Nichtparametrische Regressionsmodelle und deren Schätzung

Ludwig-Maximilians-Universität München
Institut: Statistik
Seminar: Einblicke in aktuelle Forschungsgebiete der Statistik

Autor:
Thomas Welchowski (5. Semester)
Korreferent:
Matthias Faber
Betreuer:
Andrea Wiencierz

19. Januar 2012

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

GAM = Generalized additive model
GCV = Generalized cross validation
GLM = Generalized linear model
IRLS = Iterative reweighted least squares
P-IRLS = Penalized iterative reweighted least squares
TPRS = Thin Plate Regression Splines
WMSE = Weighted mean squared error

1 Einleitung

Unter nichtparametrischen Methoden wird in dieser Arbeit die Modellierung von nichtparametrischer Regression verstanden. Der wesentliche Unterschied zu parametrischen Modellierung (z. B. einfache lineare Regression) ist, dass die Strukturkomponente nicht direkt subjektiv durch den Anwender spezifiziert wird. Dadurch sind nichtparametrische Modelle flexibler als vergleichsweise parametrische Strukturen, da diese weniger Annahmen benötigen. Man gibt lediglich eine allgemeine Funktionsklasse an, welche dann anhand der Daten geschätzt wird. Dies ist insbesondere sinnvoll, falls bei einer Problemstellung keine hinreichenden konkreten Zusammenhänge oder Theorien bestehen, welche eine genaue Spezifikation des Prädiktors erlauben. Anstatt dessen schätzt man Funktionen, welche sich als Linearkombinationen darstellen lassen. Siehe dazu Kapitel 2 oder (Fahrmeir et al., 2009, vgl. S. 320-340).

Der Schwerpunkt dieser Arbeit liegt in der Erklärung des Modellierungsansatzes GAM (generalized additive model) und basiert im wesentlichen auf dem Buch Generalized additive models Wood (2006). Zuerst wird der allgemeine Aufbau eines GAMs erläutert. Im darauf folgenden Kapitel ist aufgezeigt, wie GAMs als penalisierte GLMs aufgefasst werden können. Dann wird der praktische Algorithmus zur Schätzung allgemein erläutert und eine spezielle Form des Glättungssplines theoretisch dargestellt: Thin Plate Regression Splines. Im nächsten Kapitel geht es um die Schätzungsmöglichkeiten für den Glättungsparameter. Am Ende des Theorieteils folgt noch eine kurzes Kapitel über Konfidenzbänder. Diese im Detail beschriebenen Verfahren werden dann im praktischen Teil anhand von simulierten Daten im Vergleich zu mit einem parametrischen Modell GLM (generalized linear model) dargestellt.

2 Schätzung nichtparametrischer Modelle: GAM

Dieses Kapitel beschäftigt sich mit GAM´s und deren Schätzung. Ein GAM wird wie folgt definiert (Wood, 2006, s. S. 119):

$$g(\mu_k) = X_k^*\gamma + f_1(x_{1k}) + f_2(x_{2k}) + f_3(x_{3k}, x_{4k}) + \ldots + f_J(x_{vk});\ k = 1, \ldots n;\ j = 1, \ldots J \tag{1}$$

$$\mu_k = E(y_k|X_k^*\gamma, x_{1k}, \ldots, x_{vk}) \text{ und } f(y_k|\theta_k, \phi_k) = \exp\left\{\frac{y_k\theta_k - b(\theta_k)}{\phi_k} + c(y_k, \phi_k)\right\} \tag{2}$$

Der erste Term $X^*\theta$ ist eine parametrische Einflussgröße, wie im GLM der lineare Prädiktor, und die restlichen Funktionen f() sind nichtparametrisch. k ist die Anzahl der Beobachtungen undDieses Modell weist u. a. folgende Eigenschaften auf:

1. Wie beim GLM können damit Response Variablen beliebigen Skalenniveaus geschätzt werden. Ein Unterschied besteht darin, dass die Hauptaufgabe der Responsefunktion nicht mehr in der Datenpassung besteht, sondern in der Abbildung in den geeigneten Wertebereich der Zielvariablen.

2. Es ist ebenfalls zugelassen, dass die Funktionen $f_j()$ von mehreren Einflussgrößen abhängen.

3. Es wird eine additiver Einfluss zwischen den einzelnen nichtparametrischen Funktionen und der Zielvariablen unterstellt.

Die Annahme der additiven Struktur wird aufgrund des Dimensionsfluchs verwendet (Fahrmeir et al., 2009, s. S. 395-396). Gegeben sei ein Würfel mit Daten der Dimension q und Kantenlänge o=1. Welche Kantenlänge l muss nun ein Teilwürfel haben, um einen Anteil p der Daten zu umfassen? Im trivialen Fall für q=1 gilt l=p. Für q=2 wird die Fläche mit $l^2 = p$ ausgedrückt. Im Allgemