0, 1, 2$ ergeben. Dieser Vergleich erlaubt eine fairere Beurteilung der Schätzungen, da die Verfahren ja lediglich die in den Daten tatsächlich vorhandene Variation untersuchen können.

Wie man sieht, werden die Varianzparameter durch `ggamm` relativ dicht an den entsprechenden empirischen Werten geschätzt. Mit `BayesX` erhält man dagegen teilweise deutliche Verzerrungen, insbesondere für die Varianzen der zufälligen Steigungen b_1 und b_2 bei bernoulliverteiltem Response. Diese werden deutlich unterschätzt, was auch die schlechte Qualität der Schätzungen erklärt, die aus den Boxplots der MSEs zu erkennen war.

Für die mit der Funktion `ggamm` erzielten Schätzungen sollen nun wieder eine Reihe zusätzlicher Betrachtungen angestellt werden. Dazu sind in den Abbildungen 5.17 bis 5.20 die mittleren Schätzungen der einzelnen Modellkomponenten wiedergegeben. Für die räumliche Funktion wurde außerdem der Bias jeweils in einer eigenen Grafik visualisiert, da sich Unterschiede zur wahren Funktion relativ

schwierig aus den mittleren Schätzungen ablesen lassen.

Für normalverteilten Response werden alle Modellkomponenten in ausreichender Qualität und nahezu unverzerrt geschätzt. Für bernoulliverteilten Response erhält man dagegen mit Ausnahme der Funktion f_1 deutlich zu glatte Schätzungen, wie dies bereits für einige Funktionen in der Simulation zu generalisierten additiven Modellen der Fall war. Bei binomialverteiltem Response wird die Verzerrung bereits geringer, ist aber immer noch relativ stark ausgeprägt. Betrachtet man die Ergebnisse für poissonverteilten Response, so werden sowohl f_1 als auch f_{spat} nahezu unverzerrt geschätzt, während für die zufälligen Effekte die Verzerrung noch erkennbar ist. Die Schätzungen weisen aber erneut eine geringere Verzerrung als bei binomialverteiltem Response auf. Die Ergebnisse bestätigen damit die erwartbare Tendenz, dass mit steigendem Informationsgehalt in den Daten die Qualität der Schätzungen ansteigt.

Verteilung		f_1	f_{spat}	b_0	b_1	b_2
Normal	frequentistisch	0.956	0.899	0.932	0.924	0.933
	bayesianisch	0.980	0.993	0.993	0.976	0.986
Bernoulli	frequentistisch	0.938	0.715	0.774	0.472	0.635
	bayesianisch	0.967	0.900	0.915	0.722	0.854
Binomial	frequentistisch	0.948	0.847	0.898	0.767	0.850
	bayesianisch	0.975	0.990	0.963	0.914	0.947
Poisson	frequentistisch	0.953	0.926	0.913	0.870	0.909
	bayesianisch	0.980	0.998	0.972	0.949	0.970

Tabelle 5.12: Mittlere Überdeckungswahrscheinlichkeiten (nominaler Sicherheitsgrad: 95%).

Zusätzlich zu den mittleren Schätzungen wurden mit Hilfe der in Kapitel 3.5 beschriebenen Vorgehensweise Konfidenzintervalle beziehungsweise Konfidenzbänder für die einzelnen Modellkomponenten zum Sicherheitsgrad 95% bestimmt. In Tabelle 5.12 sind die mittleren Überdeckungswahrscheinlichkeiten wiedergegeben. Wie man sieht, sind diese für f_1 für alle vier Verteilungen zufriedenstellend. Für bernoulli- beziehungsweise binomialverteilten Response schneiden die bayesianischen Varianten etwas besser ab, während sie für normal- und poissonverteilten Response eher zu konservativ sind.

Für die räumliche Funktion f_{spat} und die zufälligen Effekte b_0, b_1 und b_2 erhält man ein etwas anderes Bild: In nahezu allen Fällen sind die frequentistischen Kon-

fiednzintervalle relativ weit vom nominalen Sicherheitsgrad entfernt, während die bayesianischen Varianten generell besser abschneiden. Dennoch erkennt man auch hier für bernoulliverteilten Response deutliche Abweichungen vom vorgegebenen Sicherheitsgrad.

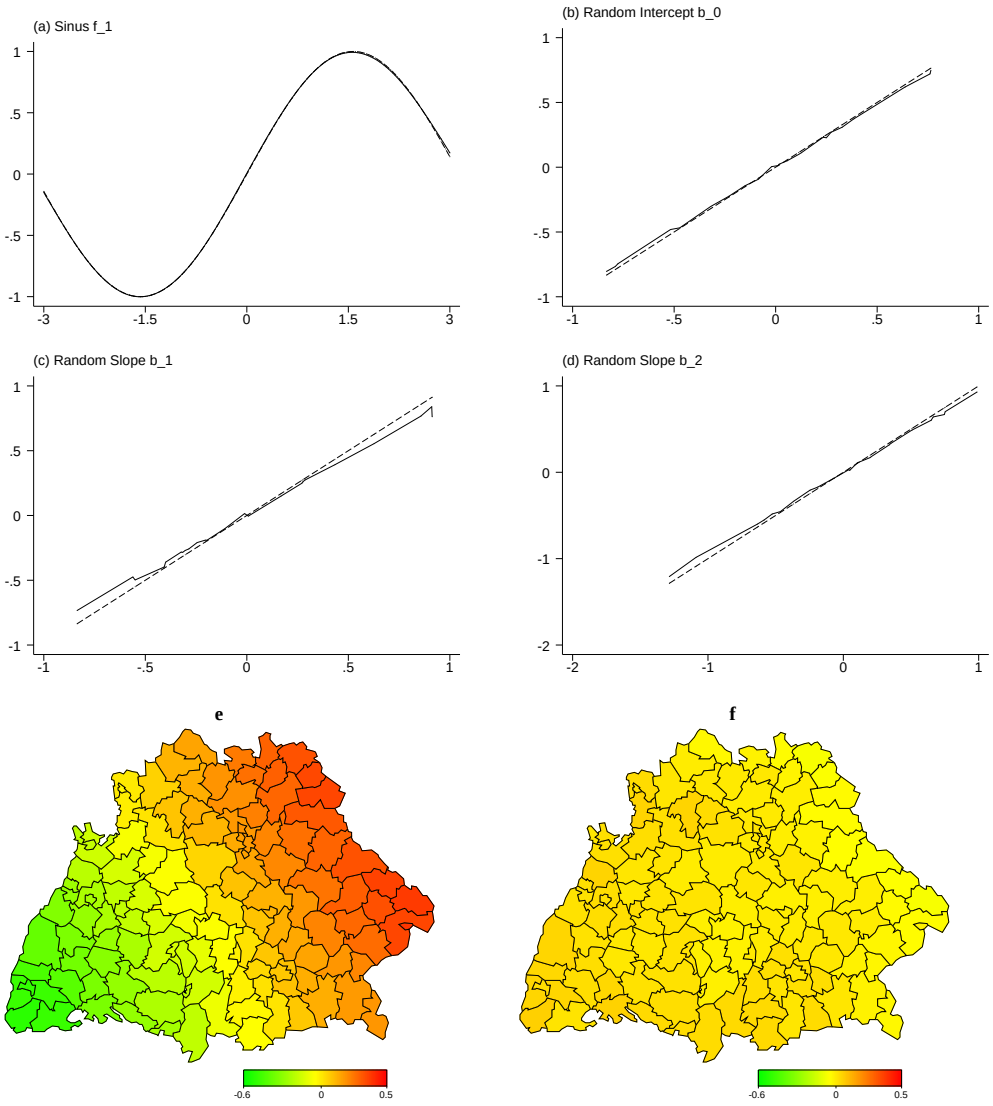

Abbildung 5.17: Normalverteilter Response: Mittlere Schätzungen der Modellkomponenten ((a)-(e)) und Bias der räumlichen Funktion (f). In den Grafiken (a) bis (d) sind die wahren Werte gestrichelt wiedergegeben.

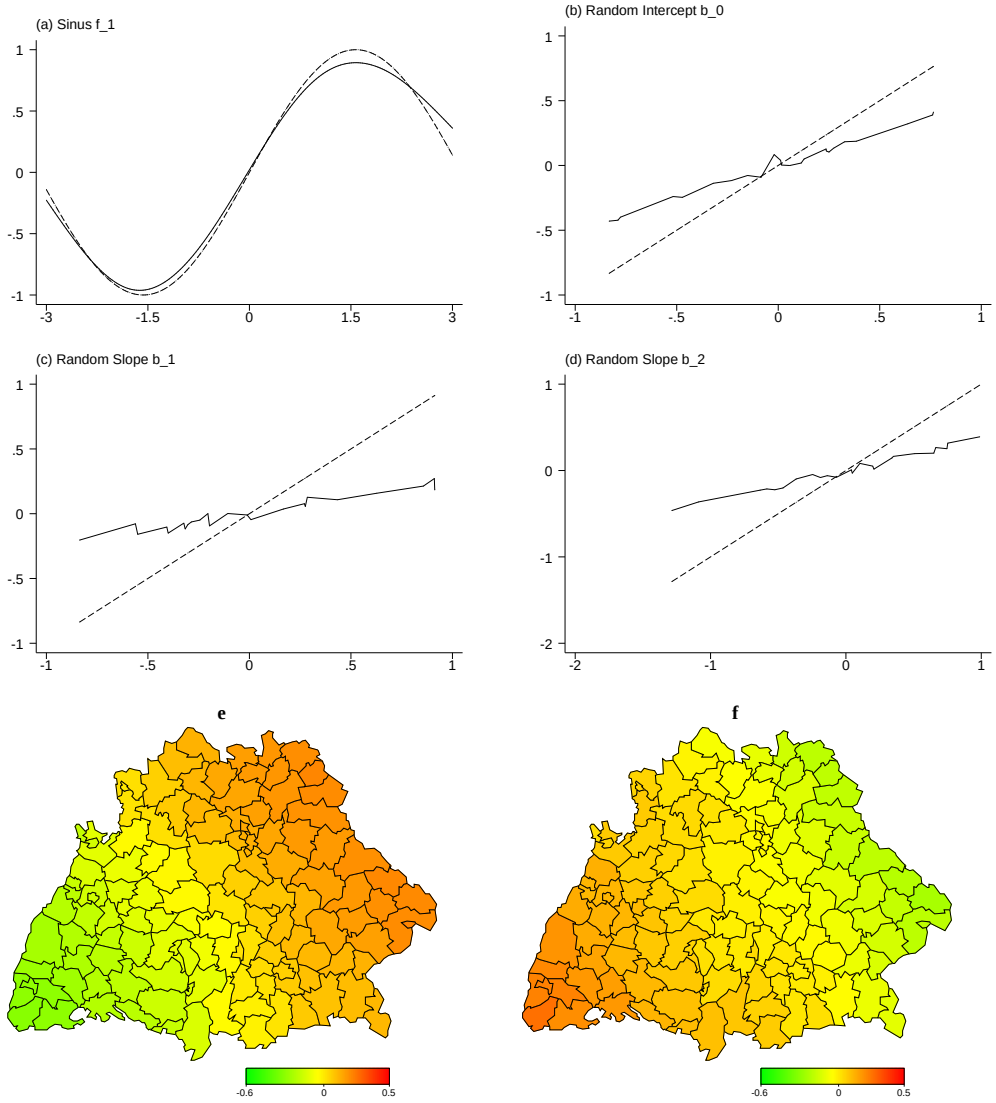

Abbildung 5.18: Bernoulliverteilter Response: Mittlere Schätzungen der Modellkomponenten ((a)-(e)) und Bias der räumlichen Funktion (f). In den Grafiken (a) bis (d) sind die wahren Werte gestrichelt wiedergegeben.

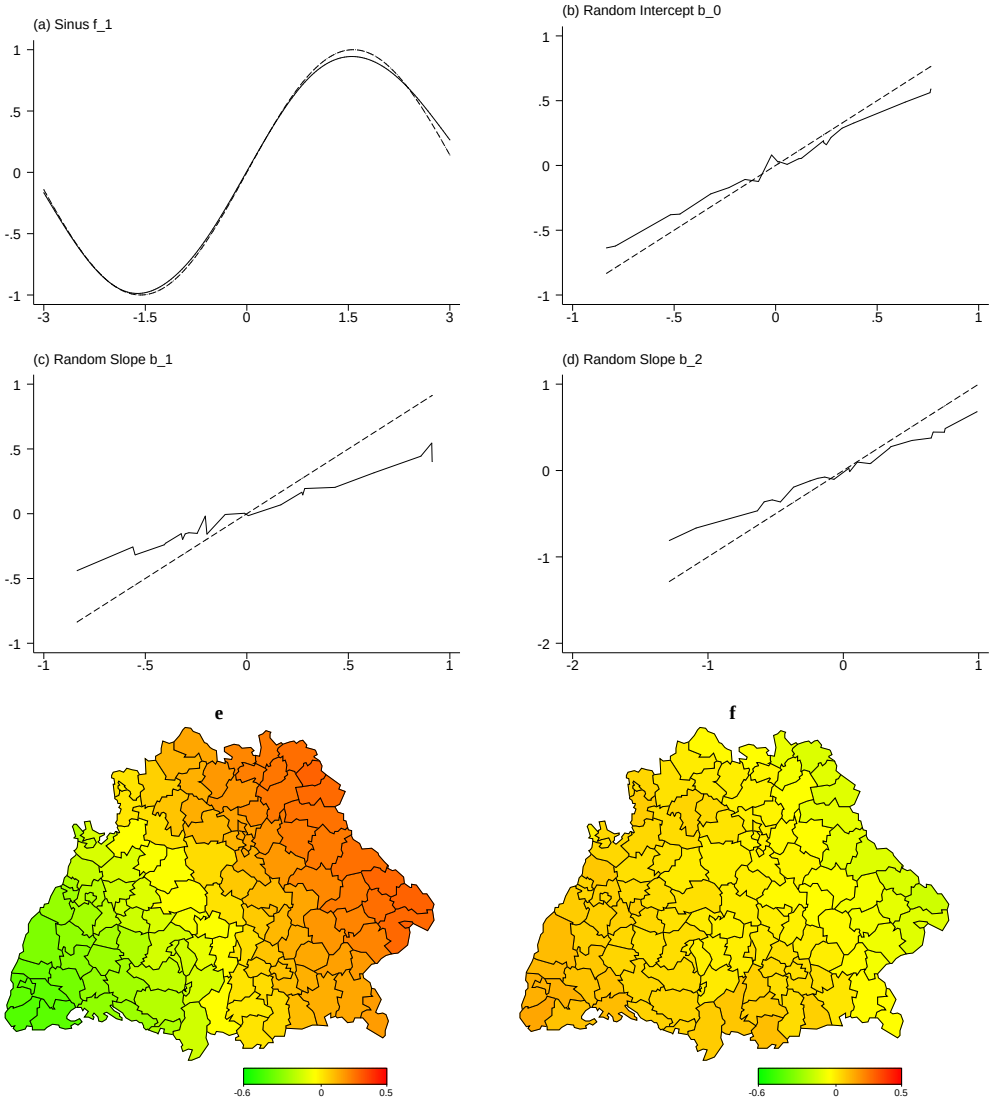

Abbildung 5.19: Binomialverteilter Response: Mittlere Schätzungen der Modellkomponenten ((a)-(e)) und Bias der räumlichen Funktion (f). In den Grafiken (a) bis (d) sind die wahren Werte gestrichelt wiedergegeben.

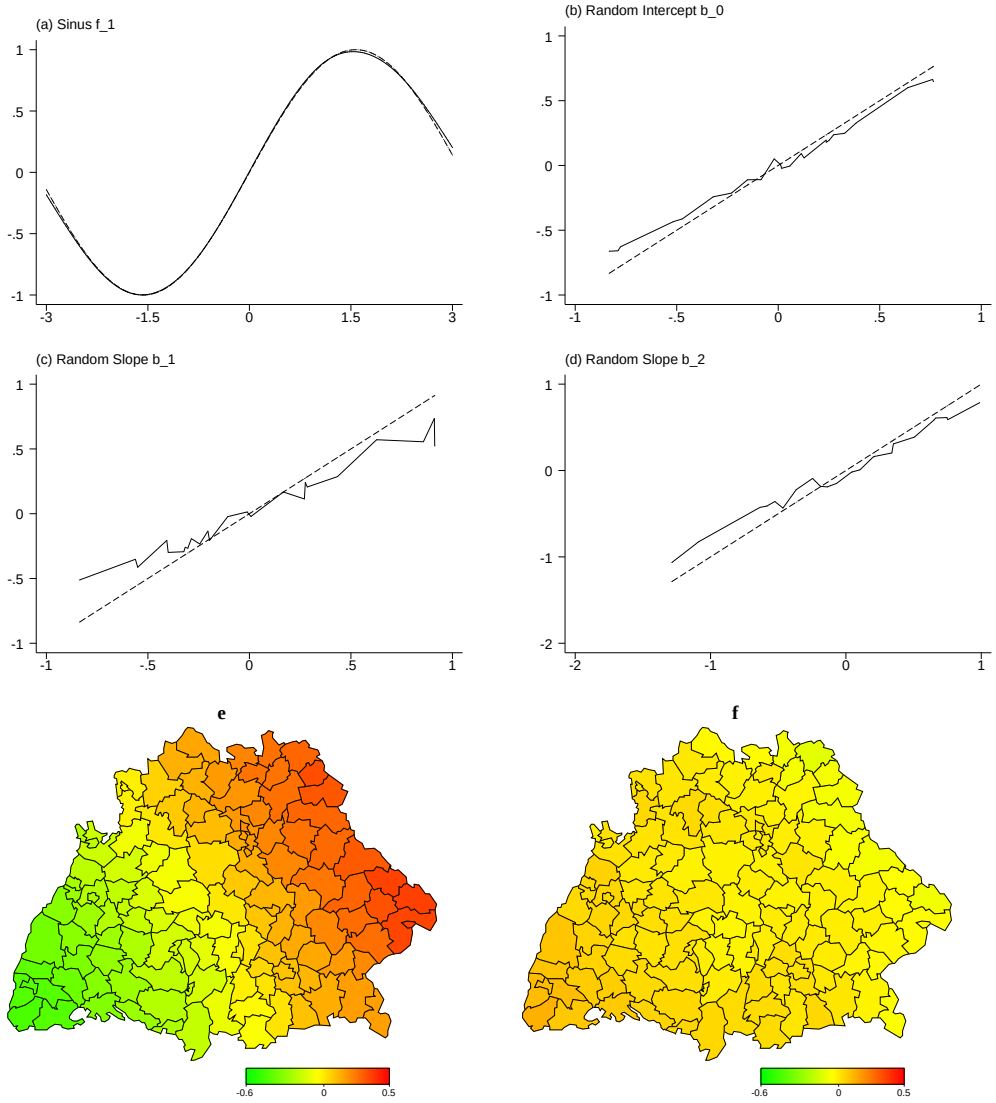

Abbildung 5.20: Poissonverteilter Response: Mittlere Schätzungen der Modellkomponenten ((a)-(e)) und Bias der räumlichen Funktion (f). In den Grafiken (a) bis (d) sind die wahren Werte gestrichelt wiedergegeben.

5.3 LQ-Tests

Nun sollen die in Kapitel 4 besprochenen Likelihood-Quotienten-Tests, mit deren Hilfe die Varianz der zufälligen Effekte eines linearen gemischten Modells auf den Wert 0 getestet werden kann, per Simulation auf ihre Güteeigenschaften untersucht werden. Die Betrachtungen sollen dabei auf die in Kapitel 4.3 beschriebene Möglichkeit beschränkt bleiben, den Zusammenhang zwischen einer als P-Spline modellierten Kovariablen und der abhängigen Variablen auf ein Polynom vom Grad $k-1$ zu testen. Mit k wird dabei wieder die Ordnung der als Penalisierung verwendeten Differenzen bezeichnet.

Speziell sollen die Möglichkeiten betrachtet werden, mit $k = 1$ auf das Vorhandensein eines Effektes der Kovariable und mit $k = 2$ auf die Linearität dieses Effekts zu testen. Es werden also für das Modell

$$y = f(x) + \varepsilon \tag{5.3}$$

die beiden Tests

$$H_0 : f(x) = \beta_0 \qquad \text{versus} \qquad H_1 : f(x) \neq \beta_0$$

und

$$H_0 : f(x) = \beta_0 + \beta_1 x \qquad \text{versus} \qquad H_1 : f(x) \neq \beta_0 + \beta_1 x$$

untersucht.

Für beide Tests wurden sowohl Modelle unter der Nullhypothese als auch unter der Alternative simuliert. Damit ist es zum einen möglich, zu beurteilen, ob die Tests das vorgegebene Signifikanzniveau einhalten und zum anderen, für gewisse Modelle einen Eindruck von der Güte der Tests zu erhalten.

In Abbildung 5.21 sind die fünf verschiedenen, für $k = 1$ verwendeten Funktionen abgebildet. Während die horizontale Funktion der Situation unter der Nullhypothese entspricht, spiegeln die übrigen Funktionen Situationen wachsender Nichtlinearität wieder. Die Funktionen wurden dabei definiert als

$$f(x) = 1 + s \cdot \sin(2\pi x)$$

mit verschiedenen Skalierungsfaktoren s. Die einzelnen Werte von s sind in Tabelle 5.13 zusammen mit dem entsprechenden empirischen Signal-Rauschen-Verhältnis wiedergegeben, das als

$$SNR = \frac{\mathrm{Var}(f(x))}{\mathrm{Var}(\varepsilon)}$$

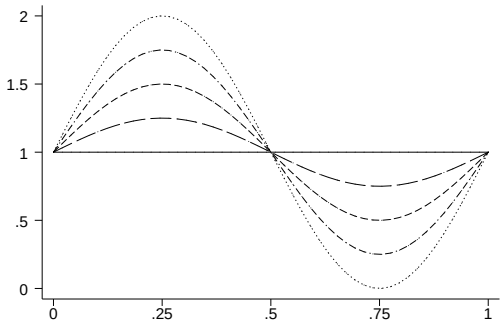

Abbildung 5.21: $f(x) = 1 + s \cdot \sin(2\pi x)$ *für Skalierungsfaktoren* $s \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1\}$.

definiert ist. Unter $\mathrm{Var}(f(x))$ ist dabei die empirische Varianz der n verschiedenen Funktionswerte $f(x_1)$ bis $f(x_n)$ zu verstehen. Für die Varianz des Fehlers ε wurde $\mathrm{Var}(\varepsilon) = \sigma^2 = 0.5$ gewählt.

s	0	0.25	0.5	0.75	1
SNR	0	0.0625	0.25	0.5625	1

Tabelle 5.13: Verschiedene Werte des Skalierungsfaktors s und entsprechendes empirisches Signal-Rauschen-Verhältnis.

Analog wurden die in Abbildung 5.22 wiedergegebenen Funktionen für den Test auf Linearität, das heißt für $k = 2$, als

$$f(x) = 1 + x + s \cdot \sin(2\pi x)$$

definiert. Die für $k = 1$ verwendeten Funktionen wurden also zusätzlich noch mit einer Geraden überlagert. Man erhält so, ausgehend von einer linearen Funktion unter H_0 wieder Funktionen mit wachsender Nichtlinearität für steigende Werte des Skalierungsfaktors. Die verschiedenen Werte des Skalierungsfaktors und des empirischen Signal-Rauschen-Verhältnisses stimmen dabei mit den in Tabelle 5.13 angegebenen Werten überein.

Für Stichproben vom Umfang $n = 50$ und $n = 100$ wurden nun für $k = 1$ und $k = 2$ und jeden Skalierungsfaktor 1000 Realisationen des Modells (5.3) erzeugt und die entsprechenden Schätzungen des Signal-Rauschen-Verhältnisses γ per Maximum-Likelihood- und Restricted-Maximum-Likelihood-Schätzung bestimmt. Zusätzlich wurden noch zwei verschiedene Knotenzahlen, nämlich $m = 20$

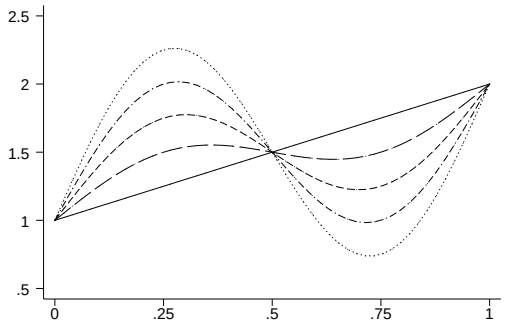

Abbildung 5.22: $f(x) = 1 + x + s \cdot \sin(2\pi x)$ *für Skalierungsfaktoren* $s \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1\}$.

und $m = 40$ zur Schätzung verwendet. Basierend auf den Schätzungen für γ wurden dann die Likelihood-Quotienten-Teststatistik LQ_n beziehungsweise die Restricted-Likelihood-Quotienten-Teststatistik RLQ_n berechnet.

Aufgrund der in Kapitel 3.4 beschriebenen Konvergenzprobleme kann dabei γ nicht tatsächlich als $\hat{\gamma} = 0$ geschätzt werden. Stattdessen wird die Schätzung bei einem kleinen Wert von $\hat{\gamma}$ gestoppt werden, falls das Maximum der Log-Likelihood beziehungsweise der Restricted-Log-Likelihood auf dem Rand des Parameterraums liegt. Dennoch ist es möglich, Modelle zu identifizieren, für die der ML- beziehungsweise REML-Schätzer tatsächlich 0 lauten müsste. In diesen Modellen erhält man nämlich negative Realisationen der jeweiligen Teststatistik. Diese Tatsache lässt sich einfach aus dem Verlauf der Restricted-Likelihood in Abbildung 3.8 ablesen.

REML				ML			
$k = 1$		$k = 2$		$k = 1$		$k = 2$	
$m = 20$	$m = 40$	$m = 20$	$m = 40$	$m = 20$	$m = 40$	$m = 20$	$m = 40$
0.632	0.646	0.441	0.421	0.457	0.456	0.430	0.425

Tabelle 5.14: Anteil der Schätzungen mit $LQ_n = 0$ *beziehungsweise* $RLQ_n = 0$ *unter* H_0 *bei Stichprobenumfang* $n = 50$.

In den Tabellen 5.14 und 5.15 sind die Anteile der Schätzungen der unter H_0 erzeugten Modelle angegeben, für die $\hat{\gamma} = 0$ und damit $LQ_n = 0$ beziehungsweise $RLQ_n = 0$ galt. Vergleicht man diese Anteile mit den Wahrscheinlichkeitsmassen der entsprechenden asymptotischen Verteilungen in Null, die in Tabelle 4.4

REML				ML			
$k=1$		$k=2$		$k=1$		$k=2$	
$m=20$	$m=40$	$m=20$	$m=40$	$m=20$	$m=40$	$m=20$	$m=40$
0.656	0.655	0.456	0.462	0.486	0.489	0.469	0.469

Tabelle 5.15: Anteil der Schätzungen mit $LQ_n = 0$ beziehungsweise $RLQ_n = 0$ unter H_0 bei Stichprobenumfang $n = 100$.

gegeben sind, so erkennt man deutliche Abweichungen für die REML-Schätzung bei $k = 2$ sowie in allen Fällen für die ML-Schätzung. Lediglich für die REML-Schätzung bei $k = 1$ erhält man dicht an den asymptotisch vorhergesagten Wahrscheinlichkeiten liegende Werte. Wie man außerdem sieht, liegen die Anteile für den größeren Stichprobenumfang $n = 100$ zwar näher an den asymptotisch erwarteten Werten, sind aber immer noch, mit Ausnahme von $k = 1$ in der REML-Schätzung, weit von diesen entfernt.

Um diese Abweichungen von der asymptotisch vorhergesagten Verteilung weiter zu untersuchen, wurden QQ-Plots der Realisationen der Likelihood-Quotienten-Teststatistiken unter der Bedingung $LQ_n > 0$ beziehungsweise $RLQ_n > 0$ gegen die jeweiligen asymptotischen Verteilungen unter der Bedingung $LQ_\infty > 0$ erzeugt. Man beachte an dieser Stelle wieder, dass die asymptotische Verteilung der Restricted-Likelihood-Quotienten-Teststatistik mit der asymptotischen Verteilung der Likelihood-Quotienten-Teststatistik übereinstimmt. In Abbildung 5.23 und 5.24 sind die QQ-Plots für die Restricted-Likelihood-Quotienten-Teststatistik RLQ_n wiedergegeben. Man erkennt eine ähnliche Struktur wie in den Tabellen 5.14 und 5.15. Für erste Differenzen ist die Anpassung der Verteilung von RLQ_n an die asymptotische Verteilung bereits für den Stichprobenumfang $n = 50$ sehr gut, während sie für zweite Differenzen und Stichprobenumfang $n = 50$ wesentlich schlechter ist. Für einen Stichprobenumfang von $n = 100$ erhält man zwar auch für zweite Differenzen bereits eine wesentlich bessere Anpassung, die aber immer noch relativ weit von der asymptotischen Verteilung entfernt ist. Für die Likelihood-Quotienten-Teststatistik LQ_n ähneln die QQ-Plots sowohl für $k = 1$ als auch für $k = 2$ den Grafiken (b) und (d), so dass an dieser Stelle auf die Wiedergabe verzichtet werden soll.

In den folgenden Tabellen 5.16 bis 5.19 sind nun die Anteile der Modelle angegeben, für die die Nullhypothese zugunsten der Alternative verworfen wurde. Für $s = 0$ entspricht dies dem Fehler erster Art. Über diesen kann kontrolliert werden,

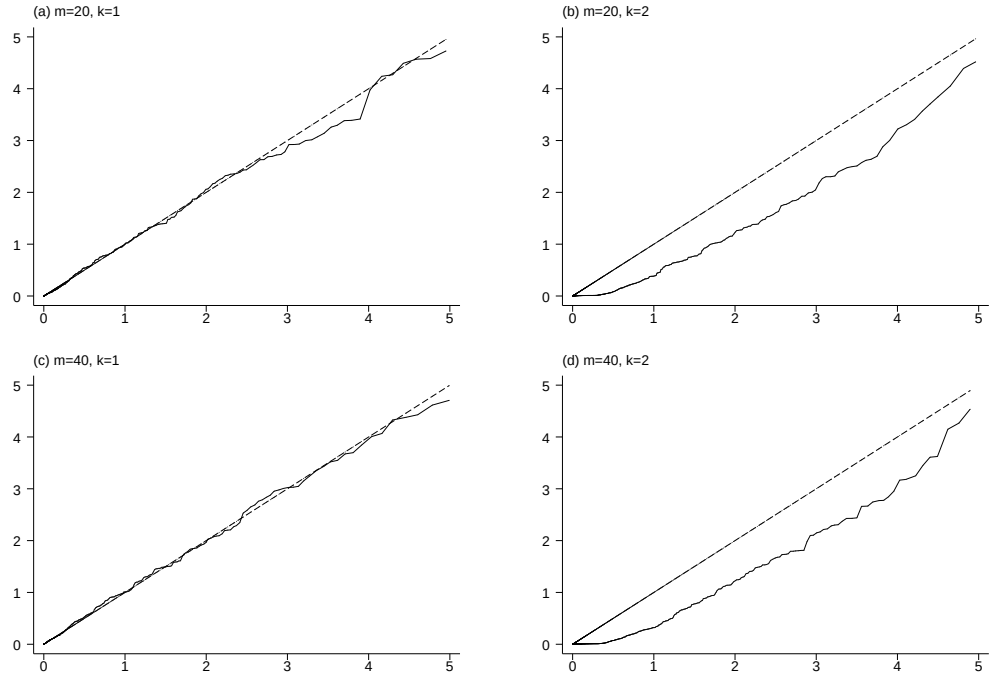

Abbildung 5.23: QQ-Plots der Restricted-Likelihood-Quotienten-Teststatistik unter der Bedingung $RLQ_n > 0$ gegen die asymptotische Verteilung unter der Bedingung $LQ_\infty > 0$ für $n = 50$. Die Quantile der asymptotischen Verteilung sind auf der horizontalen Achse abgetragen.

wie gut der Test das Signifikanzniveau einhält, das jeweils in der zweiten Spalte der Tabellen angegeben ist. Für $s > 0$ befindet man sich dagegen im Bereich der Alternative, so dass die in der Tabelle angegebenen Anteile der Güte des Tests für diesen speziellen Skalierungsfaktor entsprechen.

In den Tabellen sind nicht nur Anteile für den in Kapitel 4.3 hergeleiteten, als exakt bezeichneten Test angegeben, sondern auch für zwei mögliche Approximationen. Diese basieren auf Mischungen zweier χ^2-verteilten Zufallsvariablen mit null und einem Freiheitsgrad und unterscheiden sich in der Wahl der Mischungsgewichte. Zum einen wurde eine $(p_0, 1 - p_0)$-Mischung verwendet, wobei p_0 aus Tabelle 4.4, also als Wahrscheinlichkeitsmasse der asymptotischen Verteilung der Likelihood-Quotienten-Teststatistik im Punkt Null, gewählt wurde. Zum anderen wurde die (0.5,0.5)-Mischung betrachtet, die sich aus der Theorie für unabhängige, identisch verteilte Beobachtungen von Self & Liang (1987) ergibt.

Man beachte, dass der in Kapitel 4.3 hergeleitete Test zwar in den folgenden

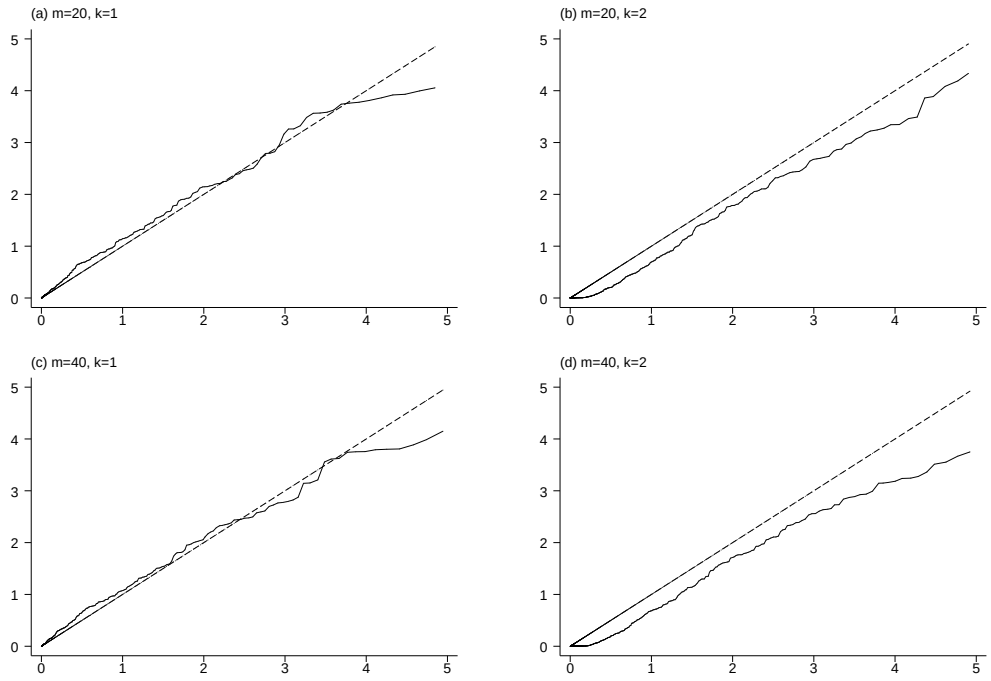

Abbildung 5.24: QQ-Plots der Restricted-Likelihood-Quotienten-Teststatistik unter der Bedingung $RLQ_n > 0$ gegen die asymptotische Verteilung unter der Bedingung $LQ_\infty > 0$ für $n = 100$. Die Quantile der asymptotischen Verteilung sind auf der horizontalen Achse abgetragen.

Tabellen als exakter Test bezeichnet wird, der Test aber auf der asymptotischen Verteilung der Likelihood-Quotienten-Teststatistik beruht. Es wird also die exakte, asymptotische Verteilung der Likelihood-Quotienten-Teststatistik verwendet, im Gegensatz zu den beschriebenen Approximationen.

Aus den Tabellen kann man ersehen, dass der exakte Test das Signifikanzniveau in allen Fällen relativ gut einhält. Es fällt jedoch auf, dass die $(p_0, 1-p_0)$-Mischung zweier χ^2-Verteilungen teilweise noch dichter am vorgegebenen Signifikanzniveau liegt, was aus den oben beschriebenen Abweichungen der Verteilungen von LQ_n beziehungsweise RLQ_n von der asymptotischen Verteilung begründet werden kann. Die $(0.5, 0.5)$-Mischung ist dagegen in allen Fällen zu konservativ.

Wie aus den Betrachtungen in Kapitel 4.3 zu erwarten war, ist der Anteil der Modelle, für die die Nullhypothese verworfen wird, unter Verwendung der exakten asymptotischen Verteilung stets größer als bei Verwendung der $(p_0, 1-p_0)$-Mischung und für diese wiederum größer als bei Verwendung der $(0.5, 0.5)$-

Mischung. Dies gilt nicht nur unter H_0, sondern auch im Bereich der Alternative.

Weiterhin fällt auf, dass bei P-Splines mit Differenzen der Ordnung $k = 2$ die Güte der Tests geringer ausfällt als für $k = 1$. Es ist also ein größeres Signal-Rauschen-Verhältnis erforderlich, damit Abweichungen von der Nullhypothese tatsächlich mit relativ großer Sicherheit durch den Test erkannt werden. Wie zu erwarten war, steigt die Güte bei größerem Stichprobenumfang schneller an, das heißt, bei gleichem Signal-Rauschen-Verhältnis erkennt man Abweichungen von der Nullhypothese bei einem Stichprobenumfang von $n = 100$ mit größerer Wahrscheinlichkeit als für $n = 50$. Obwohl dies natürlich keinesfalls einem theoretischen Nachweis von Güteeigenschaften entspricht, erhält man zumindest einen Hinweis auf die mögliche Konsistenz der Tests.

Insgesamt lässt sich festhalten, dass die Unterschiede zwischen dem auf der exakten asymptotischen Verteilung basierenden Test und den beiden Approximationen relativ gering ausfällt. Damit stellt sich die Frage, ob die relativ aufwändige Bestimmung der exakten asymptotischen Verteilung tatsächlich notwendig ist, oder ob eine weniger exakte, aber dafür leichter zu bestimmende Approximation nicht die gleichen Dienste leisten könnte. Insbesondere ließen sich basierend auf der Theorie von Self & Liang (1987) auch Tests in komplexeren Modellen bestimmen, mit denen mehrere Varianzparameter simultan getestet werden könnten. Eine Untersuchung dieser Frage könnte beispielsweise im Rahmen einer weiteren Simulationsstudie erfolgen, die jedoch in dieser Arbeit nicht mehr berücksichtigt werden konnte.

$m = 20,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.054	0.263	0.808	0.995	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.049	0.245	0.799	0.993	1.000
$(0.5, 0.5)$-Approx.	0.05	0.029	0.194	0.746	0.989	0.999
exakter Test	0.01	0.011	0.108	0.584	0.958	0.999
$(p_0, 1-p_0)$-Approx.	0.01	0.006	0.090	0.552	0.948	0.999
$(0.5, 0.5)$-Approx.	0.01	0.005	0.073	0.497	0.927	0.998
$m = 40,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.049	0.273	0.815	0.996	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.043	0.253	0.804	0.995	1.000
$(0.5, 0.5)$-Approx.	0.05	0.033	0.209	0.751	0.990	1.000
exakter Test	0.01	0.009	0.103	0.590	0.960	0.999
$(p_0, 1-p_0)$-Approx.	0.01	0.007	0.092	0.571	0.955	0.999
$(0.5, 0.5)$-Approx.	0.01	0.007	0.071	0.512	0.938	0.998
$m = 20,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.046	0.113	0.323	0.694	0.942
$(p_0, 1-p_0)$-Approx.	0.05	0.035	0.100	0.296	0.658	0.930
$(0.5, 0.5)$-Approx.	0.05	0.021	0.076	0.25	0.584	0.906
exakter Test	0.01	0.012	0.038	0.166	0.431	0.844
$(p_0, 1-p_0)$-Approx.	0.01	0.009	0.030	0.140	0.383	0.807
$(0.5, 0.5)$-Approx.	0.01	0.006	0.023	0.109	0.336	0.762
$m = 40,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.047	0.116	0.333	0.703	0.955
$(p_0, 1-p_0)$-Approx.	0.05	0.040	0.104	0.309	0.678	0.945
$(0.5, 0.5)$-Approx.	0.05	0.025	0.077	0.264	0.611	0.915
exakter Test	0.01	0.014	0.044	0.171	0.467	0.858
$(p_0, 1-p_0)$-Approx.	0.01	0.011	0.032	0.149	0.415	0.830
$(0.5, 0.5)$-Approx.	0.01	0.006	0.021	0.117	0.351	0.787

Tabelle 5.16: Anteil der Schätzungen, für die die Nullhypothese zugunsten der Alternative verworfen wurde bei Restricted-Maximum-Likelihood-Schätzung und Stichprobenumfang $n = 50$. In der zweiten Spalte ist jeweils das Signifikanzniveau des Tests angegeben.

$m = 20,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.060	0.280	0.820	0.996	1.000
$(p_0, 1 - p_0)$-Approx.	0.05	0.053	0.259	0.807	0.995	1.000
$(0.5, 0.5)$-Approx.	0.05	0.034	0.207	0.759	0.989	0.999
exakter Test	0.01	0.013	0.114	0.614	0.962	0.999
$(p_0, 1 - p_0)$-Approx.	0.01	0.009	0.100	0.569	0.956	0.999
$(0.5, 0.5)$-Approx.	0.01	0.006	0.080	0.515	0.936	0.999
$m = 40,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.057	0.286	0.830	0.997	1.000
$(p_0, 1 - p_0)$-Approx.	0.05	0.051	0.260	0.810	0.995	1.000
$(0.5, 0.5)$-Approx.	0.05	0.035	0.222	0.768	0.991	1.000
exakter Test	0.01	0.012	0.106	0.609	0.963	0.999
$(p_0, 1 - p_0)$-Approx.	0.01	0.011	0.099	0.589	0.959	0.999
$(0.5, 0.5)$-Approx.	0.01	0.007	0.081	0.529	0.943	0.998
$m = 20,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.049	0.131	0.347	0.723	0.949
$(p_0, 1 - p_0)$-Approx.	0.05	0.042	0.108	0.321	0.687	0.941
$(0.5, 0.5)$-Approx.	0.05	0.027	0.082	0.277	0.619	0.917
exakter Test	0.01	0.014	0.044	0.193	0.469	0.864
$(p_0, 1 - p_0)$-Approx.	0.01	0.011	0.037	0.161	0.423	0.838
$(0.5, 0.5)$-Approx.	0.01	0.006	0.025	0.127	0.370	0.794
$m = 40,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.057	0.129	0.346	0.735	0.963
$(p_0, 1 - p_0)$-Approx.	0.05	0.043	0.111	0.324	0.703	0.956
$(0.5, 0.5)$-Approx.	0.05	0.029	0.082	0.277	0.643	0.930
exakter Test	0.01	0.014	0.044	0.193	0.499	0.876
$(p_0, 1 - p_0)$-Approx.	0.01	0.012	0.037	0.164	0.459	0.856
$(0.5, 0.5)$-Approx.	0.01	0.007	0.025	0.127	0.386	0.815

Tabelle 5.17: Anteil der Schätzungen, für die die Nullhypothese zugunsten der Alternative verworfen wurde bei Maximum-Likelihood-Schätzung und Stichprobenumfang $n = 50$. In der zweiten Spalte ist jeweils das Signifikanzniveau des Tests angegeben.

$m = 20,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.053	0.540	0.985	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.049	0.522	0.982	1.000	1.000
$(0.5, 0.5)$-Approx.	0.05	0.030	0.461	0.975	1.000	1.000
exakter Test	0.01	0.006	0.299	0.933	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.01	0.003	0.276	0.924	1.000	1.000
$(0.5, 0.5)$-Approx.	0.01	0.002	0.235	0.909	1.000	1.000
$m = 40,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.054	0.550	0.987	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.047	0.530	0.984	1.000	1.000
$(0.5, 0.5)$-Approx.	0.05	0.028	0.458	0.978	1.000	1.000
exakter Test	0.01	0.006	0.301	0.934	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.01	0.006	0.283	0.930	1.000	1.000
$(0.5, 0.5)$-Approx.	0.01	0.004	0.241	0.917	1.000	1.000
$m = 20,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.072	0.207	0.663	0.978	0.999
$(p_0, 1-p_0)$-Approx.	0.05	0.055	0.175	0.634	0.970	0.999
$(0.5, 0.5)$-Approx.	0.05	0.032	0.132	0.548	0.947	0.999
exakter Test	0.01	0.012	0.072	0.420	0.908	0.999
$(p_0, 1-p_0)$-Approx.	0.01	0.005	0.056	0.379	0.889	0.998
$(0.5, 0.5)$-Approx.	0.01	0.003	0.046	0.325	0.857	0.994
$m = 40,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.068	0.203	0.661	0.981	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.053	0.182	0.632	0.976	1.000
$(0.5, 0.5)$-Approx.	0.05	0.030	0.133	0.567	0.957	1.000
exakter Test	0.01	0.008	0.071	0.441	0.913	0.999
$(p_0, 1-p_0)$-Approx.	0.01	0.004	0.056	0.396	0.899	0.999
$(0.5, 0.5)$-Approx.	0.01	0.001	0.044	0.341	0.867	0.998

Tabelle 5.18: Anteil der Schätzungen, für die die Nullhypothese zugunsten der Alternative verworfen wurde bei Restricted-Maximum-Likelihood-Schätzung und Stichprobenumfang $n = 100$. In der zweiten Spalte ist jeweils das Signifikanzniveau des Tests angegeben.

$m = 20,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.057	0.549	0.986	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.050	0.528	0.982	1.000	1.000
$(0.5, 0.5)$-Approx.	0.05	0.030	0.467	0.976	1.000	1.000
exakter Test	0.01	0.006	0.311	0.936	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.01	0.005	0.285	0.929	1.000	1.000
$(0.5, 0.5)$-Approx.	0.01	0.002	0.242	0.912	1.000	1.000
$m = 40,\ k = 1$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.056	0.558	0.987	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.048	0.541	0.984	1.000	1.000
$(0.5, 0.5)$-Approx.	0.05	0.030	0.462	0.981	1.000	1.000
exakter Test	0.01	0.006	0.309	0.937	1.000	1.000
$(p_0, 1-p_0)$-Approx.	0.01	0.006	0.291	0.933	1.000	1.000
$(0.5, 0.5)$-Approx.	0.01	0.005	0.249	0.920	1.000	1.000
$m = 20,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.075	0.216	0.674	0.978	0.999
$(p_0, 1-p_0)$-Approx.	0.05	0.059	0.188	0.647	0.973	0.999
$(0.5, 0.5)$-Approx.	0.05	0.036	0.142	0.561	0.953	0.999
exakter Test	0.01	0.013	0.078	0.440	0.911	0.999
$(p_0, 1-p_0)$-Approx.	0.01	0.008	0.060	0.395	0.898	0.999
$(0.5, 0.5)$-Approx.	0.01	0.003	0.049	0.347	0.867	0.995
$m = 40,\ k = 2$						
	Sig.Niv.	$s = 0$	$s = 0.25$	$s = 0.5$	$s = 0.75$	$s = 1$
exakter Test	0.05	0.072	0.214	0.672	0.982	1.000
$(p_0, 1-p_0)$-Approx.	0.05	0.058	0.192	0.647	0.980	1.000
$(0.5, 0.5)$-Approx.	0.05	0.034	0.140	0.579	0.960	1.000
exakter Test	0.01	0.009	0.074	0.454	0.915	0.999
$(p_0, 1-p_0)$-Approx.	0.01	0.006	0.063	0.413	0.905	0.999
$(0.5, 0.5)$-Approx.	0.01	0.002	0.046	0.362	0.880	0.998

Tabelle 5.19: Anteil der Schätzungen, für die die Nullhypothese zugunsten der Alternative verworfen wurde bei Maximum-Likelihood-Schätzung und Stichprobenumfang $n = 100$. In der zweiten Spalte ist jeweils das Signifikanzniveau des Tests angegeben.

6 Datenanalysen

Nachdem in Kapitel 5.1 und 5.2 die statistischen Eigenschaften der in Kapitel 3 vorgestellten Verfahren im Rahmen von Simulationsstudien untersucht wurden, sollen nun anhand dreier Datenbeispiele mögliche Anwendungen dieser Verfahren demonstriert werden. Die Schätzungen wurden dabei wieder mit der S-Plus-Implementation `ggamm` durchgeführt, die in Anhang B genauer beschrieben ist.

6.1 Disease-Mapping

6.1.1 Modell

Unter der Bezeichnung Disease-Mapping werden eine Reihe von Ansätzen zusammengefasst, deren Ziel die Analyse der räumlichen Variation des Erkrankungs- beziehungsweise Mortalitätsrisikos bezüglich einer bestimmten Krankheit ist. Insbesondere sollen Regionen mit erhöhtem Risiko identifiziert werden, um Hinweise auf unbekannte Risikofaktoren zu erhalten, aber auch die Identifizierung von Regionen mit unterdurchschnittlichem Risiko ist von Interesse, weil diese Hinweise auf geeignete Präventions- oder Behandlungsmaßnahmen liefern können. Beide Ziele werden beispielsweise mit dem Krebsatlas der Bundesrepublik Deutschland verfolgt, der die Mortalität bezüglich bestimmter Krebsarten untersucht (Becker & Wahrendorf 1997).

Im Folgenden werden stets die Bezeichnungen Mortalitätsrisiko und Mortalitätsrate verwendet, um nicht zwischen der Analyse von Erkrankungen und Todesfällen unterscheiden zu müssen. Die beschriebenen Verfahren sind jedoch problemlos auf beide Fragestellungen anwendbar. Die Darstellung beruht im Wesentlichen auf Mollié (1996), teilweise auch auf Knorr-Held & Raßer (2000).

Die Datengrundlage der Analyse bildet die beobachtete Zahl der Todesfälle y_i, $i = 1, \ldots, n$ bezüglich einer bestimmten Krankheit in n verschiedenen Regionen eines zusammenhängenden Gebietes, also beispielsweise in den Kreisen Deutschlands, wie es für den Krebsatlas der Fall ist. Zusätzlich ist die erwartete Zahl von Todesfällen e_i, $i = 1, \ldots, n$ aufgrund bekannter Risikofaktoren, der Altersstruktur und der Bevölkerungszahl der jeweiligen Region gegeben.

Für relativ seltene Krankheiten ist es plausibel anzunehmen, dass sich die Zahl

der Todesfälle bezüglich dieser Krankheit durch die Annahme

$$y_i \sim \text{Po}(e_i r_i)$$

beschreiben lässt, wobei $r_1, \ldots, r_n$ die unbekannten, regionenspezifischen und um bekannte Risikofaktoren bereinigten Mortalitätsraten darstellen. Häufigstes Beispiel für solche Krankheiten sind verschiedene (seltene) Krebserkrankungen, wie beispielsweise Mundhöhlenkrebs.

Üblich sind bis heute zwei einfache Verfahren zur Visualisierung der räumlichen Variation des Mortalitätsrisikos: Die Berechnung von p-Werten unter der Annahme, dass $r_i = 1$ für $i = 1, \ldots, n$ gilt und die Bestimmung der sogenannten Standardmortalitätsraten (SMR). Im Krebsatlas werden beispielsweise die Standardmortalitätsraten verwendet.

Der Berechnung von p-Werten liegt die Idee zugrunde, zu bestimmen, wie wahrscheinlich die beobachtete Zahl (oder eine noch extremere Zahl) von Todesfällen in einer Region unter der Annahme ist, dass bereits alle Risikofaktoren bekannt sind und keine räumliche Variation vorliegt. Das heißt, man bestimmt die Wahrscheinlichkeit

$$p_i = \mathbb{P}(Y_i \geq y_i),$$

wobei angenommen wird, dass $Y_i \sim \text{Po}(e_i)$ gilt. Unterschreitet der p-Wert einen sehr kleinen Wert, so wurden in diesem Kreis überzufällig mehr Todesfälle beobachtet, als unter Berücksichtigung der bekannten Risikofaktoren zu erwarten gewesen wären. Formal kann dies als Durchführung des Tests von $H_0 : r_i = 1$ gegen $H_1 : r_i > 1$ aufgefasst werden. Mit dieser Vorgehensweise werden nur Kreise mit überdurchschnittlichem Mortalitätsrisiko identifiziert, analog wäre aber auch die Identifizierung von Regionen mit unterdurchschnittlichem Risiko möglich.

Problematisch ist dieses Vorgehen, weil die p-Werte wesentlich durch die erwartete Zahl von Todesfällen bestimmt werden. Bei einer großen Zahl erwarteter Todesfälle sind bereits kleinere Abweichungen von der erwarteten Anzahl statistisch signifikant, so dass mit Hilfe von p-Werten eventuell lediglich Regionen mit großer Bevölkerung identifiziert werden. Außerdem werden möglicherweise vorhandene räumliche Korrelationen nicht berücksichtigt. Insbesondere ist es nicht möglich, festzustellen, ob räumliche Korrelationen vorliegen und wie stark diese sind. Räumliche Korrelationen sind dabei ein Indiz für eine räumliche Struktur der zugrunde liegenden, unbekannten Risikofaktoren.

Die Standardmortalitätsraten erhält man als Maximum-Likelihood-Schätzer in einem nonparametrischen Ansatz zur Schätzung der Risikofaktoren r_i. Sie sind gegeben durch

$$SMR_i = \frac{y_i}{e_i}.$$

Dieser Schätzer weist jedoch ebenfalls entscheidende Nachteile auf: Die Standardabweichung des Schätzers ist $s_i = \sqrt{y_i}/e_i$ und damit proportional zu $\frac{1}{e_i}$, das heißt bei einer geringen erwarteten Zahl an Todesfällen ist die Schätzung des Mortalitätsrisikos sehr unsicher. Dies gilt insbesondere für Regionen mit geringer Einwohnerzahl, so dass die Betrachtung der Standardmortalitätsraten eventuell nur zur Identifizierung von Regionen mit wenig Einwohnern führt. Wie bei der Analyse mit Hilfe von p-Werten werden auch hier räumliche Korrelationen nicht berücksichtigt.

Beide Verfahren sind also aufgrund ihrer Abhängigkeit von der erwarteten Zahl an Todesfällen und der Nichtberücksichtigung räumlicher Korrelationen zur Betrachtung der räumlichen Variation des Mortalitätsrisikos problematisch und können ein verzerrtes Bild der zugrunde liegenden Risikostruktur wiedergeben.

Um dem ersten Problem zu begegnen bietet sich ein Verfahren von Clayton & Kaldor (1987) an, das die Standardmortalitätsraten, ähnlich wie die Schätzer für zufällige Effekte in gemischten Modellen, hin zu einem globalen oder lokalen Mittelwert schrumpft. Dabei muss vorab bekannt sein, ob die unbekannten Risikofaktoren eine räumliche Struktur aufweisen oder unstrukturiert über die Regionen variieren. Die Stärke der Schrumpfung richtet sich nach der Reliabilität der Daten, das heißt für Daten mit großer erwarteter Fallzahl liegt die Schätzung für r_i nahe an der Standardmortalitätsrate, während sie in Regionen mit geringer erwarteter Fallzahl stärker geschrumpft wird.

Da häufig nicht im Voraus bekannt ist, ob die unbekannten Risikofaktoren räumlich variieren oder nicht, wurde der Ansatz von Clayton & Kaldor durch Besag et al. (1991) dahingehend erweitert, dass sowohl strukturierte als auch unstrukturierte Effekte berücksichtigt werden. Das Modell ermöglicht es außerdem, zu entscheiden, welcher der beiden Effekte überwiegt beziehungsweise ob beide Effekte gleichwertig vorliegen. Dieser Ansatz bietet also gegenüber den Standardmortalitätsraten und der Verwendung von p-Werten die gewünschten Eigenschaften.

Dazu werden die Risikofaktoren r_i parametrisiert durch

$$r_i = \exp(\beta_0 + f_{i,spat} + b_i). \tag{6.1}$$

Für f_{spat} nimmt man ein Markov-Zufallsfeld an, das heißt $f_{i,spat}$ steht für den räumlich strukturierten Anteil von r_i, während b_i als regionenspezifischer, unkorrelierter zufälliger Effekt modelliert wird, das heißt die b_i werden als unabhängig und identisch $\mathrm{N}(0, \nu^2)$-verteilt angenommen. Bezeichnet man mit τ_{spat} den inversen Glättungsparameter des Markov-Zufallsfeldes, so lässt sich theoretisch anhand des Verhältnisses τ_{spat}/ν^2 entscheiden, welcher der beiden räumlichen Effekte überwiegt. Möchte man diesen Vergleich jedoch mit Hilfe der Schätzungen durchführen, so ist zu beachten, dass die Trennung des strukturierten und des unstrukturierten Effekts häufig nur eingeschränkt möglich ist (vergleiche die Simulationsstudie in Lang & Fahrmeir (2001)), die Entscheidung also eine relativ große Unsicherheit birgt.

Für y_i erhält man mit (6.1) ein log-lineares Poisson-Modell mit linearem Prädiktor $\eta_i = \log(e_i) + \beta_0 + f_{i,spat} + b_i$, das heißt ein log-lineares Poisson-Modell mit $\log(e_i)$ als sogenanntem Offset.

Das Modell mit Offset unterscheidet sich von den in Kapitel 3 behandelten Modellen für poissonverteilten Response durch den bekannten Anteil $\log(e_i)$ des linearen Prädiktors. Äquivalent erhält man $\mathbb{E}(y_i) = e_i \cdot \exp(\beta_0 + f_{i,spat} + b_i)$ mit dem bekannten Faktor e_i.

Die Verwendung eines vorab bekannten Anteils des linearen Prädiktors in Form des Offsets lässt sich folgendermaßen motivieren (Tutz (2000) Kapitel 7.2.4): Üblicherweise führt man die Poissonverteilung als Verteilung einer Zufallsgröße ein, die die in einem festen Zeitintervall eingetretenen Ereignisse eines bestimmten Typs zählt. Die Anzahl der Ereignisse hängt dabei von der Länge des Zeitintervalls Δt und der Intensitätsrate r mit der diese Ereignisse eintreten ab. Als Verteilung erhält man die $\mathrm{Po}(\Delta t \cdot r)$-Verteilung. Übertragen auf das Modell des Disease-Mappings hängt die Anzahl der beobachteten Todesfälle zum einen von der Einwohnerzahl der betrachteten Region und bekannten Risikofaktoren und zum anderen von unbekannten Risikofaktoren ab. Der Offset $\log(e_i)$ dient dann dazu, die Regionen bezüglich bekannter Risikofaktoren und der Einwohnerzahl vergleichbar zu machen.

Die Schätzung eines log-linearen Modells mit Offset ist mit leichten Verände-

rungen über ein generalisiertes geoadditives gemischtes Modell, wie in Kapitel 3 beschrieben, möglich. Dazu muss man lediglich die Arbeitsbeobachtungen $\tilde{y}_i$ und Arbeitsgewichte w_i modifizieren. Definiert man $\tilde{\eta}_i$ als den in Kapitel 3.1 eingeführten linearen Prädiktor und $\eta_i = \log(e_i) + \tilde{\eta}_i$ als den um den Offset erweiterten linearen Prädiktor, so werden die üblichen Definitionen (2.17) und (2.18) durch

$$\tilde{y}_i(\eta_i) = \tilde{\eta}_i + D_i(\eta_i)^{-1}(y_i - \mu_i(\eta_i))$$

und

$$w_i(\eta_i) = D_i(\eta_i)^2/\sigma_i^2(\eta_i)$$

ersetzt (Gamerman (1997), Abschnitt 3). Man beachte, dass der erste Summand von $\tilde{y}_i(\eta_i)$ nicht η_i, sondern $\tilde{\eta}_i$ ist. An allen anderen Stellen der Definitionen wird $\tilde{\eta}_i$ durch η_i ersetzt. Die Schätzung der Parameter erfolgt nun genau wie in Kapitel 3 beschrieben.

Theoretisch ist auch die direkte Einbeziehung von regionenspezifischen Kovariablen in das obige Modell mit Hilfe der in Kapitel 3 vorgestellten Methoden möglich. So könnte die Altersverteilung einer Region oder der Anteil der Raucher an der Gesamtbevölkerung unmittelbar berücksichtigt werden, statt dieses Vorwissen mit Hilfe der erwarteten Todesfälle e_i auszudrücken.

6.1.2 Anwendung

Nun soll das durch (6.1) beschriebene Modell für das Mortalitätsrisiko bezüglich Mundhöhlenkrebs geschätzt werden. Die Daten bestehen aus der Zahl der im Zeitraum von 1986 bis 1990 an Mundhöhlenkrebs gestorbenen Männer in allen 544 damaligen Kreisen West- und Ostdeutschlands sowie der erwarteten Anzahl von Todesfällen aufgrund bekannter Risikofaktoren für diese Kreise. Die erwartete Zahl an Todesfällen beruht dabei im Wesentlichen auf der Altersstruktur des entsprechenden Kreises. Diese Daten werden auch in Knorr-Held & Raßer (2000) Abschnitt 3.2 analysiert, allerdings mit dem Ziel der Identifizierung von Clustern mit konstantem Risiko. Die folgende Datenbeschreibung richtet sich weitgehend nach der dort gegebenen.

Insgesamt wurden 15466 Todesfälle aufgrund von Mundhöhlenkrebs registriert. Die größte Anzahl weist West-Berlin mit 501 Fällen auf, die minimale Zahl von

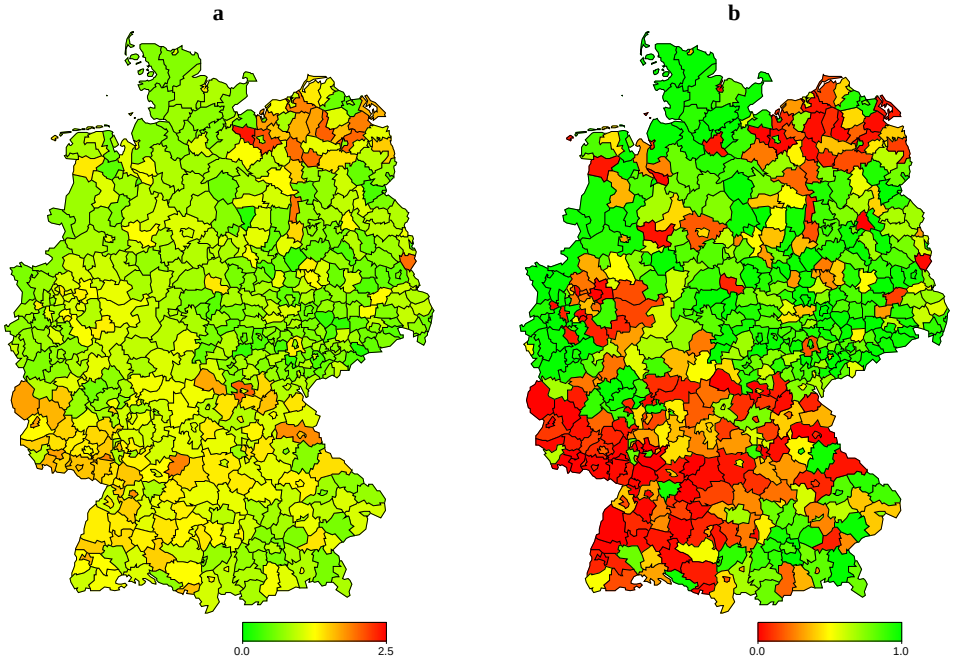

Abbildung 6.1: Standardmortalitätsraten (a) und p-Werte (b) für das Mortalitätsrisiko bezüglich Mundhöhlenkrebs unter Männern von 1986 bis 1990.

Todesfällen in einem Kreis ist 1. Die Mortalitätsrate für Gesamtdeutschland beträgt 40.9 Fälle pro 100000 Einwohner. Die berechneten Standardmortalitätsraten und p-Werte für den Test auf $r_i = 1$ sind in Abbildung 6.1 visualisiert. Bei den Standardmortalitätsraten fallen einige Kreise mit erhöhtem Risiko in Mecklenburg-Vorpommern auf, die auch bei den p-Werten als Kreise mit größerem Risiko identifiziert werden. Zusätzlich erkennt man bei den p-Werten Regionen erhöhten Risikos in Teilen von Rheinland-Pfalz, dem Saarland, Baden-Württemberg, Bayern und Hessen sowie im Ruhrgebiet. Aus den oben aufgeführten Gründen sind die Ergebnisse jedoch schwer zu interpretieren, da besonders die Kreise in Mecklenburg-Vorpommern nur eine geringe Bevölkerungszahl und damit auch eine geringe erwartete Zahl an Todesfällen aufweisen. Außerdem ist es schwierig in den Daten eine klare räumliche Struktur zu erkennen.

Minimale und maximale Standardmortalitätsrate sind 0.15 beziehungsweise 2.40, die Standardabweichung der logarithmierten Standardmortalitätsraten (unter Annahme konstanten Risikos für Gesamtdeutschland) beträgt 0.387. Auffällig ist, dass die beiden extremsten Werte in Regionen auftreten, die nur 6.85 beziehungsweise 4.17 erwartete Todesfälle aufweisen. Die Standardabweichung $s_i = \sqrt{(y_i)}/e_i$

der maximalen Standardmortalitätsrate beträgt darüberhinaus 0.76 und gehört damit zu den 1% größten Standardabweichungen die beobachtet wurden.

Betrachtet man nur Regionen mit mehr als 20 erwarteten Todesfällen, so reduziert sich die Standardabweichung der logarithmierten Standardmortalitätsraten auf 0.270, minimale und maximale Standardmortalitätsrate betragen 0.47 beziehungsweise 1.82. Hier zeigt sich wieder das oben beschriebene Problem, dass die Variabilität der Standardmortalitätsraten von der erwarteten Zahl von Todesfällen abhängt.

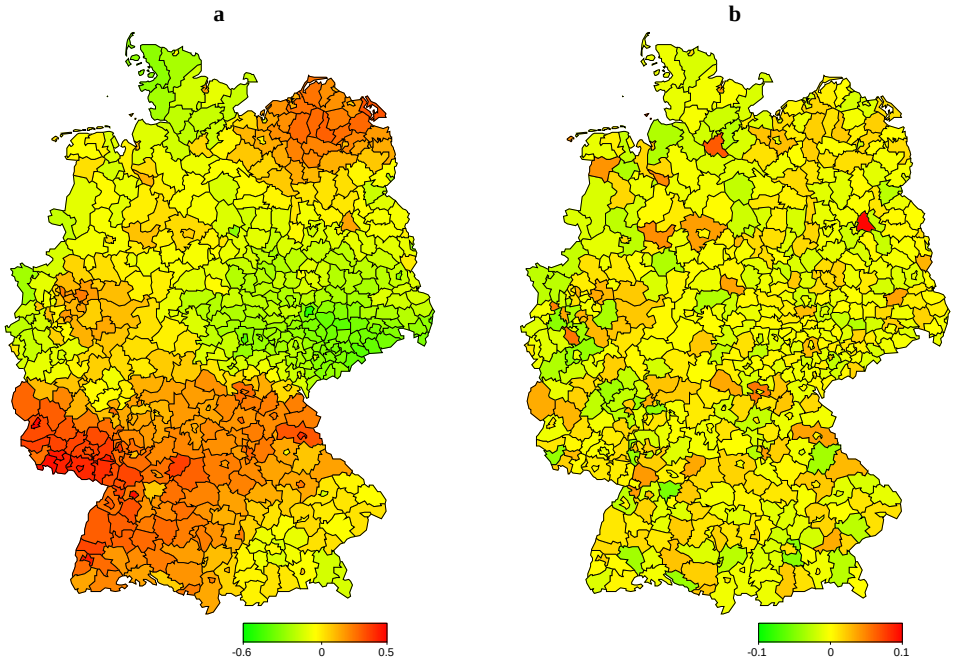

Abbildung 6.2: Schätzungen des strukturierten (a) und des unstrukturierten (b) räumlichen Effekts.

In der Analyse mit Hilfe der in Kapitel 3 beschriebenen Verfahren muss Rügen aus dem Datensatz entfernt werden, da der Kreis Rügen keine Verbindung zu anderen Kreisen besitzt. Die Nachbarschaftsmatrix weist dann ein Rangdefizit von 2 auf und somit ist eine Reparametrisierung wie in Kapitel 3.3 beschrieben nicht mehr möglich.

Abbildung 6.2 zeigt die resultierenden Schätzungen für den räumlich strukturierten beziehungsweise den räumlich unstrukturierten Effekt. Die zugehörigen geschätzten Hyperparameter findet man in Tabelle 6.1, die auch die Ergebnisse einer voll-bayesianischen Analyse mit Hilfe des Programms `BayesX` enthält.

	ggamm	BayesX
τ_{spat}	0.0668	0.0680
ν^2	0.0044	0.0037

Tabelle 6.1: Schätzungen des inversen Glättungsparameters des strukturierten Effekts und der Varianz des unstrukturierten Effekts.

Wie man sieht, überwiegt der räumlich strukturierte Effekt deutlich den räumlich unstrukturierten Effekt. Ein erhöhtes Mortalitätsrisiko weisen Mecklenburg-Vorpommern sowie das Saarland und Teile von Rheinland-Pfalz, Baden-Württemberg und Franken auf. Ein unterdurchschnittliches Mortalitätsrisiko besteht dagegen in Sachsen sowie Teilen von Sachsen-Anhalt, Thüringen, Brandenburg und Schleswig-Holstein. Die Beobachtung eines erhöhten Risikos in Mecklenburg-Vorpommern stimmt überein mit der Tatsache, dass Alkoholkonsum der bedeutendste Risikofaktor für Mundhöhlenkrebs ist und Mecklenburg-Vorpommern den größten Pro-Kopf-Verbrauch an Alkohol in Deutschland besitzt (Becker & Wahrendorf 1997). Nach Blot, Devesa, McLaughlin & Fraumeni Jr. (1994) weist der nordöstliche Teil Frankreichs entlang der deutschen Grenze die höchste Inzidenz bezüglich Mundhöhlenkrebs in Europa auf. Diese Region mit erhöhtem Risiko scheint sich nach den obigen Ergebnissen in Deutschland fortzusetzen.

Bei Betrachtung des räumlich unstrukturierten Effekts fällt insbesondere das erhöhte Risiko für einige Großstädte wie Berlin, Hamburg oder Bremen auf. Dies deutet auf unbeobachtete Risikofaktoren hin, die vermutlich mit dem höheren Urbanisierungsgrad zusammenhängen dürften.

Um einen Vergleich mit den Standardmortalitätsraten zu ermöglichen, sind in Abbildung 6.3 (a) die Schätzungen $\hat{r}_i = \exp(\hat{\beta}_0 + \hat{f}_{i,spat} + \hat{b}_i)$ abgebildet. Der lineare Prädiktor $\hat{\tilde{\eta}}_i = \hat{\beta}_0 + \hat{f}_{i,spat} + \hat{b}_i$ weist dabei eine Standardabweichung von 0.204 auf und liegt damit deutlich unter der Standardabweichung der logarithmierten Standardmortalitätsraten. Auch die minimale und maximale geschätzte Mortalitätsrate fallen mit 0.560 und 1.571 deutlich weniger extrem aus als bei den Standardmortalitätsraten. Hier zeigt sich die Glättungseigenschaft sowohl des Markov-Zufallsfeldes als auch des zufälligen Effekts.

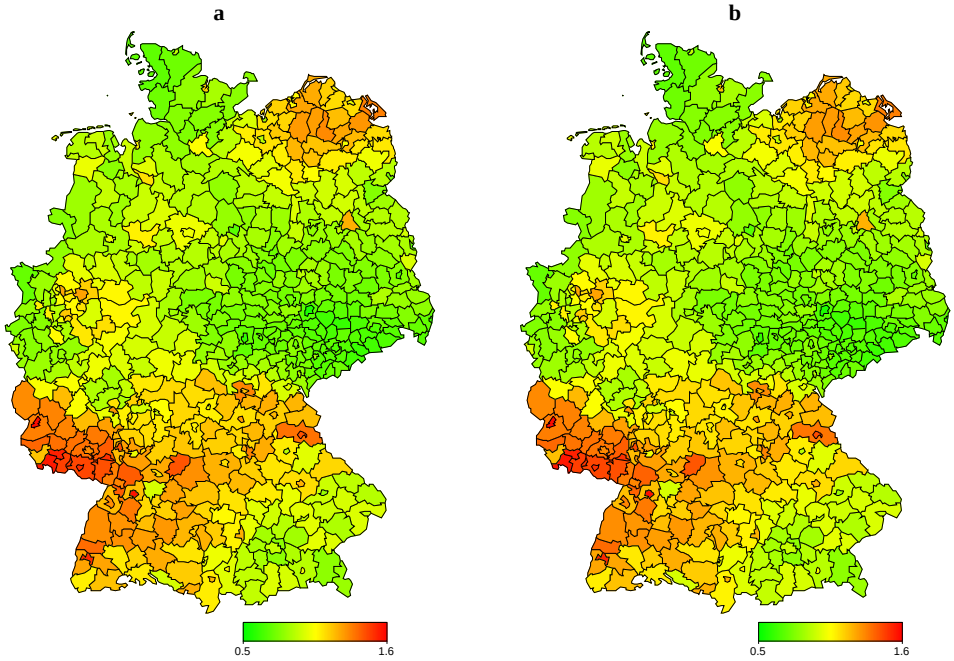

Abbildung 6.3: Geschätzte Mortalitätsraten $\hat{r}_i = \exp(\hat{\beta}_0 + \hat{f}_{i,spat} + \hat{b}_i)$ *aus der Schätzung mit ggamm (a) und geschätzte Mortalitätsraten* $\hat{r}_i$ *aus der vollen Bayes-Schätzung (b) .*

6.1.3 Vergleich mit voller Bayes-Schätzung

Zusätzlich zur Schätzung mit Hilfe der in Kapitel 3 beschriebenen Verfahren, die als empirischer Bayes-Ansatz betrachtet werden können, wurden die Daten auch mit `BayesX`, das heißt in einem voll-bayesianischen Ansatz analysiert. Problematisch ist hier die Wahl der Hyperparameter für die Prioris der Varianzparameter. Außerdem müssen Samplingpfade und Autokorrelationen betrachtet werden, um die Burnin-Phase und die notwendige Iterationszahl zu bestimmen. Man vergleiche hierzu beispielsweise die Ausführungen in Brezger (2000).

Im vorliegenden Beispiel wurden für die Priori-Verteilungen der Varianzparameter die Standardeinstellungen aus `BayesX` verwendet, das heißt inverse Gammaverteilungen mit Parametern $a = 1$ und $b = 0.005$. Die Burnin-Phase betrug 5000 Iterationen, die anschließende Schätzphase 50000 Iterationen mit Ausdünnungsparameter 50. Man erhält so eine Stichprobe von jeweils 1000 Zufallszahlen aus den marginalen Posterioris der gesuchten Parameter.

Betrachtet man die Samplingpfade des inversen Glättungsparameters τ_{spat} und

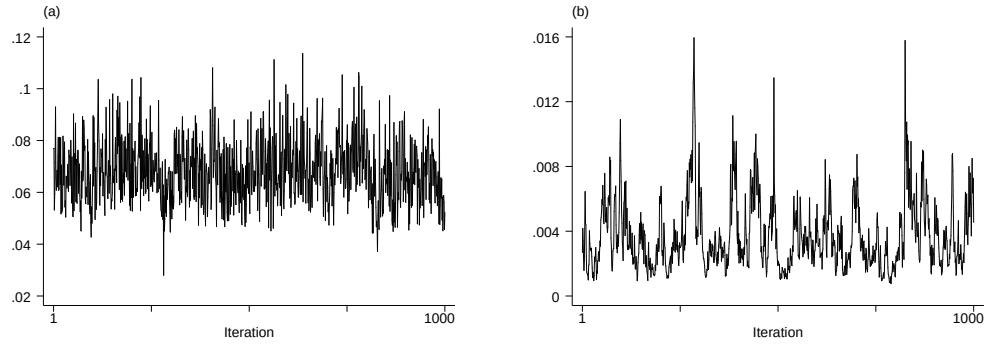

Abbildung 6.4: Samplingpfade des inversen Glättungsparameters des strukturierten Effekts (a) und der Varianz des unstrukturierten Effekts (b).

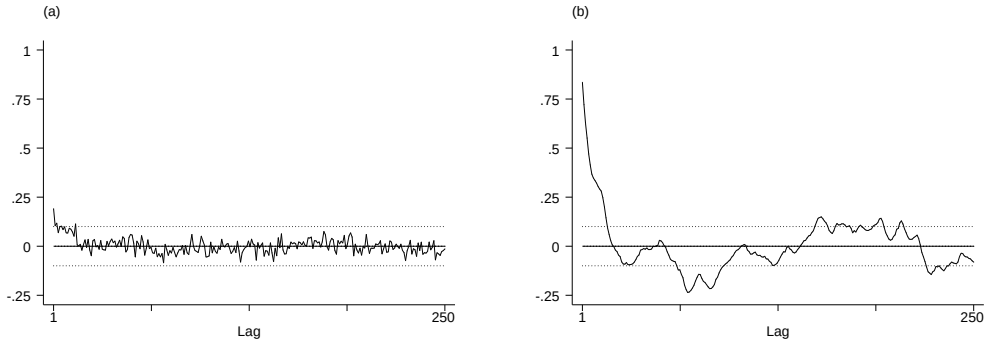

Abbildung 6.5: Autokorrelationsfunktionen des inversen Glättungsparameters des strukturierten Effekts (a) und der Varianz des unstrukturierten Effekts (b). Die gepunkteten Linien bezeichnen die Werte -0.1 und 0.1.

des Varianzparameter ν^2 (Abbildung 6.4), so scheint das Mixing im Falls von τ_{spat} durchaus akzeptabel, während der Verlauf für ν^2 ein deutlich schlechteres Mixing aufweist. Ähnlich verhält es sich mit den Autokorrelationsfunktionen der beiden Parameter (Abbildung 6.5): Während die Autokorrelation für τ_{spat} recht schnell gegen Null geht und dann zufällig um Null schwankt, weist die Autokorrelationsfunktion für ν^2 bereits einen wesentlich höheren Wert beim Lag 1 auf und zeigt auch darüberhinaus nicht das erwünschte Verhalten.

Die Punktschätzungen für den strukturierten und den unstrukturierten räumlichen Effekt stimmen, wie aus Abbildung 6.6 ersichtlich, mit den im vorigen Abschnitt erhaltenen Schätzungen weitgehend überein. Man erhält also, trotz der beschriebenen Probleme im vollen Bayes-Ansatz nahezu identische Ergebnisse. Im Gegensatz zu den empirischen Bayes-Schätzungen bleibt jedoch bei alleiniger Betrachtung des vollen Bayes-Ansatzes die Gültigkeit der Ergebnisse aufgrund

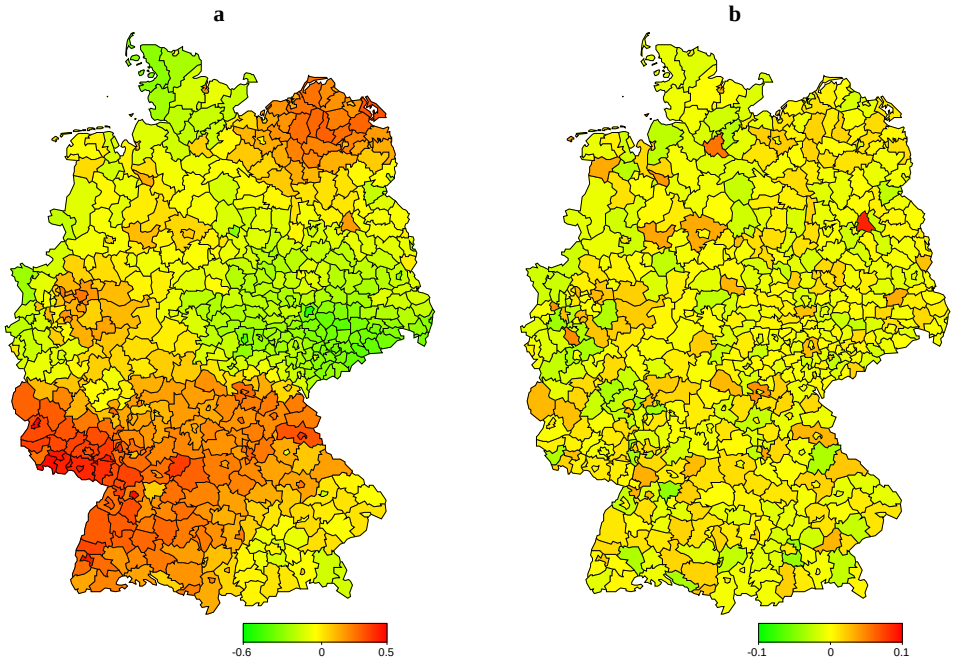

Abbildung 6.6: Strukturierter (a) und unstrukturierter (b) räumlicher Effekt aus der vollen Bayes-Schätzung mit BayesX.

des schlechten Mixings fraglich. Im empirischen Bayes-Ansatz werden dagegen Konvergenz- und Mixingprobleme völlig vermieden.

Häufig wird als Vorteil der vollen Bayes-Schätzung mit Hilfe von MCMC-Verfahren vorgebracht, dass nicht nur Schätzungen der ursprünglich interessierenden Parameter, sondern, bei Speicherung der Stichproben aus den marginalen Posterioris, auch Schätzungen des Posteriori-Erwartungswertes von Funktionalen dieser Parameter möglich sind. In dem in Kapitel 3 behandelten Ansatz ist man dafür auf das so genannte Plug-In-Verfahren angewiesen, das heißt, die Schätzungen der ursprünglich interessierenden Parameter werden in die entsprechenden Funktionale eingesetzt. Beispielsweise wurden die in Abbildung 6.3 (a) wiedergegebenen Schätzungen der Mortalitätsraten r_i basierend auf den Schätzungen für β_0, $f_{spat,i}$ und b_i durch Einsetzen in die Formel $r_i = \exp(\beta_0 + f_{spat,i} + b_i)$ bestimmt. Zum Vergleich wurden auch Schätzungen der Mortalitätsraten r_i aus dem vollen Bayes-Ansatz unter Verwendung der gespeicherten Samplingpfade berechnet. Diese Schätzungen der Posteriori-Erwartungswerte der r_i sind in Abbildung 6.3 (b) wiedergegeben und unterscheiden sich praktisch nicht von den aus dem Plug-In-Verfahren erhaltenen Schätzungen. Dennoch besitzen die voll-bayesianischen

Schätzer eine stärkere Rechtfertigung aufgrund ihrer Eigenschaft als Posteriori-Erwartungswerte der entsprechenden Parameter.

6.2 Waldschadensdaten

6.2.1 Datenbeschreibung

Als zweites Beispiel für die Anwendungsmöglichkeiten der in Kapitel 3 vorgestellten Modelle soll nun der Einfluss verschiedener Kovariablen auf den Schädigungsgrad von Bäumen im Forstgebiet Rothenbuch im Spessart untersucht werden. Das Beobachtungsgebiet umfasst einen Bereich von 15 Kilometern von Osten nach Westen und 10 Kilometern von Norden nach Süden mit 84 Beobachtungspunkten. Damit ist das Gebiet im Verhältnis zu den Gebieten, die beispielsweise in den Untersuchungen der Bayerischen Landesanstalt für Wald- und Forstwirtschaft betrachtet werden, relativ klein, ermöglicht hierdurch aber die Schätzung eines räumlich strukturierten Effektes, da die Beobachtungspunkte wesentlich näher zusammen liegen.

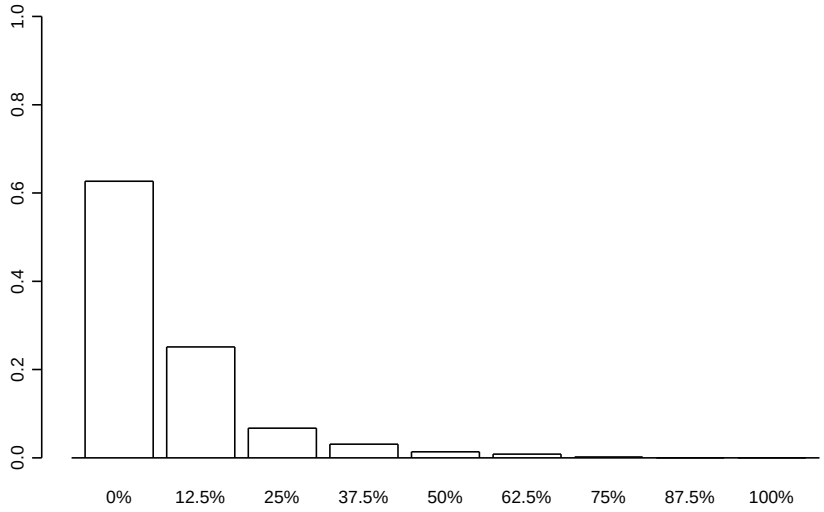

Abbildung 6.7: Häufigkeitsverteilung der verschiedenen Schädigungsstufen.

Analysiert werden soll lediglich die Schädigung einer Baumart, nämlich der Buche. Der Schädigungsgrad wird gemessen über den Entlaubungsgrad des Baumes, der im Frühjahr geschätzt wird, wenn noch die gesamte Baumkrone einsehbar ist, aber bereits erkennbar ist, wie viele Blätter der Baum später haben wird. Dies ist besonders bei der Buche wichtig, weil bei dieser die Schädigung im oberen

Teil der Baumkrone beginnt und sich dann nach unten fortsetzt. Ursprünglich wurde die Schädigung auf einer ordinalen Skala mit 9 möglichen Ausprägungen gemessen: Beginnend mit gesunden Bäumen (0% Schädigung) erstreckt sich die Skala in 12.5%-Schritten bis zur 100%igen Entlaubung. Abbildung 6.7 zeigt die Häufigkeitsverteilung dieser Variablen.

Zur Analyse wurde die ursprüngliche Variable zusammengefasst zu den zwei Kategorien ‚0' (0% Schädigung) und ‚1' (12.5% bis 100% Schädigung). Durch diese weitere Kategorisierung geht natürlich Information verloren, eine adäquate Analyse über multivariate kategoriale Modelle ist aber bisher mit den in Kapitel 3 vorgestellten Verfahren nicht möglich. Außerdem sind Bäume mit großem Entlaubungsgrad eher selten, so dass die Beschränkung auf die Unterscheidung gesunder und kranker Bäume keine allzu großen Auswirkungen auf die Schätzergebnisse haben sollte. Die Kategorie ‚0' tritt 971 mal auf, die Kategorie ‚1' 578 mal.

Abbildung 6.8: Zeitliche Entwicklung des Anteils geschädigter Bäume.

Als Kovariablen stehen die Kalenderzeit t, der Standort des i-ten Baumes S_i sowie das Bestandsalter A_{it} und der Beschirmungsgrad B_{it} des i-ten Baumes zum Zeitpunkt t zur Verfügung. Die Kalenderzeit wird in Jahren gemessen und reicht von 1983 bis 2001. Abbildung 6.8 zeigt die zeitliche Entwicklung des Anteils geschädigter Bäume. Wie man sieht erreicht dieser Anteil 1987 mit 57.5% seine maximale Ausprägung. Anschließend fällt er bis 1992 auf ein Minimum von circa 23% ab, um sich dann im Verlauf der 90er Jahre auf einem relativ konstanten Niveau von etwa 37.5% zu stabilisieren.

Es wurden Bäume an 84 Standorten beobachtet. In Abbildung 6.9 wird zum einen die Verteilung der Standorte über das Forstgebiet visualisiert und zum

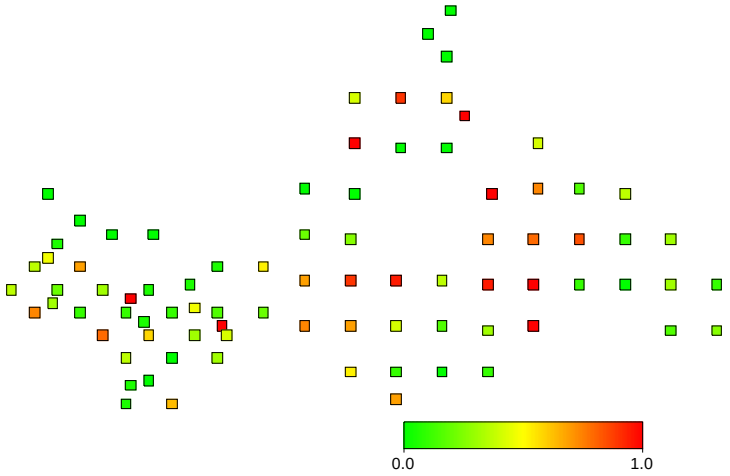

Abbildung 6.9: Anteil der Zeitpunkte zu denen ein Baum als geschädigt eingestuft wurde.

anderen der Anteil geschädigter Bäume pro Beobachtungspunkt wiedergegeben. Man beachte, dass es sich hierbei jeweils um den selben Baum handelt, dass also aufgetragen wird, zu wieviel Prozent der gemessenen Zeitpunkte der Baum als geschädigt eingestuft wurde. Die Lücke innerhalb der beobachteten Standpunkte entspricht der Lage der Ortschaft Rothenbuch.

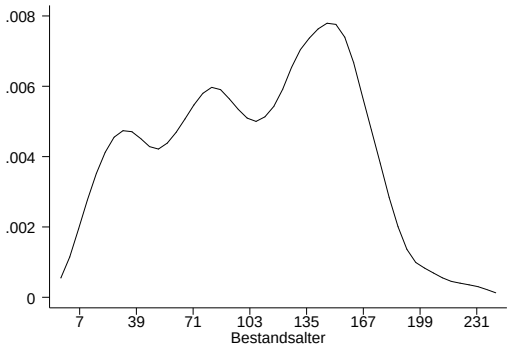

Abbildung 6.10: Kerndichteschätzer für die Verteilung des Bestandsalters.

Das Bestandsalter misst nicht das Alter des beobachteten Baumes, sondern das durchschnittliche Alter des diesen Baum umgebenden Bestands. Das Bestandsalter entwickelt sich also nicht linear mit der Kalenderzeit, sondern kann auch im folgenden Jahr geringer sein als im vorhergehenden, wenn viele ältere Bäume gefällt wurden. Der Wertebereich umfasst Bestandsalter von 7 Jahren bis 231 Jahren. Abbildung 6.10 zeigt die Verteilung des Bestandsalters mit Hilfe eines

Kerndichteschätzers.

Der Beschirmungsgrad misst die Dichte der Baumkrone eines Baumes. Konkret wird in Schritten von 10 Prozent gemessen, wieviel Sonnenlicht noch durch die Laubdecke fällt.

6.2.2 Schätzung

Bezeichne y_{it} die Schädigung des Baumes i ($i = 1, \ldots, 84$) zum Zeitpunkt t ($t = 1983, \ldots, 2001$). Dann modelliert man die Wahrscheinlichkeit für das Vorliegen einer Schädigung ($y_{it} = 1$) im Jahr t und für Baum i durch das folgende Logit-Modell:

$$\text{logit}(y_{it}) = \eta_{it} = \beta_0 + f_1(t) + f_2(A_{it}) + f_3(B_{it}) + f_{spat}(S_i).$$

Mit $\text{logit}(y_{it})$ bezeichnet man dabei die logarithmierte Chance

$$\log\left(\frac{\mathbb{P}(y_{it} = 1)}{\mathbb{P}(y_{it} = 0)}\right).$$

Die Funktionen f_1 bis f_3 werden wie in Kapitel 3 beschrieben als P-Splines vom Grad 3 mit zweiten Differenzen als Penalisierung modelliert, für den Effekt des Standorts des Baumes wird ein Markov-Zufallsfeld angenommen. Zur Schätzung des Markov-Zufallsfeldes werden zwei Bäume als benachbart betrachtet, wenn ihre Standorte weniger als 1.2 Kilometer voneinander entfernt sind. Im Gegensatz zum Disease-Mapping wird der räumliche Effekt nicht in zwei Komponenten aufgeteilt, weil relativ viele Bäume zu allen Zeitpunkten geschädigt oder zu allen Zeitpunkten nicht geschädigt sind (vergleiche Abbildung 6.9).

Abbildung 6.11 zeigt die geschätzten Funktionen f_1 und f_2. Beide Funktionen weisen einen deutlich nichtlinearen Verlauf auf, was die Modellierung als glatte Funktionen bestätigt. Wie man sieht, gibt der Effekt der Kalenderzeit den bereits aus den Rohdaten erkennbaren Trend wieder: Mitte der 80er Jahre erreicht die Schädigung den stärksten Grad, um danach bis zu Beginn der 90er Jahre auf ein Minimum abzufallen. Im Verlauf der 90er Jahre verweilt der Schädigungsgrad dann auf einem relativ konstanten Niveau.

Der Effekt des Bestandsalters erreicht ein erstes Maximum bei einem Alter von circa 65 Jahren, um anschließend bis zu einem Bestandsalter von 90 Jahren wieder abzufallen. Anschließend steigt die Wahrscheinlichkeit für eine Schädigung

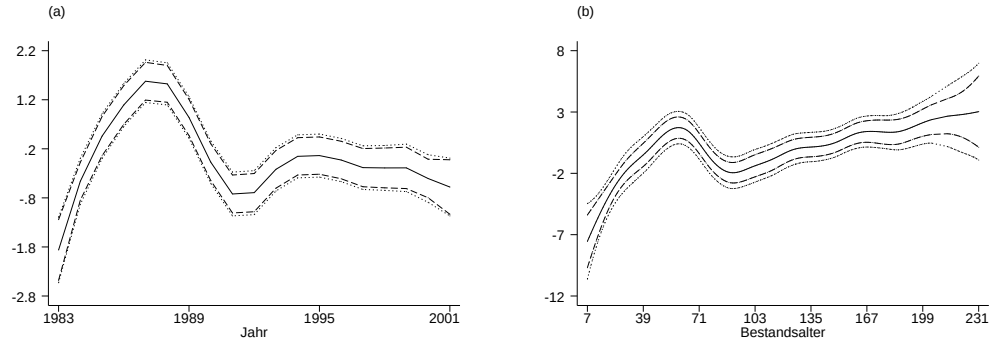

Abbildung 6.11: Effekt des Kalenderjahres (a) und des Bestandsalters (b) zusammen mit punktweisen 95%-Konfidenzbändern (frequentistisch (---) und bayesianisch (· · ·)).

wieder an, wenn auch nicht so steil wie im Bereich von 7 bis 65 Jahren, und erreicht zuletzt wieder das Niveau, das auch bei einem Bestandsalter von 65 Jahren erreicht wurde.

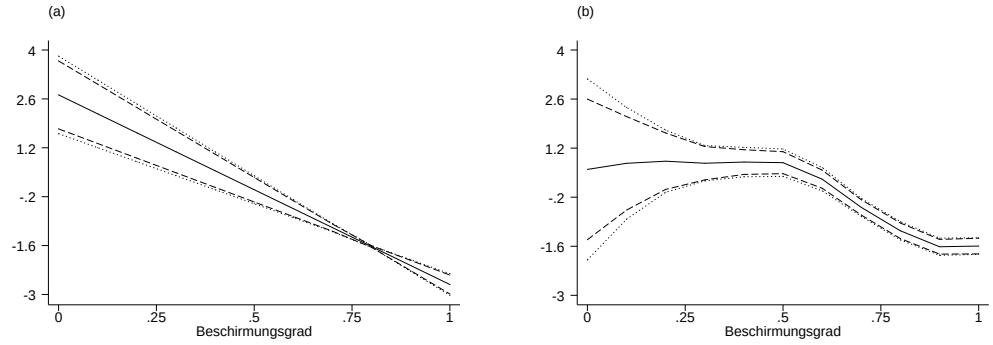

Abbildung 6.12: Effekt des Beschirmungsgrads bei Berücksichtigung des räumlichen Effekts (a) und ohne räumlichen Effekt (b) zusammen mit punktweisen 95%-Konfidenzbändern (frequentistisch (---) und bayesianisch (· · ·)).

Wie aus Abbildung 6.12 (a) ersichtlich ist, wird der Effekt des Beschirmungsgrads nahezu linear geschätzt. Der Regressionsparameter des Linearanteils von f_3 beträgt -5.4, das heißt mit steigendem Beschirmungsgrad ist eine geringere Wahrscheinlichkeit für die Schädigung eines Baumes zu beobachten. Der Verlauf der Konfidenzbänder, die im Bereich des mittleren Beschirmungsgrads von 0.78 sehr nahe an die Funktionsschätzung heranreichen, resultiert dabei aus der an die Funktionen gestellten Zentrierungsbedingung.

Abbildung 6.13 zeigt die Schätzung des räumlich strukturierten Effekts. Wie zu

erkennen ist, sind die Bäume in der Umgebung der Ortschaft Rothenbuch (rund um die Lücke innerhalb der Standorte) stärker geschädigt als die übrigen Bäume. Dies deutete sich auch schon in Abbildung 6.9 an, dort war der räumliche Trend aber weniger deutlich ausgeprägt.

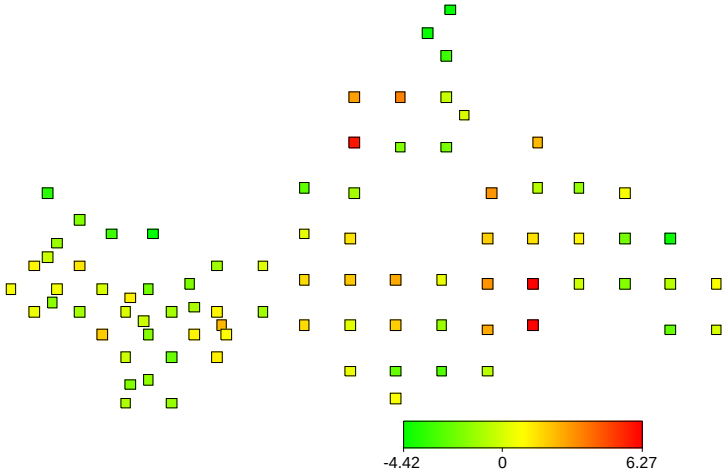

Abbildung 6.13: Strukturierter räumlicher Effekt.

Um die Notwendigkeit der Einbeziehung eines räumlichen Effektes zu überprüfen, wurden Klassifikationstabellen für das Modell mit räumlichem Effekt und ein Modell, das auf die räumliche Komponente verzichtet, bestimmt. Die Ergebnisse sind in Tabelle 6.2 wiedergegeben.

	$\hat{y}_{it}$	
y_{it}	0	1
0	900	71
1	113	465

	$\hat{y}_{it}$	
y_{it}	0	1
0	846	125
1	207	371

Tabelle 6.2: Klassifikation bei Analyse mit und ohne räumliche Komponente.

Wie man sieht, ist die Klassifikation bei Berücksichtigung des räumlichen Effekts deutlich besser. Dies drückt sich auch in den Fehlklassifikationsraten von 11.9% (mit räumlicher Komponente) beziehungsweise 21.4% (ohne räumliche Komponente) aus. Die räumliche Komponente ist also zur adäquaten Modellierung erforderlich und das Modell mit räumlichem Effekt stellt einen deutlichen Fortschritt gegenüber einem einfachen generalisierten additiven Modell dar. Man beachte auch, dass ohne die räumliche Komponente der Effekt des Beschirmungsgrads noch deutlich nichtlinear ist, wie in Abbildung 6.12 (b) zu sehen ist. Eine Ver-

nachlässigung des räumlichen Effekts hätte also auch hier zu einer fehlerhaften Interpretation geführt.

6.3 Mietspiegel München

6.3.1 Datenbeschreibung

Als drittes Beispiel sollen nun noch die dem Münchner Mietspiegel zugrunde liegenden Daten untersucht werden. Ziel der Erstellung von Mietspiegeln ist es, die ortsüblichen Vergleichsmieten zu bestimmen, nach denen sich die Miethöhe zu richten hat. Üblicherweise werden Mietspiegel in der Form von Tabellen angegeben, aus denen sich, ausgehend von gewissen Merkmalen einer Wohnung, die interessierenden Vergleichsmieten ablesen lassen. Die Erstellung von Mietspiegeln über die Bildung von Mittelwerten in Teilgruppen mit bestimmten Kovariablenkombinationen ist dabei problematisch, weil hierfür eine sehr große Stichprobe erforderlich ist, die häufig nahe an einer Vollerhebung liegen kann. Daher bietet sich die Erstellung eines Mietspiegels mit Hilfe von Regressionsverfahren an, da so die gesamte Information einer Stichprobe ausgenutzt werden kann und sich aus den erhaltenen Ergebnissen wieder ein Tabellenmietspiegel erstellen lässt.

Der Münchner Mietspiegel basiert auf eine Stichprobe von 3082 Wohnungen, so dass die Analyse dieser Daten, aufgrund der in Kapitel 2.2 beschriebenen Beschränkungen des Stichprobenumfangs, am oberen Rand der mit Hilfe der Funktion `ggamm` numerisch noch bewältigbaren Problemstellungen liegt.

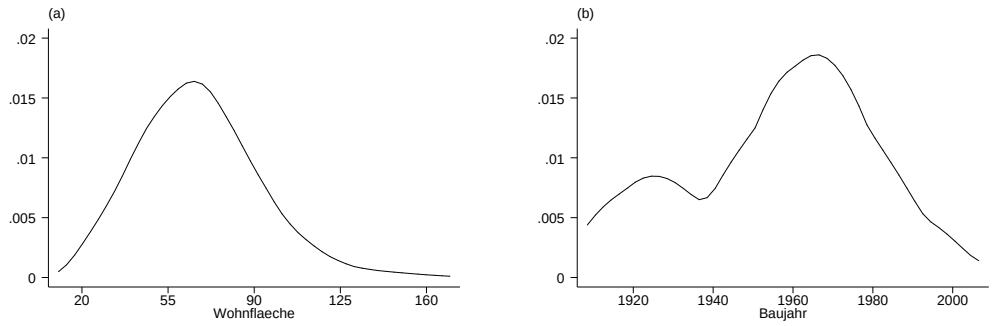

Abbildung 6.14: Kerndichteschätzer für die Verteilung der Wohnfläche (a) und des Baujahrs (b) in der Stichprobe.

Als abhängige Variable wird in der Analyse des Mietspiegels die Nettomiete pro

Quadratmeter ($nmqm$) verwendet. Als metrische Einflussgrößen sind die Wohnfläche einer Wohnung in Quadratmetern (wfl) sowie das Baujahr des Hauses, in dem sich die Wohnung befindet (bj), bekannt. Abbildung 6.14 zeigt für beide Variablen die Verteilung in der Stichprobe mit Hilfe von Kerndichteschätzern.

Als weitere Information ist die Zugehörigkeit der Wohnungen zu den 380 Münchner Bezirksvierteln (B) vorhanden, so dass die Schätzung eines räumlich strukturierten Effektes möglich ist. Man vergleiche Abbildung 3.2 in Kapitel 3.1, in der die mittleren Mieten pro Bezirksviertel abgebildet sind. Durch die Berücksichtigung eines räumlich strukturierten Effekts lassen sich beispielsweise unbeobachtete Charakteristika der Lage einer Wohnung auffangen.

Zusätzlich sind eine Reihe kategorialer Variablen gegeben, die Eigenschaften der jeweiligen Wohnung beschreiben und die in Dummy-Kodierung in das Modell eingehen. Beispiele hierfür sind das Vorhandensein starker Verkehrsbelastung ($verkehr$) oder die Ausstattung des Bades einer Wohnung ($besbad$ = besondere Zusatzausstattung, $kbad$ = kein Bad vorhanden, $zwbad$ = zweites Bad vorhanden, $kbadkach$ = kein gekacheltes Bad). Insgesamt sind 27 solche kategorialen Einflussgrößen vorhanden, die zwar im Modell berücksichtigt werden, auf die aber im Folgenden nicht intensiver eingegangen wird.

Als weitere kategoriale Kovariable ist die Einstufung der Lage einer Wohnung in drei Kategorien bekannt. Die Einstufung innerhalb dieser Kategorien erfolgte durch Experten einer Arbeitsgruppe der Stadt München und besitzt die drei möglichen Ausprägungen durchschnittliche Lage, gute Lage und sehr gute Lage. Wohnungen schlechter Wohnlage kamen nur so selten in der erhobenen Stichprobe vor, dass sie vorab aus der Analyse ausgeschlossen wurden. Aus dieser Kovariablen ergeben sich die beiden Dummy-Variablen $lagegut$, mit $lagegut = 1$ für Wohnungen guter Wohnlage, und $lagesgut$, mit $lagesgut = 1$ für Wohnungen sehr guter Wohnlage. Von Interesse ist es nun, zu überprüfen, wie gut die Einstufung der Wohnungen in die drei Kategorien erfolgt, beziehungsweise inwieweit die räumliche Variation der Nettomieten pro Quadratmeter auf Kennzeichen, die dieser Einstufung zugrunde liegen, zurückgeführt werden kann. Um diese Fragen beurteilen zu können, soll nun zunächst ein Modell, das zwar einen räumlich strukturierten Effekt, aber nicht die Variablen $lagegut$ und $lagesgut$ enthält, geschätzt werden und dann mit einem Modell, das sowohl den räumlichen Effekt als auch $lagegut$ und $lagesgut$ enthält, verglichen werden.

6.3.2 Modell ohne Experteneinschätzung der Wohnlage

Ohne die Variablen *lagegut* und *lagesgut* erhält man für die Nettomiete pro Quadratmeter das Modell

$$\eta_i = \beta_0 + f_1(wfl_i) + f_2(bj_i) + f_{spat}(B_i) + \beta_1 verkehr_i + \beta_2 besbad_i + \ldots,$$

das sich unter der zusätzlichen Annahme $nmqm_i \sim \mathrm{N}(\eta_i, \sigma^2)$ mit Hilfe der in Kapitel 3 beschriebenen Verfahren bestimmen lässt.

Häufig tritt jedoch in Datensituationen mit normalverteiltem Response das Problem heteroskedastischer Varianzen auf, das heißt, für die bedingte Kovarianzmatrix des Responsevektors y gilt nicht mehr $\mathrm{Var}(y|b) = \sigma^2 I_n$, sondern $\mathrm{Var}(y|b) = \Sigma$ mit einer positiv definiten Kovarianzmatrix Σ. Oft unterstellt man, dass es sich auch bei Σ um eine Diagonalmatrix handelt, so dass man $\Sigma = \sigma^2 \,\mathrm{diag}(\omega_1, \ldots, \omega_n)$ erhält. Sind die Gewichte ω_i vorab bekannt, so können sie in der Schätzung berücksichtigt werden, indem in den individuellen Likelihood-Beiträgen die entsprechenden Gewichte eingesetzt werden. Man vergleiche hierzu die Darstellung der Dichte einer einfachen Exponentialfamilie in (2.10).

In der Regel sind die Gewichte jedoch nicht vorab bekannt, so dass sie mit Hilfe der Residuen $\hat{\varepsilon}_i = y_i - \hat{y}_i$ geschätzt werden müssen. Dabei geht man davon aus, dass die Gewichte umgekehrt proportional zu den quadrierten Residuen sind, das heißt, man nimmt an, dass

$$\omega_i \propto \frac{1}{\hat{\varepsilon}_i^2}$$

gilt. Da eine direkte Bestimmung über die Residuen jedoch zu einem überparametrisierten Modell führt, schätzt man die Gewichte als vorhergesagte Werte aus einem Regressionsmodell, das die quadrierten Residuen als abhängige Variable besitzt. Genauer benutzt man häufig eine geeignete Transformation r_i der quadrierten Residuen, um für die vorhergesagten Werte positive Werte zu erhalten. Als Kovariablen werden die gleichen Variablen verwendet, die auch im eigentlich interessierenden Modell betrachtet werden. Bei Verwendung des Logarithmus als Transformation erhält man so den folgenden Algorithmus zur Bestimmung der Gewichte:

Algorithmus 7 (Bestimmung der Gewichte ω_i)

(i) Bestimme die Residuen $\hat{\varepsilon}_i = y_i - \hat{y}_i$ aus der ungewichteten Schätzung.

(ii) Berechne $r_i = \log(\hat{\varepsilon}_i^2)$ und schätze ein Regressionsmodell mit r_i als abhängiger Variable.

(iii) Die geschätzten Gewichte ergeben sich als $\hat{\omega}_i = 1/\exp(\hat{r}_i)$ mit den aus der Schätzung in (ii) erhaltenen vorhergesagten Werten $\hat{r}_i$.

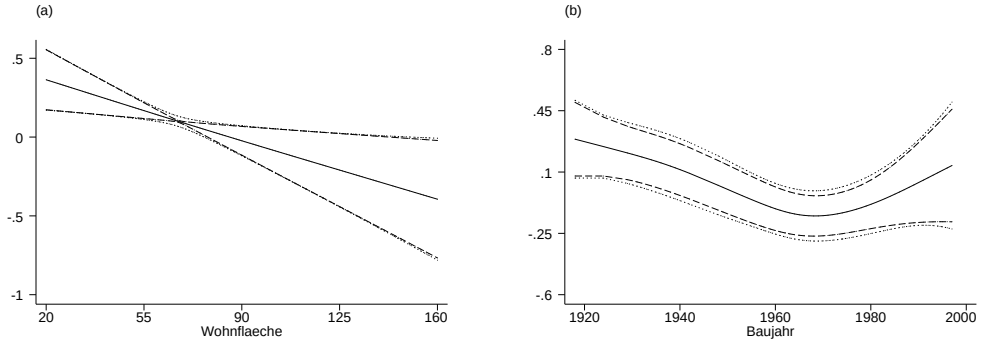

Abbildung 6.15: Effekt der Wohnfläche (a) und des Baujahrs (b) in der Schätzung der Gewichte ω_i zusammen mit punktweisen 95%-Konfidenzbändern (frequentistisch (---) und bayesianisch ($\cdots$)).

Die Anwendung des Algorithmus besitzt auch den Vorteil, überprüfen zu können, ob im untersuchten Modell die Annahme homoskedastischer Varianzen verletzt ist. Lassen sich in der Regression für r_i keine signifikanten Einflüsse nachweisen, so kann man davon ausgehen, dass für die Varianzen die Annahme der Homoskedastizität erfüllt ist.

Nun soll das beschriebene Verfahren auf die Analyse der Mietspiegeldaten angewendet werden. In den Abbildungen 6.15 und 6.16 sind die aus dem Modell für die logarithmierten, quadrierten Residuen r_i resultierenden Schätzungen der nonparametrischen Effekte und des räumlichen Effekts wiedergegeben. Wie man sieht, erhält man relativ deutliche Hinweise auf heteroskedastische Varianzen, so dass die Schätzung des Modells für die Nettomiete pro Quadratmeter durch einen gewichteten Ansatz ratsam erscheint. Betrachtet man insbesondere die räumliche Schätzung in Abbildung 6.16, so erkennt man, dass offensichtlich für Wohnungen im Innenstadtbereich und in Schwabing eine größere Variation der Nettomieten pro Quadratmeter vorhanden ist als in den Randgebieten Münchens.

Die gewichtete Schätzung des Modells mit der Nettomiete pro Quadratmeter als abhängiger Variable führt zu den in Abbildung 6.17 wiedergegebenen Schätzungen der nonparametrischen Effekte. Wie zu erwarten war, nimmt die Nettomiete

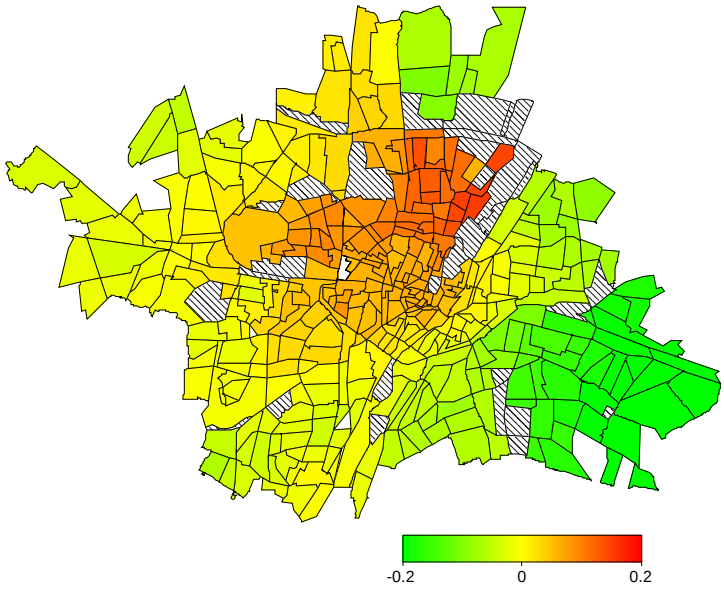

Abbildung 6.16: Strukturierter räumlicher Effekt in der Schätzung der Gewichte ω_i.

pro Quadratmeter mit steigender Wohnfläche ab. Die Abhängigkeit vom Baujahr verläuft über den Zeitraum von 1920 bis circa 1950 annähernd horizontal, um dann stark anzusteigen. Während also für Wohnungen in älteren Häusern die Nettomiete nahezu vom Baujahr unabhängig ist, besitzen Wohnungen in neueren Häusern eine deutlich vom Baujahr abhängige Miete. Außerdem erkennt man anhand der Wertebereiche der geschätzten Effekte, dass das Baujahr im Verhältnis zur Wohnfläche einen deutlich geringeren Einfluss auf die Nettomiete besitzt.

Abbildung 6.18 zeigt den räumlichen Effekt der Lage einer Wohnung innerhalb Münchens. Auffällig sind dabei die Bezirke mit deutlich unterdurchschnittlichen Nettomieten im Münchner Norden, sowie die erhöhten Mieten im Osten Münchens. Ebenfalls überdurchschnittlich sind, wie zu erwarten war, die Mieten im Zentrum Münchens.

6.3.3 Modell mit Experteneinschätzung der Wohnlage

Um nun die Qualität der Experteneinschätzung der Wohnlagen der Münchner Wohnungen zu überprüfen, wurde das obige Modell für die Nettomiete pro Quadratmeter um die Variablen *lagegut* und *lagesgut* erweitert. Man erhält damit

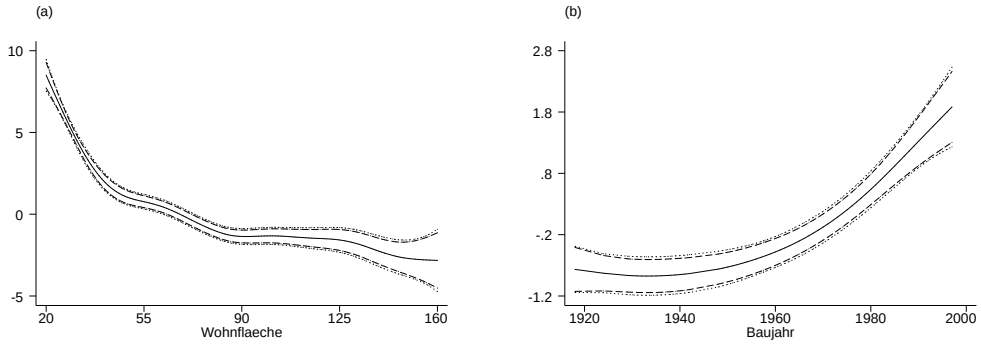

Abbildung 6.17: Effekt der Wohnfläche (a) und des Baujahrs (b) ohne Berücksichtigung der Experteneinschätzung der Wohnlage zusammen mit punktweisen 95%-Konfidenzbändern (frequentistisch (---) und bayesianisch (· · ·)).

den linearen Prädiktor

$$\eta_i = \beta_0 + f_1(wfl_i) + f_2(bj_i) + f_{spat}(B_i) + \beta_1 lagegut_i + \beta_2 lagesgut_i + \ldots$$

Wie zuvor wurden auch für dieses Modell entsprechend Algorithmus 7 Gewichte bestimmt und dann die gewichtete Schätzung durchgeführt. Die Ergebnisse dieser Schätzung sind für die nonparametrischen Effekte in Abbildung 6.19 und für den räumlichen Effekt in Abbildung 6.20 wiedergegeben. Wie man sieht, ändern sich die Schätzungen der Effekte der Wohnfläche und des Baujahrs praktisch nicht. Für den räumlichen Effekt bleibt zwar der grundsätzliche Verlauf, mit niedrigeren Mieten im Norden sowie höheren Mieten im Osten und im Zentrum Münchens erhalten, der Effekt fällt aber deutlich geringer aus. Die Einschätzung der Wohnlage über die beiden Variablen *lagegut* und *lagesgut* erklärt also offenbar einen Teil der räumlichen Variation, erfasst aber wohl noch nicht alle Kennzeichen, so dass die Schätzung eines strukturierten räumlichen Effekts nützlich sein kann, um auch diese geeignet zu berücksichtigen.

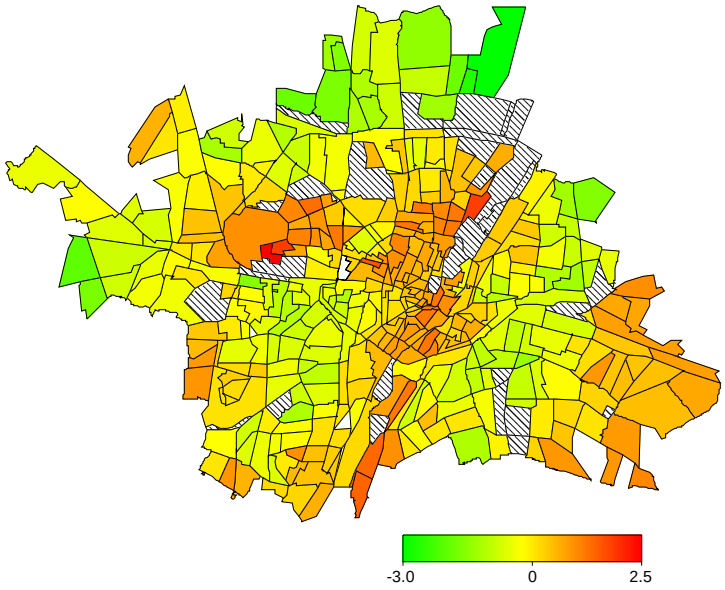

Abbildung 6.18: Strukturierter räumlicher Effekt ohne Berücksichtigung der Experteneinschätzung der Wohnlage.

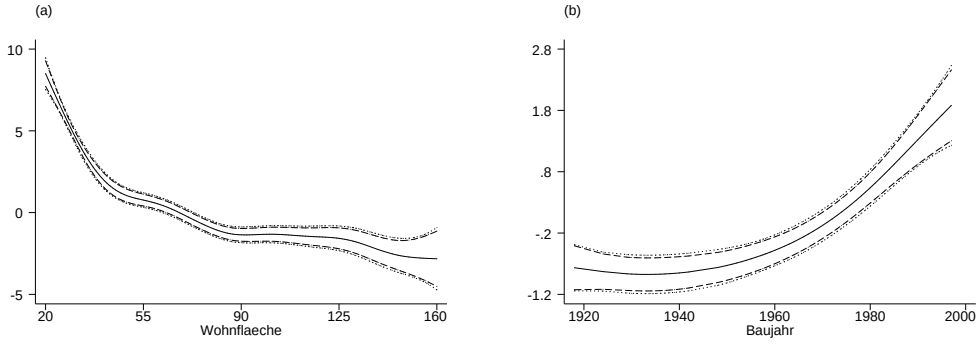

Abbildung 6.19: Effekt der Wohnfläche (a) und des Baujahrs (b) bei Berücksichtigung der Experteneinschätzung der Wohnlage zusammen mit punktweisen 95%-Konfidenzbändern (frequentistisch (---) und bayesianisch (· · ·)).

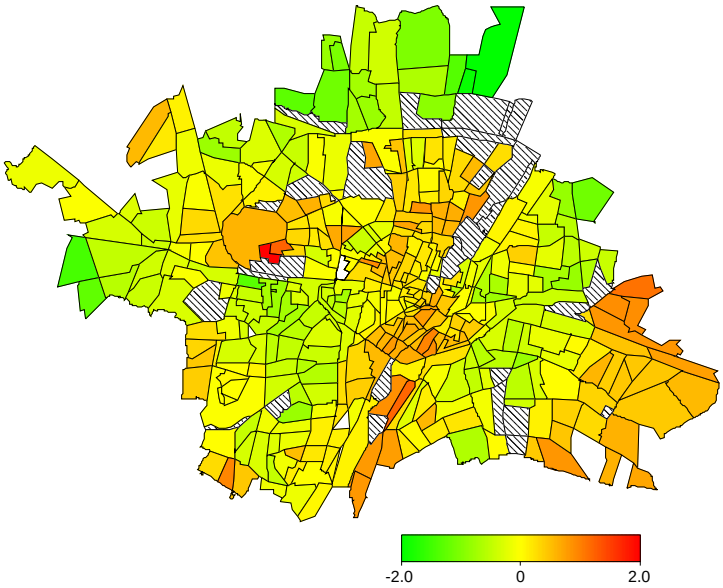

Abbildung 6.20: Strukturierter räumlicher Effekt bei Berücksichtigung der Experteneinschätzung der Wohnlage.

7 Zusammenfassung und Ausblick

In dieser Arbeit wurde eine Möglichkeit vorgestellt, generalisierte geoadditive gemischte Modelle mit Hilfe von P-Splines und Markov-Zufallsfeldern zu bestimmen. Dabei war es durch eine geeignete Reparametrisierung des Modells nicht nur möglich, Schätzungen der Regressionskoeffizienten der einzelnen Modellkomponenten zu erhalten, sondern auch die Hyperparameter des Modells, also Glättungs- und Varianzparameter simultan zu bestimmen. Damit ist die vollständige Schätzung einer weiten Klasse von Regressionsmodellen möglich, die nicht nur die nonparametrische Modellierung des Einflusses metrischer Kovariablen, sondern auch die adäquate Einbeziehung von Korrelationen zwischen den Beobachtungen, beispielsweise basierend auf räumlichen Informationen, erlauben.

Wie man im Rahmen der Simulationsstudien in Kapitel 5 gesehen hat, ist das resultierende Verfahren im Hinblick auf die Schätzqualität einigen weiteren Möglichkeiten zur Bestimmung generalisierter additiver Modelle zumindest ebenbürtig, bei normalverteiltem Response und niedrigem Signal-Rauschen-Verhältnis sowie für poissonverteilten Response gehört es sogar zu den besten der getesteten Verfahren. Darüber hinaus ist die Anwendung eben nicht auf generalisierte additive Modelle beschränkt, sondern erlaubt auch die Einbeziehung räumlicher und zufälliger Effekte. Diese Flexibilität des vorgestellten Ansatzes wurde auch durch die Analyse realer Daten in äußerst verschiedenen Fragestellungen in Kapitel 6 unterstrichen.

Prinzipiell ist die in Kapitel 3.3 vorgestellte Reparametrisierung sogar auf beliebige Penalisierungsansätze anwendbar. Man erhält so die Möglichkeit, die beschriebenen Modelle um weitere Komponenten zu erweitern, deren Schätzung ebenfalls auf der Maximierung einer penalisierten Likelihood beruht. Ein Beispiel hierfür ist etwa die Verallgemeinerung von P-Splines auf die Schätzung von Oberflächen, so dass die Modellierung von Interaktionen metrischer Kovariablen möglich wird. Ein solcher Ansatz ist in Lang & Brezger (2002) beschrieben und könnte im Rahmen zukünftiger Aktivitäten der bestehenden Implementation hinzugefügt werden. Ein weiteres Beispiel ist die Berücksichtigung flexibler Saisonkomponenten zur Analyse von Longitudinaldaten, wie sie etwa in Knorr-Held (1996), Kapitel 4.1 oder Fahrmeir & Tutz (2001) Kapitel 8.1 beschrieben werden.

Eingeschränkt wird die Anwendbarkeit der vorgestellten Verfahren durch den maximalen Stichprobenumfang, für den die notwendigen Berechnungen numerisch noch durchführbar sind. Eventuell lassen sich diese Beschränkungen durch eine vorteilhaftere Bestimmung der zur REML-Schätzung verwendeten Größen aufheben oder zumindest nach oben verschieben. Eine verhältnismäßig einfache Möglichkeit, die Berechnungen zu beschleunigen, ohne jedoch die Bestimmung von $n \times n$-Matrizen vollständig vermeiden zu können, bestünde eventuell in der Approximation der erwarteten Fisher-Informationsmatrix durch eine Diagonalmatrix. Hiermit ließe sich zumindest die Zahl der zu berechnenden $n \times n$-Matrizen deutlich verringern. Eine solche Approximation müsste jedoch durch zusätzliche Simulationsstudien untersucht werden, um eventuelle Verschlechterungen der Schätzqualität feststellen zu können, so dass im Rahmen dieser Arbeit darauf verzichtet werden musste.

Ein weiterer Punkt, der noch ausführlicher zu untersuchen wäre, sind die in Kapitel 4 behandelten Testmöglichkeiten, die in einfachen Modellen beispielsweise erlaubten, den Zusammenhang zwischen einer Kovariablen und der abhängigen Variablen auf Linearität zu testen. Hier wäre eine Erweiterung auf Modelle mit mehreren Varianzkomponenten und nicht normalverteilten Daten wünschenswert. Da dies kaum oder nur unter großem Aufwand basierend auf der vorgestellten Theorie möglich zu sein scheint, wäre eine bayesianische Betrachtung des Testproblems interessant. Durch die Verwendung einer modifizierten Priori-Verteilung für die Varianzparameter, die es mit positiver Wahrscheinlichkeit erlaubt, einzelne Varianzparameter als 0 zu schätzen, wäre in einem solchen Ansatz prinzipiell die Bestimmung der Posteriori-Wahrscheinlichkeiten für Nullhypothese und Alternative möglich.

Neben den bereits im Rahmen dieser Arbeit vorgestellten Modellen und den oben angesprochenen weiteren Penalisierungsansätzen, wären zusätzlich noch verschiedenartige Erweiterungen der vorgestellten Verfahren denkbar. Eine Möglichkeit, den Umfang der schätzbaren Modelle wesentlich zu vergrößern bestünde etwa in der Verallgemeinerung auf Modelle mit variierenden Koeffizienten, wie sie in Hastie & Tibshirani (1993) vorgeschlagen wurden. Zwei weitere mögliche Erweiterungen betreffen die zugelassenen Verteilungen der Response-Variablen. In Kapitel 2.1 wurde für normalverteilten Response stets davon ausgegangen, dass für die bedingte Kovarianzmatrix $\mathrm{Var}(y|b) = \sigma^2 I_n$ gilt. In Kapitel 6.3 wurde zwar

eine Möglichkeit vorgestellt, diese Annahme etwas zu lockern, interessant wäre aber dennoch die Verallgemeinerung zu $\mathrm{Var}(y|b) = \Sigma$ mit einer positiv definiten Kovarianzmatrix Σ, die geeignet über einen Vektor von Varianzparametern parametrisiert werden kann. Dann könnten nämlich die Varianzparameter simultan per Restricted-Maximum-Likelihood-Schätzung bestimmt werden, statt ein zweistufiges Verfahren zur Schätzung geeigneter Gewichte anwenden zu müssen.

Ebenfalls wünschenswert wäre eine Erweiterung der Modellierungsmöglichkeiten auf multivariate Modelle, wie sie etwa in Fahrmeir & Lang (2001a, 2001b) behandelt werden. Diese Erweiterung würde es beispielsweise erlauben, die in Kapitel 6.2 analysierten Waldschadensdaten in einer genaueren Kategorisierung, das heißt mit drei oder mehr geordneten Kategorien als abhängiger Variablen, zu untersuchen. Da der numerische Aufwand für eine solche Schätzung wesentlich größer ist als in univariaten Modellen, wäre hierfür die Implementation der behandelten Verfahren in einer Programmiersprache wie beispielsweise C++ vorteilhaft.

Anhang

A REML-Schätzung

Im Folgenden sollen einige in Kapitel 2.2 präsentierte Resultate in Bezug auf die Schätzung von Varianzparametern in einem linearen gemischten Modell mit Hilfe der Restricted-Likelihood detailliert hergeleitet werden. Zunächst werden dazu in A.1 einige Definitionen und Sätze beziehungsweise Rechenregeln zusammengestellt, die später benötigt werden. In A.2 wird dann die Verteilung der Fehlerkontraste betrachtet bevor in den Abschnitten A.3 bis A.5 die zur Schätzung von $\hat{\vartheta}_{REML}$ notwendigen Größen bestimmt werden.

A.1 Definitionen und Rechenregeln

Transformationssatz für Dichten: Sei y ein n-dimensionaler Zufallsvektor mit Dichte $f(y) = f(y_1, \ldots, y_n)$. Weiterhin sei $T : \mathbb{R}^n \to \mathbb{R}^n$ mit $y \mapsto T(y) = z$ eine in jeder Variablen partiell differenzierbare Abbildung. Dann existiert die Jacobische Funktionalmatrix

$$\Delta_T(y) = \begin{pmatrix} \frac{\partial T_1}{\partial y_1} & \cdots & \frac{\partial T_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial T_n}{\partial y_1} & \cdots & \frac{\partial T_n}{\partial y_n} \end{pmatrix}.$$

Ist T eineindeutig und stetig differenzierbar auf ganz $\mathbb{R}^n$ mit nirgends verschwindender Funktionaldeterminante $|\Delta_T(y)|$, so gilt für die Dichte $g(z)$ des Zufallsvektors $z = T(y)$

$$g(z) = \frac{f(T^{-1}(z))}{\mathrm{abs}(|\Delta_T(T^{-1}(z))|)}. \tag{A.1}$$

Ist speziell $T(y) = Ay$ eine lineare Abbildung mit einer invertierbaren $n \times n$-Matrix A, so vereinfacht sich (A.1) zu:

$$g(z) = \frac{f(A^{-1}z)}{\mathrm{abs}(|A|)}. \tag{A.2}$$

Rechenregeln für Determinanten: Sei A eine $n \times n$-Matrix. Dann gilt:

$$|A| = |A'|, \tag{A.3}$$

$$|A'A| = |AA'| = |A|^2. \tag{A.4}$$

Lässt sich A zerlegen zu

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

mit $|A_{11}| \neq 0$, so gilt

$$|A| = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}|. \tag{A.5}$$

(Toutenburg (2003) Seite 479/480)

Erwartungswert quadratischer Formen: Seien y ein n-dimensionaler Zufallsvektor und A eine $n \times n$-Matrix. Dann gilt

$$\mathbb{E}(y'Ay) = \text{spur}(A\,\text{Var}(y)) + \mathbb{E}(y)'A\,\mathbb{E}(y). \tag{A.6}$$

Insbesondere gilt mit $\mathbb{E}(y) = 0$

$$\mathbb{E}(y'Ay) = \text{spur}(A\,\text{Var}(y)). \tag{A.7}$$

(Pruscha (2000) Seite 105)

Spur des Produkts zweier Matrizen: Seien A eine $n_1 \times n_2$-Matrix und B eine $n_2 \times n_1$-Matrix. Dann gilt:

$$\text{spur}(AB) = \text{spur}(BA) \tag{A.8}$$

Ableitung einer Matrix nach einem Skalar: Sei $A = A(x)$, $x \in \mathbb{R}$ eine $n \times n$-Matrix, deren Einträge funktional von einem Skalar x abhängen. Dann definiert man

$$\frac{\partial A(x)}{\partial x} = \begin{pmatrix} \frac{\partial a_{11}(x)}{\partial x} & \cdots & \frac{\partial a_{1n}(x)}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_{n1}(x)}{\partial x} & \cdots & \frac{\partial a_{nn}(x)}{\partial x} \end{pmatrix}.$$

Ableitung der Inversen: Sei $A(x)$ eine invertierbare Matrix. Dann gilt

$$\frac{\partial A(x)^{-1}}{\partial x} = -A(x)^{-1}\frac{\partial A(x)}{\partial x}A(x)^{-1}. \tag{A.9}$$

(McCulloch & Searle (2001) Seite 298)

Ableitung der Spur einer Matrix: Es gilt

$$\frac{\partial \operatorname{spur}(A(x))}{\partial x} = \operatorname{spur}\left(\frac{\partial A(x)}{\partial x}\right). \tag{A.10}$$

(Toutenburg (2003) Seite 517/18)

Ableitung der logarithmierten Determinante einer Matrix: Sei $A(x)$ eine invertierbare Matrix. Dann gilt

$$\frac{\partial \log(|A(x)|)}{\partial x} = \operatorname{spur}\left(A(x)^{-1}\frac{\partial A(x)}{\partial x}\right). \tag{A.11}$$

(McCulloch & Searle (2001) Seite 299)

A.2 Verteilung der Fehlerkontraste

Nun soll die Verteilung der wie in Kapitel 2.2 definierten Fehlerkontraste $u = A'y$ hergeleitet werden. Dazu wird im Folgenden die Abhängigkeit der Kovarianzmatrix $V = \operatorname{Var}(y)$ von ϑ notationell unterdrückt. Die Herleitung richtet sich weitgehend nach Diggle et al. (1994) Seite 64-68, sowie nach Harville (1974).

Für y gilt

$$y \sim \mathrm{N}(X\beta, V)$$

und damit für die Dichte der Verteilung von y

$$p(y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} |V|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)\right].$$

Als Schätzer für β erhält man den gewichteten Kleinste-Quadrate-Schätzer

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y = G'y$$

wobei im Folgenden die Matrix G durch

$$G = V^{-1}X(X'V^{-1}X)^{-1}$$

definiert sei. Für $\hat{\beta}$ gilt dann

$$\hat{\beta} \sim \mathrm{N}(\beta, (X'V^{-1}X)^{-1})$$

und für die Dichte der Verteilung von $\hat{\beta}$

$$p(\hat{\beta}) = \left(\frac{1}{2\pi}\right)^{\frac{p+1}{2}} |X'V^{-1}X|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\hat{\beta}-\beta)'X'V^{-1}X(\hat{\beta}-\beta)\right]. \qquad \text{(A.12)}$$

Die Fehlerkontraste u sind nun definiert durch $u = A'y$, wobei die $n \times n-p-1$-Matrix A aus der Zerlegung $AA' = I - X(X'X)^{-1}X'$ mit $A'A = I$ stammt. Im Folgenden wird stets die so definierte Matrix verwendet. Die resultierende Verteilung ändert sich jedoch bei Verwendung einer anderen Matrix nur um eine Normierungskonstante, solange die Matrix so gewählt wird, dass man $n-p-1$ linear unabhängige Fehlerkontraste erhält (Verbeke & Molenberghs (2000) Kapitel 5.3).

Für die so definierten Fehlerkontraste gilt dann die gewünschte Eigenschaft $\mathbb{E}(u) = 0$, denn

$$\begin{aligned}
\mathbb{E}(u) &= A'\,\mathbb{E}(y) = A'X\beta = \underbrace{A'A}_{=I}A'X\beta \\
&= A'\underbrace{(I - X(X'X)^{-1}X')}_{=AA'}X\beta \\
&= A'X\beta - A'X\underbrace{(X'X)^{-1}X'X}_{=I}\beta \\
&= 0.
\end{aligned}$$

Darüberhinaus sind u und $\hat{\beta}$ stochastisch unabhängig, denn es gilt

$$\begin{aligned}
\mathrm{Cov}(u,\hat{\beta}) &= \mathbb{E}(u(\hat{\beta}-\beta)') \\
&= \mathbb{E}(A'y(y'G' - \beta') \\
&= A'\,\mathbb{E}(yy')G' - A'\,\mathbb{E}(y)\beta' \\
&= A'(\mathrm{Var}(y) + \mathbb{E}(y)\,\mathbb{E}(y)')G' - A'\,\mathbb{E}(y)\beta' \\
&= A'(V + X\beta\beta'X')G' - A'X\beta\beta' \\
&= A'VG' + A'X\beta\beta'X'G' - A'X\beta\beta' \\
&= A'VG' + A'X\beta\beta'\underbrace{X'V^{-1}X(X'V^{-1}X)^{-1}}_{=I} - A'X\beta\beta' \\
&= A'VG'
\end{aligned}$$

$$
\begin{aligned}
&= A'VV^{-1}X(X'V^{-1}X)^{-1} \\
&= A'X(X'V^{-1}X)^{-1} \\
&= A'AA'X(X'V^{-1}X)^{-1} \\
&= A'\underbrace{(I - X(X'X)^{-1}X')X}_{=0}(X'V^{-1}X)^{-1} \\
&= 0.
\end{aligned}
$$

Damit sind u und $\hat{\beta}$ unkorreliert und aufgrund der gemeinsamen Normalverteilung auch unabhängig.

Betrachtet man nun die Transformation $T(y) = (A, G)'y = (u', \hat{\beta}')'$, so gilt nach dem Transformationssatz für Dichten (A.2)

$$
\begin{aligned}
p(u, \hat{\beta}) &= \frac{1}{\mathrm{abs}(|(A, G)'|)} p(T^{-1}(u, \hat{\beta})) \\
&= \frac{1}{\mathrm{abs}(|(A, G)'|)} p(y).
\end{aligned}
$$

Für $|(A, G)'|$ erhält man

$$
\begin{aligned}
|(A, G)'| &\overset{(A.3)(A.4)}{=} |(A, G)'(A, G)|^{\frac{1}{2}} \\
&= \left| \begin{pmatrix} A'A & A'G \\ G'A & G'G \end{pmatrix} \right|^{\frac{1}{2}} \\
&\overset{(A.5)}{=} |A'A|^{\frac{1}{2}} |G'G - G'A \underbrace{(A'A)^{-1}}_{=I} A'G|^{\frac{1}{2}} \\
&= |I|^{\frac{1}{2}} |G'G - G'(I - X(X'X)^{-1}X')G|^{\frac{1}{2}} \\
&= |G'X(X'X)^{-1}X'G|^{\frac{1}{2}} \\
&= |(X'V^{-1}X)^{-1}X'V^{-1}X(X'X)^{-1}X'V^{-1}X(X'V^{-1}X)^{-1}|^{\frac{1}{2}} \\
&= |X'X|^{-\frac{1}{2}}.
\end{aligned}
$$

Beachtet man noch, dass $X'X$ positiv semidefinit ist, so ergibt sich zusammen mit der Unabhängigkeit von u und $\hat{\beta}$ für die Dichte der Verteilung von u

$$
p(u) = |X'X|^{\frac{1}{2}} \frac{p(y)}{p(\hat{\beta})}.
$$

Um das Verhältnis $p(y)/p(\hat{\beta})$ bestimmen zu können, benötigt man noch das fol-

gende Ergebnis:

$$\begin{aligned}
&(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) + (\hat{\beta} - \beta)'X'V^{-1}X(\hat{\beta} - \beta) \\
&= y'V^{-1}y - 2\hat{\beta}'X'V^{-1}y + \hat{\beta}'X'V^{-1}X\hat{\beta} \\
&\quad +\hat{\beta}'X'V^{-1}X\hat{\beta} - 2\beta'X'V^{-1}X\hat{\beta} + \beta'X'V^{-1}X\beta \\
&= y'V^{-1}y - 2y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y \\
&\quad +2y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y \\
&\quad -2\beta'X'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y + \beta'X'V^{-1}X\beta \\
&= y'V^{-1}y - 2\beta'X'V^{-1}y + \beta'X'V^{-1}X\beta \\
&= (y - X\beta)'V^{-1}(y - X\beta).
\end{aligned}$$

Damit erhält man insgesamt

$$p(u) = \left(\frac{1}{2\pi}\right)^{\frac{n-p-1}{2}} |X'X|^{\frac{1}{2}}|V|^{-\frac{1}{2}}|X'V^{-1}X|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})\right].$$

A.3 Score-Funktion

Nun sollen die zur numerischen Berechnung von $\hat{\vartheta}_{REML}$ benötigten Größen hergeleitet werden. Bezeichne dazu zunächst wieder

$$l^*(\vartheta) = -\frac{1}{2}\log|V| - \frac{1}{2}\log|X'V^{-1}X| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) \qquad \text{(A.13)}$$

die logarithmierte Restricted-Likelihood. Außerdem sei die Matrix P wieder definiert durch

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}. \qquad \text{(A.14)}$$

Man beachte, dass auch P und $\hat{\beta}$ wie die Kovarianzmatrix V von ϑ abhängen. Diese Abhängigkeit wird aber im Folgenden ebenfalls notationell unterdrückt, um die Formeln und Herleitungen nicht unnötig zu verkomplizieren.

Zunächst werden nun einige Hilfsergebnisse bestimmt, die es anschließend erlauben, die Ableitung der logarithmierten Restricted-Likelihood nach einem Element des Vektors ϑ und damit die Elemente der Score-Funktion herzuleiten.

Zunächst berechnet sich $\frac{\partial}{\partial \vartheta_j} \log |X'V^{-1}X|$ als

$$\begin{aligned}
\frac{\partial}{\partial \vartheta_j} \log |X'V^{-1}X| &\overset{(A.11)}{=} \operatorname{spur}\left((X'V^{-1}X)^{-1}\frac{\partial X'V^{-1}X}{\partial \vartheta_j}\right) \\
&= \operatorname{spur}\left((X'V^{-1}X)^{-1}X'\frac{\partial V^{-1}}{\partial \vartheta_j}X\right) \\
&\overset{(A.9)}{=} -\operatorname{spur}\left((X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}X\right) \\
&\overset{(A.8)}{=} -\operatorname{spur}\left(V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_j}\right). \quad \text{(A.15)}
\end{aligned}$$

Für $\frac{\partial}{\partial \vartheta_j}(y - X\hat{\beta})$ erhält man

$$\begin{aligned}
\frac{\partial}{\partial \vartheta_j}(y - X\hat{\beta}) &= -\frac{\partial X\hat{\beta}}{\partial \vartheta_j} \\
&= -\frac{\partial}{\partial \vartheta_j}X(X'V^{-1}X)^{-1}X'V^{1}y \\
&= -X\frac{\partial (X'V^{-1}X)^{-1}}{\partial \vartheta_j}X'V^{-1}y - X(X'V^{-1}X)^{-1}X'\frac{\partial V^{-1}}{\partial \vartheta_j}y \\
&\overset{(A.9)}{=} X(X'V^{-1}X)^{-1}X'\frac{\partial V^{-1}}{\partial \vartheta_j}X(X'V^{-1}X)^{-1}X'V^{-1}y \\
&\quad +X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}y \\
&\overset{(A.9)}{=} X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}(y - X\underbrace{(X'V^{-1}X)^{-1}X'V^{-1}y}_{=\hat{\beta}}) \\
&= X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}(y - X\hat{\beta}). \quad \text{(A.16)}
\end{aligned}$$

Darüberhinaus gilt

$$\begin{aligned}
&V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}(y - X\hat{\beta}) \\
&= V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y \\
&\quad -V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}X(X'V^{-1}X)^{-1}\underbrace{X'V^{-1}X(X'V^{-1}X)^{-1}}_{=I}X'V^{-1}y \\
&= 0. \quad \text{(A.17)}
\end{aligned}$$

Mit Hilfe von (A.16) und (A.17) lässt sich nun $\frac{\partial}{\partial\vartheta_j}(y-X\hat{\beta})'V^{-1}(y-X\hat{\beta})$ bestimmen:

$$\begin{aligned}
&\frac{\partial}{\partial\vartheta_j}(y-X\hat{\beta})'V^{-1}(y-X\hat{\beta}) \\
&\quad= \quad \frac{\partial}{\partial\vartheta_j}\left[(y-X\hat{\beta})'\right]V^{-1}(y-X\hat{\beta})+(y-X\hat{\beta})'\frac{\partial}{\partial\vartheta_j}\left[V^{-1}(y-X\hat{\beta})\right] \\
&\overset{(A.16)}{=} \quad (y-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}(y-X\hat{\beta}) \\
&\qquad +(y-X\hat{\beta})'\frac{\partial V^{-1}}{\partial\vartheta_j}(y-X\hat{\beta})+(y-X\hat{\beta})'V^{-1}\frac{\partial}{\partial\vartheta_j}\left[y-X\hat{\beta}\right] \\
&\overset{(A.9)(A.16)}{=} \quad (y-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}(y-X\hat{\beta}) \\
&\qquad -(y-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}(y-X\hat{\beta}) \\
&\qquad (y-X\hat{\beta})'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}(y-X\hat{\beta}) \\
&\overset{(A.17)}{=} \quad -(y-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}(y-X\hat{\beta}). \qquad (A.18)
\end{aligned}$$

Unter Verwendung der Hilfsergebnisse ist es nun leicht, den j-ten Eintrag der Score-Funktion $s^*(\vartheta)=(s_j^*(\vartheta))_{j=1,\ldots,d}$ zu berechnen:

$$\begin{aligned}
s_j^*(\vartheta) = \quad & \frac{\partial l^*(\vartheta)}{\partial\vartheta_j} \\
\overset{(A.13)}{=} \quad & -\frac{1}{2}\frac{\partial}{\partial\vartheta_j}\log|V|-\frac{1}{2}\frac{\partial}{\partial\vartheta_j}\log|X'V^{-1}X| \\
& -\frac{1}{2}\frac{\partial}{\partial\vartheta_j}\left[(y-X\hat{\beta})'V^{-1}(y-X\hat{\beta})\right] \\
\overset{(A.11)(A.15)(A.18)}{=} \quad & -\frac{1}{2}\,\mathrm{spur}\left(V^{-1}\frac{\partial V}{\partial\vartheta_j}\right)+\frac{1}{2}\,\mathrm{spur}\left(V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial\vartheta_j}\right) \\
& +\frac{1}{2}(y-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}(y-X\hat{\beta}) \\
= \quad & -\frac{1}{2}\,\mathrm{spur}\left((V^{-1}-V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})\frac{\partial V}{\partial\vartheta_j}\right) \\
& +\frac{1}{2}(y-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}(y-X\hat{\beta}) \\
\overset{(A.14)}{=} \quad & -\frac{1}{2}\,\mathrm{spur}\left(P\frac{\partial V}{\partial\vartheta_j}\right)+\frac{1}{2}(y-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\vartheta_j}V^{-1}(y-X\hat{\beta}) \qquad (A.19)
\end{aligned}$$

A.4 Beobachtete Fisher-Information

Zur Schätzung von ϑ benötigt man nicht nur die Score-Funktion, sondern auch die beobachtete oder die erwartete Fisher-Information. Um die beobachtete Fisher-Information bestimmen zu können, werden zunächst wieder einige Hilfsergebnisse vorgestellt.

Für die Ableitung der Matrix P gilt

$$\begin{aligned}
\frac{\partial P}{\partial \vartheta_k} &= \frac{\partial V^{-1}}{\partial \vartheta_k} - \frac{\partial}{\partial \vartheta_k} V^{-1} X (X'V^1X)^{-1} X'V^{-1} \\
&\overset{(A.9)}{=} -V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1} - \frac{\partial V^{-1}}{\partial \vartheta_k} X (X'V^{-1}X)^{-1} X'V^{-1} \\
&\quad -V^{-1}X\frac{\partial}{\partial \vartheta_k}[(X'V^{-1}X)^{-1}X'V^{-1}] \\
&\overset{(A.9)}{=} -V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1} + V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1} X (X'V^{-1}X)^{-1} X'V^{-1} \\
&\quad -V^{-1}X\frac{\partial (X'V^{-1}X)^{-1}}{\partial \vartheta_k} X'V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'\frac{\partial V^{-1}}{\partial \vartheta_k} \\
&\overset{(A.9)}{=} -V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1} + V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1} X (X'V^{-1}X)^{-1} X'V^{-1} \\
&\quad -V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} \\
&\quad +V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1} \\
&= -P\frac{\partial V}{\partial \vartheta_k}P. \qquad \text{(A.20)}
\end{aligned}$$

Außerdem erhält man

$$\begin{aligned}
&\frac{\partial}{\partial \vartheta_k}\left[(y-X\hat\beta)'V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}(y-X\hat\beta)\right] \\
&= \frac{\partial}{\partial \vartheta_k}\left[(y-X\hat\beta)'\right]V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}(y-X\hat\beta) \\
&\quad +(y-X\hat\beta)'\frac{\partial}{\partial \vartheta_k}\left[V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}(y-X\hat\beta)\right] \\
&\overset{(A.16)}{=} (y-X\hat\beta)'V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}(y-X\hat\beta) \\
&\quad +(y-X\hat\beta)'\frac{\partial}{\partial \vartheta_k}\left[V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}\right](y-X\hat\beta) \\
&\quad +(y-X\hat\beta)'V^{-1}\frac{\partial V}{\partial \vartheta_j}V^{-1}\frac{\partial}{\partial \vartheta_k}\left[y-X\hat\beta\right]
\end{aligned}$$

$$
\begin{aligned}
\overset{(A.9)(A.16)}{=} \quad & (y - X\hat{\beta})' V^{-1} \frac{\partial V}{\partial \vartheta_k} V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} \frac{\partial V}{\partial \vartheta_j} V^{-1} (y - X\hat{\beta}) \\
& + (y - X\hat{\beta})' \left(-V^{-1} \frac{\partial V}{\partial \vartheta_k} V^{-1} \frac{\partial V}{\partial \vartheta_j} V^{-1} + V^{-1} \frac{\partial^2 V}{\partial \vartheta_j \partial \vartheta_k} V^{-1} \right. \\
& \left. - V^{-1} \frac{\partial V}{\partial \vartheta_j} V^{-1} \frac{\partial V}{\partial \vartheta_k} V^{-1} \right) (y - X\hat{\beta}) \\
& + (y - X\hat{\beta})' V^{-1} \frac{\partial V}{\partial \vartheta_j} V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} \frac{\partial V}{\partial \vartheta_k} V^{-1} (y - X\hat{\beta}) \\
= \quad & (y - X\hat{\beta})' V^{-1} \left(\frac{\partial^2 V}{\partial \vartheta_j \partial \vartheta_k} - 2 \frac{\partial V}{\partial \vartheta_j} P \frac{\partial V}{\partial \vartheta_k} \right) V^{-1} (y - X\hat{\beta}). \qquad \text{(A.21)}
\end{aligned}
$$

Damit ergibt sich für die einzelnen Einträge $F^*_{obs,jk}(\vartheta)$ der beobachteten Fisher-Information

$$
\begin{aligned}
F^*_{obs,jk}(\vartheta) &= -\frac{\partial^2 l^*(\vartheta)}{\partial \vartheta_j d\vartheta_k} = -\frac{\partial s^*_j(\vartheta)}{\partial \vartheta_k} \\
\overset{(A.19)(A.10)}{=} \quad & \frac{1}{2} \operatorname{spur} \left(\frac{\partial}{\partial \vartheta_k} \left[P \frac{\partial V}{\partial \vartheta_j} \right] \right) - \frac{1}{2} \frac{\partial}{\partial \vartheta_k} \left[(y - X\hat{\beta})' V^{-1} \frac{\partial V}{\partial \vartheta_j} V^{-1} (y - X\hat{\beta}) \right] \\
= \quad & \frac{1}{2} \operatorname{spur} \left(\frac{\partial P}{\partial \vartheta_k} \frac{\partial V}{\partial \vartheta_j} + P \frac{\partial^2 V}{\partial \vartheta_j \partial \vartheta_k} \right) \\
& - \frac{1}{2} \frac{\partial}{\partial \vartheta_k} \left[(y - X\hat{\beta})' V^{-1} \frac{\partial V}{\partial \vartheta_j} V^{-1} (y - X\hat{\beta}) \right] \\
\overset{(A.20)(A.21)(A.8)}{=} \quad & \frac{1}{2} \operatorname{spur} \left(P \frac{\partial^2 V}{\partial \vartheta_j \partial \vartheta_k} - P \frac{\partial V}{\partial \vartheta_j} P \frac{\partial V}{\partial \vartheta_k} \right) \\
& - \frac{1}{2} (y - X\hat{\beta})' V^{-1} \left(\frac{\partial^2 V}{\partial \vartheta_j \partial \vartheta_k} - 2 \frac{\partial V}{\partial \vartheta_j} P \frac{\partial V}{\partial \vartheta_k} \right) V^{-1} (y - X\hat{\beta}). \qquad \text{(A.22)}
\end{aligned}
$$

A.5 Erwartete Fisher-Information

Nun soll noch die erwartete Fisher-Information hergeleitet werden. Dazu bestimmt man zunächst $\operatorname{Var}(y - X\hat{\beta})$:

$$
\begin{aligned}
\operatorname{Var}(y - X\hat{\beta}) &= \operatorname{Var}(y) + \operatorname{Var}(X\hat{\beta}) - 2 \operatorname{Cov}(y, X\hat{\beta}) \\
&= V + X \operatorname{Var}(\hat{\beta}) X' - 2 \operatorname{Cov}(y, X(X' V^{-1} X)^{-1} X' V^{-1} y) \\
&= V + X(X' V^{-1} X)^{-1} X' - 2X(X' V^{-1} X)^{-1} X' V^{-1} \operatorname{Var}(y) \\
&= V + X(X' V^{-1} X)^{-1} X' - 2X(X' V^{-1} X)^{-1} X' V^{-1} V \\
&= V - X(X' V^{-1} X)^{-1} X'. \qquad \text{(A.23)}
\end{aligned}
$$

Dieses Ergebnis verwendet man zusammen mit $\mathbb{E}(y - X\hat{\beta}) = 0$, um an geeigneter Stelle Formel (A.7) anwenden zu können. Für die einzelnen Einträge $F^*_{jk}(\vartheta)$ der erwarteten Fisher-Information erhält man so

$$
\begin{aligned}
F^*_{jk}(\vartheta) &= \mathbb{E}(F^*_{obs,jk}(\vartheta)) \\
&\overset{(A.22)}{=} \frac{1}{2}\,\text{spur}\left(P\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} - P\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right) \\
&\quad -\frac{1}{2}\,\mathbb{E}\left[(y - X\hat{\beta})'V^{-1}\left(\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} - 2\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right)V^{-1}(y - X\hat{\beta})\right] \\
&\overset{(A.7)}{=} \frac{1}{2}\,\text{spur}\left(P\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} - P\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right) \\
&\quad -\frac{1}{2}\,\text{spur}\left(V^{-1}\left(\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} - 2\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right)V^{-1}\,\text{Var}(y - X\hat{\beta})\right) \\
&\overset{(A.23)}{=} \frac{1}{2}\,\text{spur}\left(P\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} - P\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right) \\
&\quad -\frac{1}{2}\,\text{spur}\left(V^{-1}\left(\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} - 2\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right)V^{-1}(V - X(X'V^{-1}X)^{-1}X')\right) \\
&\overset{(A.8)}{=} \frac{1}{2}\,\text{spur}\left(P\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} - P\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k} - P\frac{\partial^2 V}{\partial\vartheta_j\partial\vartheta_k} + 2P\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right) \\
&= \frac{1}{2}\,\text{spur}\left(P\frac{\partial V}{\partial\vartheta_j}P\frac{\partial V}{\partial\vartheta_k}\right).
\end{aligned}
$$

B GGAMM Software-Beschreibung

Im Rahmen dieser Arbeit wurden die in Kapitel 3 und insbesondere in Kapitel 3.3 und 3.4 beschriebenen Verfahren zur Schätzung generalisierter geoadditiver gemischter Modelle mit Hilfe von P-Splines und Markov-Zufallsfeldern in S-Plus implementiert. Die Implementation besteht aus den Dateien

- `ggamm.s`,
- `ggammc.s`,
- `helpfunctions.s` und
- `mat.dll`,

die im Verzeichnis `functions` auf der beiliegenden CD-Rom zu finden sind. Während die beiden Dateien `ggamm.s` und `ggammc.s` unterschiedliche Versionen der zur Schätzung verwendeten Funktion mit Namen `ggamm` und `ggammc` beinhalten, werden in der Datei `helpfunctions.s` eine Reihe von Hilfsfunktionen definiert. Um die Geschwindigkeit zu erhöhen, mit der das Programm ausgeführt wird, werden in der mit `ggammc` bezeichneten Implementation einige Berechnungen nicht mit Hilfe von in S-Plus programmierten Funktionen, sondern durch in der Programmiersprache C abgefasste Funktionen durchgeführt. Diese C-Funktionen sind in der Datei `mat.dll` enthalten.

Die Implementation `ggammc` ist für kleine bis mittlere Datensätze (bis circa 1000 Beobachtungen) vorzuziehen, da die Berechnungen mit einer zum Teil deutlich größeren Geschwindigkeit ausgeführt werden. Für große Datensätze ist dagegen die reine S-Plus-Implementation `ggamm` vorteilhafter, da ab einer gewissen Beobachtungszahl der Datenaustausch zwischen S-Plus und C den Zeitvorteil wieder aufhebt, der durch die schnelleren Berechnungen entstanden ist. Außerdem vermeidet diese Implementation gewisse Speicherplatzprobleme, die durch die gleichzeitige Verwendung originärer S-Plus-Funktionen und in C programmierter Funktionen entstehen können.

Neben den beiden S-Plus-Implementationen enthält das Verzeichnis `functions` auch eine Implementation für das Programmpaket R, die in den beiden Dateien `ggamm.r` und `helpfunctions.r` enthalten ist. Die Installation und Benutzung erfolgt hier völlig analog zur S-Plus-Version, so dass auf diese Implementation im

Folgenden nicht weiter eingegangen wird. Es sei jedoch darauf hingewiesen, dass sich durch die Verwendung der R-Implementation die Rechenzeiten im Vergleich zur reinen S-Plus-Implementation in etwa halbieren. Da in R keine Möglichkeit besteht, C-Funktionen einzubinden, existiert hier lediglich die Programm-Version `ggamm`.

B.1 Installation und Aufruf

Um die Funktionen `ggamm` und `ggammc` in S-Plus zu definieren, müssen zunächst durch die Ausführung des Kommandos

```
> source("e:\\functions\\helpfunctions.s")
```

die Hilfsfunktionen installiert werden. Eventuell ist dabei noch zusätzlich der Pfad zu dem Verzeichnis, in dem sich diese Datei befindet, entsprechend zu verändern. Man beachte, dass durch den Aufruf des `source`-Kommandos auch die in `mat.dll` enthaltenen Funktionen in S-Plus eingelesen werden. Damit dies möglich ist, muss in der Datei `helpfunctions.s` der korrekte Pfad, unter dem `mat.dll` zu finden ist, in den Aufruf

```
> dll.load("e:\\functions\\mat.dll",c("spur2",...),"cdecl")
```

eingesetzt werden. Man beachte außerdem, dass die in C programmierten Funktionen nur temporär in S-Plus definiert werden. Das heißt, nach Beendigung von S-Plus sind die zugehörigen Funktionen beim nächsten Programmaufruf nicht mehr verfügbar und müssen wieder über den obigen `source`-Befehl definiert werden.

Je nachdem, welche der beiden Funktionen `ggamm` und `ggammc` gewählt wird, ist dann zusätzlich noch der Befehl

```
> source("e:\\functions\\ggamm.s")
```

beziehungsweise

```
> source("e:\\functions\\ggammc.s")
```

auszuführen. Durch die Ausführung beider Befehle werden sowohl `ggamm` als auch `ggammc` in S-Plus installiert.

Die beiden Implementationen `ggamm` und `ggammc` unterscheiden sich, abgesehen von ihrem Namen, weder in ihrem Aufruf, das heißt in den möglichen Argumenten, noch in ihrem Rückgabewert. Der Aufruf besitzt prinzipiell die Form

```
> test<-ggamm(dep=...,fix=...,smooth=...,reg=...,regions=...,
  random=...,id=...,...)
```

beziehungsweise

```
> test<-ggammc(dep=...,fix=...,smooth=...,reg=...,regions=...,
  random=...,id=...,...)
```

Die Ergebnisse sind dann im S-Plus-Objekt `test` gespeichert und stehen damit beispielsweise zur Visualisierung der Schätzungen (vergleiche Abschnitt B.4) zur Verfügung. Ein genauer Überblick über die einzelnen Argumente der Funktionen `ggamm` und `ggammc` wird im folgenden Abschnitt gegeben. In Abschnitt B.5 wird anhand einiger Beispiele genauer erläutert, wie der Funktionsaufruf für verschiedene Modelle zu erfolgen hat.

B.2 Argumente

Das einzige notwendige Argument ist die abhängige Variable `dep`. Werden keine weiteren Argumente übergeben, so wird lediglich ein konstanter Effekt geschätzt. Man beachte dabei, dass für den Fall, dass nur parametrische Effekte spezifiziert werden, `ggamm` lediglich die S-Plus-Funktion `glm` aufruft und die resultierenden Ergebnisse in die Form des `ggamm`-typischen Rückgabewerts bringt.

`dep` Vektor der abhängigen Variablen y.

`family` Exponentialfamilie, aus der die Verteilung der abhängigen Variablen stammt.

- `"normal"` Normalverteilung.
- `"binomial"` Binomialverteilung. Gilt $y_i \sim B(n_i, p_i)$, so muss y_i/n_i als abhängige Variable angegeben werden. n_i wird dann in der Gewichtsvariable `weight` übergeben.
- `"poisson"` Poissonverteilung.
- `"gamma"` Gammaverteilung.

Voreinstellung: `"normal"`

`link` Linkfunktion. Für Normal- und Poissonverteilung wird automatisch die natürliche Linkfunktion, für die Gammaverteilung der Logarithmus als Linkfunktion verwendet. Die Option spielt also nur bei binomialverteiltem Response eine Rolle.

`"logit"` Umkehrfunktion der Verteilungsfunktion der logistischen Verteilung als Linkfunktion.

`"probit"` Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung als Linkfunktion.

Voreinstellung: `"logit"`

`dispers` Wahrheitswert, der angibt, ob der Dispersionsparameter ϕ geschätzt werden soll. Für Normalverteilung und Gammaverteilung ist die Schätzung des Skalenparameters zwingend erforderlich.

Voreinstellung: `F`

`fix` Matrix der Ausprägungen der parametrisch zu modellierenden Kovariablen.

`smooth` Matrix der Ausprägungen der als P-Splines zu modellierenden Kovariablen.

`nknot` Vektor, der für jede Variable in `smooth` angibt, wie groß die Knotenzahl m sein soll.

Voreinstellung: 20

`ord` Vektor, der für jede Variable in `smooth` angibt, Differenzen welcher Ordnung als Penalisierung verwendet werden sollen. Möglich sind nur Differenzen erster und zweiter Ordnung.

Voreinstellung: 2

`deg` Vektor, der für jede Variable in `smooth` angibt, von welchem Grad die verwendete B-Spline-Basis sein soll.

Voreinstellung: 3

`plotf` Wahrheitswert, der angibt, ob die Funktionsschätzungen automatisch gezeichnet werden sollen.

Voreinstellung: `F`

`include.lin` Wahrheitswert, der angibt, ob in die Abbildung der Funktionsschätzungen der lineare Anteil der Funktion gezeichnet werden soll. Wird nur für Funktionen ausgeführt, für die `ord`=2 gilt.

Voreinstellung: `F`

`reg` Vektor, der die beobachteten räumlichen Kovariablenausprägungen (in der Regel Kennziffern verschiedener Regionen) enthält.

`pmatrix` Strafmatrix K_{spat} des räumlichen Effekts.

`regions` Vektor der möglichen räumlichen Kovariablenausprägungen in der `pmatrix` entsprechenden Reihenfolge.

`plotmap` Wahrheitswert, der angibt, ob die geschätzte räumliche Funktion automatisch gezeichnet werden soll.

Voreinstellung: `F`

`map` Kartenobjekt, mit dessen Hilfe der geschätzte räumliche Effekt gezeichnet werden kann. Vergleiche Kapitel 5 in Brezger et al. (2002).

`random` Vektor der Kovariablenausprägungen, deren Effekte als zufällig modelliert werden sollen. Ein zufälliger Intercept wird über einen n-dimensionalen Vektor aus Einsen angegeben.

`id` Gruppierungsvariable der zufälligen Effekte.

`weight` Vektor von Gewichten.

`offset` Vektor, der die Ausprägungen des Offsets enthält.

`sig` Sicherheitsgrad der zu bestimmenden Konfidenzintervalle und Konfidenzbänder.
Voreinstellung: 0.95

`eps` Schätzgenauigkeit.
Voreinstellung: 0.00001

`noprint` Wahrheitswert. Falls `noprint` wahr ist, werden während der Schätzung keine Informationen über die Iterationen und die resultierenden Schätzergebnisse ausgegeben.
Voreinstellung: `F`

`nowarnings` Wahrheitswert. Falls `nowarnings` wahr ist, werden keine Warnungen ausgegeben.
Voreinstellung: `F`

`startValue1` Startwert für die inversen Glättungsparameter von P-Splines und räumlichen Effekten.
Voreinstellung: 0.1

`startValue2` Startwert für die Varianzparameter der zufälligen Effekte.
Voreinstellung: 0.5

`maxit` Maximale Iterationszahl.
Voreinstellung: 400

`ranktest` Wahrheitswert, der angibt, ob die in Kapitel 2.3.3 definierte Fisher-Informationsmatrix zur Schätzung von β und b vor jeder Iteration auf ein Rangdefizit getestet werden soll. Verlangsamt die Berechnungen.
Voreinstellung: `F`

`diagmult` Wert, mit dem die Diagonalelemente der Fisher-Information multipliziert werden sollen, wenn ein Rangdefizit vorliegt. Vergleiche Kapitel 2.3.3.
Voreinstellung: 1.0005

`outfile` Erlaubt die Spezifikation eines Pfades, in den die resultierenden Schätzungen geschrieben werden. Beispielsweise werden die Schätzungen der fixen Effekte bei einer Spezifikation von `outfile="c:\\temp\\results"` in der Datei `c:\\temp\\results_fixedEffects.raw` gespeichert.

`log.like` Wahrheitswert, der angibt, ob die (Restricted-) Log-Likelihood des Modells berechnet werden soll. Nur im Normalverteilungsfall implementiert.
Voreinstellung: `F`

`lrtest` Wahrheitswert, der angibt, ob basierend auf der (hier im Allgemeinen nicht zutreffenden) Theorie für unabhängige, identisch verteilte Beobachtungen aus Self & Liang (1987) approximative LQ-Tests der nonparametrischen, räumlichen und zufälligen Effekte über die zugehörigen Varianzparameter durchgeführt werden sollen (vergleiche Kapitel 4).
Voreinstellung: `F`

`lowerlim` Gibt an, ab welchem Wert des Kriteriums (3.9) die Schätzung eines Varianzparameters gestoppt werden soll. Vergleiche Kapitel 3.4.
Voreinstellung: 0.001

`method` Schätzverfahren für die Varianzparameter.
`"ML"` Maximum-Likelihood
`"REML"` Restricted-Maximum-Likelihood
Voreinstellung: `"REML"`

B.3 Rückgabewert

Der Rückgabewert sowohl der Funktion `ggamm` als auch der Funktion `ggammc` besteht aus einer Liste von S-Plus-Objekten, die jeweils die Schätzungen verschiedener Modellkomponenten enthalten. In der folgenden Übersicht werden diese Objekte kurz beschrieben. Man beachte, dass einige Objekte, wie beispielsweise `spatialEffects` oder `log.like` nur erzeugt werden, wenn sie im analysierten Modell definiert sind beziehungsweise im Funktionsaufruf angefordert wurden.

`fixedEffects` Matrix der Schätzungen für die parametrisch modellierten Effekte. Zusätzlich enthalten sind Konfidenzintervalle, Standardabweichungen und p-Werte. Wurden nonparametrisch modellierte Funktionen mit über Differenzen der Ordnung $k = 2$ penalisierten P-Splines geschätzt, so sind zusätzlich die zugehörigen Parameter des Linearanteils der Schätzungen wiedergegeben.

`iterations` Zur Schätzung benötigte Iterationen.

`log.like` Log-Likelihood beziehungsweise Restricted-Log-Likelihood des geschätzten Modells.

`predict` Schätzung des linearen Prädiktors und des Erwartungswertes für die verschiedenen Beobachtungen.

`randomEffects`	Matrix der Schätzungen der zufälligen Effekte. Falls mehrere zufällige Effekte bestimmt wurden, werden auch mehrere, entsprechend nummerierte Matrizen erzeugt. Enthält zusätzlich punktweise Konfidenzintervalle, Standardabweichungen und p-Werte.
`smoothedEffects`	Matrix der Schätzungen der nonparametrischen Effekte. Enthält zusätzlich punktweise Konfidenzintervalle, Standardabweichungen und p-Werte sowie bei über Differenzen der Ordnung $k = 2$ penalisierten P-Splines den Linearanteil der Schätzung und die Abweichung vom Linearanteil. Falls mehrere nonparametrische Effekte bestimmt wurden, werden auch mehrere, entsprechend nummerierte Matrizen erzeugt.
`spatialEffects`	Matrix der Schätzungen des räumlichen Effekts. Enthält zusätzlich punktweise Konfidenzintervalle, Standardabweichungen und p-Werte.
`theta`	Matrix der geschätzten Varianz- und Glättungsparameter.
`errors`	Zeichenkette, die eventuell aufgetretene Fehler beschreibt.

B.4 Visualisierung der geschätzten Effekte

Zur nachträglichen Visualisierung der geschätzten nonparametrischen und räumlichen Effekte stehen zwei Funktionen zur Verfügung. Sind die Schätzergebnisse im Objekt `test` gespeichert, so können über den Aufruf

```
> plotf(test)
```

die enthaltenen nonparametrischen Schätzungen zusammen mit den frequentistischen und bayesianischen Konfidenzintervallen gezeichnet werden. Zusätzlich besitzt die Funktion noch das Argument `include.lin`. Wird im Funktionsaufruf die Option `include.lin=T` gewählt, so werden für die einzelnen Funktionen zusätzlich die linearen Anteile der Funktionsschätzung in den Grafiken berücksichtigt.

Schätzungen räumlicher Funktionen können über den Aufruf

```
> plotmap(test,m)
```

gezeichnet werden, wenn im Objekt `m` die entsprechende Karte gespeichert ist. Man vergleiche Kapitel 5 in Brezger et al. (2002) für eine Beschreibung, wie die zugehörigen Kartenobjekte in S-Plus definiert werden. Die Funktion `plotmap`

besteht im wesentlichen im Aufruf der von Andreas Brezger geschriebenen Funktion `drawmap`, die ebenfalls in der Datei `helpfunctions.s` enthalten ist. Für diese existieren noch eine Reihe zusätzlicher Optionen, die in Abschnitt 7.4.2.3 in Brezger et al. (2002) beschrieben werden. Die Funktion `plotmap` dient lediglich dazu, mit etwas weniger Schreibarbeit einen ersten Eindruck von der räumlichen Schätzung zu erhalten.

Sowohl `plotf` als auch `plotmap` sind in der Datei `helpfunctions.s` enthalten und werden automatisch durch den Aufruf

```
> source("e:\\functions\\helpfunctions.s")
```

definiert.

B.5 Beispiele

Anhand zweier Beispiele, soll nun noch kurz die Schätzung generalisierter additiver Modelle und generalisierter geoadditiver gemischter Modelle mit Hilfe von `ggamm` beschrieben werden. Zur Verwendung der Funktion `ggammc` sind lediglich der `source`-Befehl sowie im Funktionsaufruf `ggamm` durch `ggammc` zu ersetzen.

In Beispiel 1 werden generalisierte additive Modelle für normal- und binomialverteilten Response simuliert, wobei der lineare Prädiktor gegeben ist durch die Summe zweier nonparametrischer Effekte.

Beispiel 1: Generalisierte additive Modelle.

```
> source("e:\\functions\\helpfunctions.s")
> source("e:\\functions\\ggamm.s")
> x1<-runif(200,-3,3)
> x2<-runif(200,-1,1)
> eta<-sin(x1)+x2^2
> y<-rnorm(200,eta,0.7)
> covariates<-cbind(x1,x2)
> test<-ggamm(dep=y,smooth=covariates,nknot=c(20,20),dispers=T,
  family="normal")
> plotf(test)
> p<-exp(eta)/(1+exp(eta))
```

```
> y<-rbinom(200,3,p)
> y<-y/3
> n<-rep(3,200)
> test<-ggamm(dep=y,smooth=covariates,nknot=c(20,20),weight=n,
  family="binomial")
> plotf(test)
```

Beispiel 2 beschreibt nun die Schätzung eines generalisierten geoadditiven gemischten Modells, dass aus einer nonparametrischen Funktion, einer räumlichen Funktion und einem zufälligen Intercept besteht. Die räumliche Funktion entspricht dabei der Funktion aus der Simulation in Kapitel 5.2 und ist auf der beiliegenden CD-Rom im Verzeichnis `examples` als `fspat.raw` abgespeichert. Im gleichen Verzeichnis befinden sich auch die zur Schätzung notwendige Strafmatrix in der Datei `pmat_bybw.raw` sowie die Kreiskennziffern der entsprechenden Kreise in der Datei `regions_bybw.raw`. Die zur Visualisierung benötigte Datei der Grenzverläufe der Kreise ist unter dem Name `bybw.bnd` gegeben.

Das in Beispiel 2 simulierte Modell besteht aus 248 Beobachtungen, so dass jeder der 124 räumlichen Funktionswerte zweimal vorkommt. Zudem werden die Beobachtungen in 31 Gruppen mit je 8 Beobachtungen eingeteilt. Jede dieser Gruppen besitzt einen zufälligen Intercept, der in der Variablen `ranint` gespeichert ist. Die Zugehörigkeit der Beobachtungen zu den Gruppen wird in der Variablen `id` definiert.

Beispiel 2: Generalisierte geoadditive gemischte Modelle.

```
> source("e:\\functions\\helpfunctions.s")
> source("e:\\functions\\ggamm.s")
> x1<-runif(248,-3,3)
> fspat<-scan("e:\\examples\\fspat.raw")
> fspat<-rep(fspat,2)
> K<-read.table("e:\\examples\\pmat_bybw.raw")
> regions<-scan("e:\\examples\\regions_bybw.raw")
> reg<-rep(regions,2)
> readbndfile("e:\\examples\\bybw.bnd","m")
> ranint<-rnorm(31,0,0.4)
```

```
> id<-1:31
> ranint<-rep(ranint,each=8)
> id<-rep(id,each=8)
> z<-rep(1,248)
> eta<-sin(x1)+0.5*fspat+ranint
> mu<-exp(eta)
> y<-rpois(248,mu)
> test<-ggamm(dep=y,smooth=x1,random=z,id=id,reg=reg,
  regions=regions,pmatrix=K,family="poisson")
> plotf(test)
> plotmap(test,m)
```

Im Verzeichnis `examples` sind neben dem hier verwendeten Beispiel auch Strafmatrizen, Regionen-Kennziffern und Boundary-Files zu den Datenanalysen aus Kapitel 6 sowie zu den Kreisen Westdeutschlands zu finden. In der Datei `readme.txt`, die ebenfalls im `examples`-Verzeichnis enthalten ist, werden die einzelnen Dateien näher erläutert.

C S-Plus-Funktionen zu LQ-Tests

In Kapitel 4 wurde beschrieben, wie die Wahrscheinlichkeiten für lokale Maxima der Likelihood-Quotienten $LQ_n(\gamma)$ und $RLQ_n(\gamma)$ beziehungsweise der asymptotischen Likelihood-Quotienten $LQ_\infty(\gamma)$ und $RLQ_\infty(\gamma)$ im Punkt Null bestimmt werden können. Außerdem wurde gezeigt, wie die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken LQ_n und RLQ_n simuliert werden können. Für die drei in Kapitel 4.2 bis 4.4 behandelten Beispiele befinden sich auf der dieser Arbeit beiliegenden CD-Rom im Verzeichnis `functions` S-Plus-Funktionen, mit deren Hilfe diese Berechnungen beziehungsweise Simulationen durchgeführt werden können. Die einzelnen Funktionen sind in der Datei `testfunctions.s` gespeichert und können über den Befehl

```
> source("e:\\functions\\testfunctions.s)
```

in S-Plus definiert werden. Im Folgenden sollen die verschiedenen Funktionen kurz vorgestellt werden.

C.1 Lokale Maxima der Likelihood-Quotienten im Punkt Null

Über Formel (4.3) und (4.6) lassen sich die Wahrscheinlichkeiten für lokale Maxima der zufälligen Funktionen $LQ_n(\gamma)$ beziehungsweise $RLQ_n(\gamma)$ im Punkt Null per Simulation bestimmen. Genauer verwendet man zur Simulation Gleichung (4.5) beziehungsweise einen analogen Ausdruck basierend auf (4.7). Man beachte, dass für das ANOVA-Beispiel aus Abschnitt 4.2 unter der Annahme, dass für den wahren Parameter $\gamma = 0$ gilt, numerisch vorteilhaftere Formeln existieren und vergleiche hierzu Gleichung (22) und (23) in Crainiceanu et al. (2002).

Zur Bestimmung der Wahrscheinlichkeiten bei endlichem Stichprobenumfang stehen die drei Funktionen

- `zero.probability.anova()`,
- `zero.probability.pspline()` und
- `zero.probability.mrf()`

zur Verfügung. Für jede dieser drei Funktionen sind eine Reihe von Argumenten zu übergeben, die im Folgenden beschrieben werden.

`zero.probability.anova()`:

`J`	Beobachtungen pro Gruppe. *Voreinstellung:* 5
`N`	Gruppenzahl. *Voreinstellung:* 5
`gamma`	Wahrer Wert des Parameters γ. *Voreinstellung:* 0
`method`	Verfahren, nach dem γ geschätzt wird. `"ML"` Maximum-Likelihood `"REML"` Restricted-Maximum-Likelihood *Voreinstellung:* `"REML"`
`nsim`	Anzahl der zur Bestimmung verwendeten Zufallszahlen. *Voreinstellung:* 1000

`zero.probability.pspline()`:

`x`	Beobachtete Ausprägungen der Kovariablen.
`nknot`	Zahl der Knoten des P-Splines. *Voreinstellung:* 20
`ord`	Ordnung der als Penalisierung verwendeten Differenzen. *Voreinstellung:* 1
`deg`	Grad des P-Splines. *Voreinstellung:* 0
`gamma`	Wahrer Wert des Parameters γ. *Voreinstellung:* 0
`method`	Verfahren, nach dem γ geschätzt wird. `"ML"` Maximum-Likelihood `"REML"` Restricted-Maximum-Likelihood *Voreinstellung:* `"REML"`
`nsim`	Anzahl der zur Bestimmung verwendeten Zufallszahlen. *Voreinstellung:* 1000

`zero.probability.mrf()`:

`K`	Strafmatrix des Markov-Zufallsfelds.
`regions`	Kennziffern der möglichen Regionen in der der Strafmatrix entsprechenden Reihenfolge.
`reg`	Ausprägungen der Kennziffern der Beobachtungen.

`gamma` Wahrer Wert des Parameters γ.
Voreinstellung: 0

`method` Verfahren, nach dem γ geschätzt wird.
`"ML"` Maximum-Likelihood
`"REML"` Restricted-Maximum-Likelihood
Voreinstellung: `"REML"`

`nsim` Anzahl der zur Bestimmung verwendeten Zufallszahlen.
Voreinstellung: 1000

Die Wahrscheinlichkeiten für lokale Maxima des asymptotischen Likelihood-Quotienten $LQ_\infty(d)$ beziehungsweise des asymptotischen Restricted-Likelihood-Quotienten $RLQ_\infty(d)$ im Punkt Null lassen sich unter H_0 über die Formeln (4.14) und (4.15) per Simulation bestimmen. Dazu verwendet man die in den Abschnitten 4.2 bis 4.4 hergeleiteten asymptotischen Eigenwerte. Für das ANOVA-Beispiel aus Abschnitt 4.2 fallen die Wahrscheinlichkeiten für lokale Maxima im Punkt Null mit den Wahrscheinlichkeitsmassen der entsprechenden asymptotischen Verteilungen im Punkt Null zusammen (Crainiceanu et al. (2002), Abschnitt 4.1) und lassen sich daher einfacher und exakt aus (4.16) und (4.17) berechnen. Insgesamt stehen wieder drei Funktionen zur Verfügung:

- `asy.zero.probability.anova()`,
- `asy.zero.probability.pspline()` und
- `asy.zero.probability.mrf()`.

Man beachte, dass zur Bestimmung der asymptotischen Eigenwerte für P-Splines angenommen wird, dass die Werte der zugrunde liegenden Kovariablen (asymptotisch) über ihren Wertebereich gleichverteilt sind und dass die Berechnungen nur für P-Splines vom Grad $l = 0$ möglich sind. Analog nimmt man für Markov-Zufallsfelder an, dass die Beobachtungen (asymptotisch) über die Regionen gleichverteilt sind. Aufgrund der Orthogonalität der Designmatrizen X und Z hängen die asymptotischen Aussagen für P-Splines und Markov-Zufallsfelder nicht vom Schätzverfahren ab.

Die Argumente der drei Funktionen sind im Einzelnen:

`asy.zero.probability.anova()`:

`N` Gruppenzahl.
Voreinstellung: 5

`method` Verfahren, nach dem γ geschätzt wird.
- `ML` Maximum-Likelihood
- `REML` Restricted-Maximum-Likelihood

Voreinstellung: `"REML"`

`asy.zero.probability.pspline()`:

`nknot` Zahl der Knoten des P-Splines.
Voreinstellung: 20

`ord` Ordnung der als Penalisierung verwendeten Differenzen.
Voreinstellung: 1

`nsim` Anzahl der zur Bestimmung verwendeten Zufallszahlen.
Voreinstellung: 1000

`asy.zero.probability.mrf()`:

`K` Strafmatrix des Markov-Zufallsfelds.

`nsim` Anzahl der zur Bestimmung verwendeten Zufallszahlen.
Voreinstellung: 1000

C.2 Asymptotische Verteilung der Likelihood-Quotienten-Teststatistiken

Die asymptotischen Verteilungen der Likelihood-Quotienten-Teststatistiken LQ_n und RLQ_n sind durch die Gleichungen (4.12) und (4.13) gegeben. Die Simulation der Verteilungen kann dann mit Hilfe der für die drei Beispiele in den Abschnitten 4.2 bis 4.4 hergeleiteten asymptotischen Eigenwerte wie in Algorithmus 6 beschrieben erfolgen. Zur genauen Bestimmung dieser Eigenwerte für P-Splines und Markov-Zufallsfelder gelten die gleichen Bemerkungen wie in C.1. Für das ANOVA-Beispiel existieren wieder numerisch vorteilhaftere Ausdrücke, die durch (4.16) und (4.17) gegeben sind und mit deren Hilfe die Verteilung einfacher simuliert werden kann. Man beachte auch wieder, dass die asymptotischen Verteilungen für P-Splines und Markov-Zufallsfelder nicht vom gewählten Schätzverfahren abhängen, LQ_n und RLQ_n also die gleiche asymptotische Verteilung besitzen.

Wieder stehen drei Funktionen zur Verfügung, die mit

- `sim.distribution.anova()`,

- `sim.distribution.pspline()` und
- `sim.distribution.mrf()`

bezeichnet sind. Die zu übergebenden Argumente der Funktionen sind im Folgenden zusammengestellt:

`sim.distribution.anova()`:

`N`	Gruppenzahl. *Voreinstellung:* 5
`nsim`	Anzahl der zu simulierenden Zufallszahlen. *Voreinstellung:* 1000
`qqplot`	Wahrheitswert, der angibt, ob QQ-Plots der simulierten Verteilungen unter der Bedingung $LQ_\infty > 0$ beziehungsweise $RLQ_\infty > 0$ gegen die χ^2_1-Verteilung gezeichnet werden sollen. *Voreinstellung:* `T`
`quantiles`	Vektor der zu bestimmenden Quantile der asymptotischen Verteilungen. *Voreinstellung:* 0.95

`sim.distribution.pspline()`:

`nknot`	Zahl der Knoten des P-Splines. *Voreinstellung:* 20
`ord`	Ordnung der als Penalisierung verwendeten Differenzen. *Voreinstellung:* 1
`nsim`	Anzahl der zu simulierenden Zufallszahlen. *Voreinstellung:* 1000
`ndgrid`	Anzahl der zur Simulation verwendeten Gitterpunkte. *Voreinstellung:* 100
`lowert`	Kleinster positiver Wert des zur Basis 10 logarithmierten Gitters. *Voreinstellung:* -3
`uppert`	Größter Wert des zur Basis 10 logarithmierten Gitters.. *Voreinstellung:* 1.2
`qqplot`	Wahrheitswert, der angibt, ob ein QQ-Plot der simulierten Verteilung unter der Bedingung $LQ_\infty > 0$ gegen die χ^2_1-Verteilung gezeichnet werden soll. *Voreinstellung:* `T`
`quantiles`	Vektor der zu bestimmenden Quantile der asymptotischen Verteilung. *Voreinstellung:* 0.95

`sim.distribution.mrf()`:

`K`	Strafmatrix des Markov-Zufallsfelds.
`nsim`	Anzahl der zu simulierenden Zufallszahlen. *Voreinstellung:* 1000
`ndgrid`	Anzahl der zur Simulation verwendeten Gitterpunkte. *Voreinstellung:* 100
`lowert`	Kleinster positiver Wert des zur Basis 10 logarithmierten Gitters. *Voreinstellung:* -3
`uppert`	Größter Wert des zur Basis 10 logarithmierten Gitters.. *Voreinstellung:* 1.2
`qqplot`	Wahrheitswert, der angibt, ob ein QQ-Plot der simulierten Verteilung unter der Bedingung $LQ_\infty > 0$ gegen die χ^2_1-Verteilung gezeichnet werden soll. *Voreinstellung:* `T`
`quantiles`	Vektor der zu bestimmenden Quantile der asymptotischen Verteilung. *Voreinstellung:* 0.95

Jede der Funktionen liefert eine Liste zurück, die aus den folgenden Komponenten besteht:

`p0`	Wahrscheinlichkeitsmasse der asymptotischen Verteilung im Punkt.
`rlr`	Vektor der Länge `nsim`, der die Realisationen aus der asymptotischen Verteilung der Restricted-Likelihood-Quotienten-Teststatistik RLQ_n enthält (nur bei ANOVA).
`lr`	Vektor der Länge `nsim`, der die Realisationen aus der asymptotischen Verteilung der Likelihood-Quotienten-Teststatistik LQ_n enthält.
`quantiles`	Vektor der angeforderten Quantile der asymptotischen Verteilung.

Zusätzlich wird für P-Splines und Markov-Zufallsfelder eine Fehlermeldung ausgegeben, falls der Wert auf dem oberen Rand des zugrunde gelegten Gitters in mindestens einem Fall als Realisation gewählt wurde. Dies erlaubt dem Benutzer, die Qualität der Simulation zu überprüfen und gegebenenfalls die übergebenen Optionen, insbesondere den Wert von `uppert`, entsprechend zu ändern.

Literatur

Azzalini, A. & Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships, *Journal of the Royal Statistical Society Series B* **55**: 549–557.

Becker, N. & Wahrendorf, J. (1997). *Krebsatlas der Bundesrepublik Deutschland 1981-1990*, Springer, Berlin.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society Series B* **36**: 192–236.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**: 1–59.

Biller, C. (2000a). Adaptive bayesian regression splines in semiparametric generalized linear models, *Journal of Computational and Graphical Statistics* **9**: 122–140.

Biller, C. (2000b). Bayesian varying-coefficient models (BVCM). Software description. Erhältlich unter www.stat.uni-muenchen.de/∼sfb386.

Biller, C. (2000c). Bayesianische Ansätze zur nonparametrischen Regression. Dissertation, Universität München.

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

Blot, W. J., Devesa, S. S., McLaughlin, J. K. & Fraumeni Jr., J. F. (1994). Oral and pharyngeal cancers, *in* R. Doll, C. S. Muir & J. F. Fraumeni Jr. (eds), *Cancer Surveys: Trends in Cancer Incidence and Mortality*, Vol. 19/20, Cold Spring Harbor Laboratory Press, New York, pp. 23–42.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistial Association* **88**: 9–25.

Brezger, A. (2000). Bayesianische P-Splines. Diplomarbeit, Universität München. Erhältlich unter www.stat.uni-muenchen.de/∼brezger.

Brezger, A., Kneib, T. & Lang, S. (2002). BayesX - Software for bayesian inference based on markov chain monte carlo simulation techniques. Erhältlich unter www.stat.uni-muenchen.de/∼lang/bayesx/bayesx.html.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical Linear Models*, Sage, Newbury Park.

Cantoni, E. & Hastie, T. (2002). Degrees of freedom tests for smoothing splines, *Biometrika* **89**: 251–263.

Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**: 671–681.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Crainiceanu, C. M. & Ruppert, D. (2002). Asymptotic distribution of likelihood ratio tests in linear mixed models.

Crainiceanu, C. M., Ruppert, D. & Vogelsang, T. J. (2002). Probability that the MLE of a variance component is zero with applications to likelihood ratio tests.

Cressie, N. (1993). *Statistics for spatial data*, Wiley, New York.

Currie, I. D. & Durban, M. (2002). Flexible smoothing with P-splines: a unified approach, *Statistical Modelling* **4**: 333–349.

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables, *Applied Statistics* **29**: 323–333. Algorithm AS 155.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). Automatic bayesian curve fitting, *Journal of the Royal Statistical Society Series B* **60**: 333–350.

Diggle, P. J., Liang, K.-Y. & Zeger, S. L. (1994). *Analysis of Longitudinal Data*, Clarendon Press, Oxford.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties, *Statistical Science* **11**: 89–121.

Fahrmeir, L. & Lang, S. (2001a). Bayesian inference for generalized additive mixed models based on markov random field priors, *Journal of the Royal Statistical Society Series C* **50**: 201–220.

Fahrmeir, L. & Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorial time-space data, *Annals of the Institute of Statistical Mathematics* **53**: 11–30.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer, New York.

Farebrother, R. W. (1990). The distribution of a quadratic form in normal variables, *Applied Statistics* **39**: 294–309. Algorithm AS 256.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion), *The Annals of Statistics* **19**: 1–141.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**: 57–68.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models, *International Statistical Review* **55**: 245–259.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**: 383–385.

Harville, D. A. (1976). Extensions of the Gauss-Markov theorem to include the estimation of random effects, *The Annals of Statistics* **4**: 384–395.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**: 320–338.

Hastie, T. J. (1996). Pseudosplines, *Journal of the Royal Statistical Society Series B* **58**: 379–396.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

Hastie, T. J. & Tibshirani, R. J. (1993). Varying-coefficient models (with discussion), *Journal of the Royal Statistical Society Series B* **55**: 757–796.

Hämmerlin, G. & Hoffmann, K.-H. (1994). *Numerische Mathematik*, Springer, Berlin.

Hobert, J. P. & Casella, G. (1996). The effect of improper priors on gibbs sampling in hierarchical linear mixed models, *Journal of the American Statistical Association* **91**: 1461–1473.

Johnson, N. L. & Kotz, S. (1970). *Continous univariate distributions Volume 2*, Wiley, New York.

Kammann, E. E. & Wand, M. P. (2003). Geoadditive models, *Journal of the Royal Statistical Society Series C (to appear)* .

Kennedy Jr., W. J. & Gentle, J. E. (1980). *Statistical Computing*, Marcel Dekker, New York.

Knorr-Held, L. (1996). Hierarchical modelling of discrete longitudinal data. Dissertation, Universität München.

Knorr-Held, L. & Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps, *Biometrics* **56**: 13–21.

Kuo, B.-S. (1999). Asymptotics of ML estimator for regression models with a stochastic trend component, *Econometric Theory* **15**: 24–49.

Laird, N. M. & Ware, J. H. (1982). Random effects models for longitudinal data, *Biometrics* **38**: 963–974.

Lang, S. & Brezger, A. (2002). Bayesian P-Splines, *Journal of Computational and Graphical Statistics (to appear)* .

Lang, S. & Fahrmeir, L. (2001). Bayesian generalized additive mixed models. A simulation study, *Discussion Paper 230, Sonderforschungsbereich 386, Universität München* . Erhältlich unter www.stat.uni-muenchen.de/∼sfb386.

Liang, K.-Y. & Self, S. G. (1996). On the asymptotic behaviour of the pseudolikelihood ratio test statistic, *Journal of the Royal Statistical Society Series B* **58**: 785–796.

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society Series B* **61**: 381–400.

Lindstrom, M. J. & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data, *Journal of the American Statistical Association* **83**: 1014–1022.

Marx, B. D. & Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood, *Computational Statistics & Data Analysis* **28**: 193–209.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, Chapman & Hall, London.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models*, Wiley, New York.

Mollié, A. (1996). Bayesian mapping of disease, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Morell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood, *Biometrics* **54**: 1560–1568.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society Series A* **135**: 370–384.

Nychka, D. W. (2000). Spatial-process estimates as smoothers, *in* M. G. Schimek (ed.), *Smoothing and Regression*, Wiley, New York.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems, *Statistical Science* **1**: 502–527.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**: 545–554.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed Effects Models in S and S-Plus*, Springer, New York.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, *Biometrika* **86**: 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix, *Biometrika* **87**: 425–435.

Pruscha, H. (2000). *Vorlesungen über Mathematische Statistik*, B. G. Teubner, Stuttgart.

Rawlings, J. O., Pantula, S. G. & Dickey, D. A. (1998). *Applied Regression Analysis. A Research Tool*, Springer, New York.

Reinsch, C. H. (1967). Smoothing by spline functions, *Numerische Mathematik* **10**: 177–183.

Rüger, B. (1999). *Test- und Schätztheorie Band I: Grundlagen*, Oldenbourg Verlag, München.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion), *Statistical Science* **6**: 15–51.

Ruppert, D. (2002). Selecting the number of knots for penalized splines, *Journal of Computational and Graphical Statistics* **11**: 735–757.

Ruppert, D. & Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting, *Australian and New Zealand Journal of Statistics* **42**: 205–223.

Self, S. G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association* **82**: 605–610.

Shephard, N. (1993). Maximum likelihood estimation of regression models with stochastic trend components, *Journal of the American Statistical Association* **88**: 590–595.

Shephard, N. G. & Harvey, A. C. (1989). On the probability of estimating a deterministic component in the local level model, *Journal of time series analysis* **11**: 339–347.

Speed, T. (1991). Comment on Robertsion (1991): "That BLUP is a good thing: The estimation of random effects", *Statistical Science* **6**: 42–44.

Sprott, D. A. (1975). Marginal and conditional sufficiency, *Biometrika* **62**: 599–605.

Stone, C. J., Hansen, M., Kooperberg, C. & Troung, Y. K. (1997). Polynomial splines and their tensor products in ectended linear modelling (with discussion), *The Annals of Statistics* **25**: 1271–1470.

Stram, D. O. & Lee, J. W. (1994). Variance component testing in the longitudinal mixed effects model, *Biometrics* **50**: 1171–1177.

Stram, D. O. & Lee, J. W. (1995). Correction to "Variance component testing in the longitudinal mixed effects model", *Biometrics* **51**: 1196.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**: 82–86.

Toutenburg, H. (2003). *Lineare Modelle*, Physica Verlag, Heidelberg.

Tutz, G. (2000). *Die Analyse kategorialer Daten*, Oldenbourg Verlag, München.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer, New York.

Wand, M. P. (2002). Smoothing and mixed models.

Wang, Y. (1998a). Mixed effects smoothing spline analysis of variance, *Journal of the Royal Statistical Society Series B* **60**: 159–174.

Wang, Y. (1998b). Smoothing spline models with correlated random errors, *Journal of the American Statistical Association* **93**: 341–348.

Whittaker, E. T. (1922/23). On a new method of graduation, *Proceedings of the Edinburgh Mathematical Society* **41**: 63–75.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple penalties, *Journal of the Royal Statistical Society Series B* **62**: 413–428.

Zhang, D., Lin, X., Raz, J. & Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data, *Journal of the American Statistial Association* **93**: 710–719.

Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 07.02.2003

(Thomas Kneib)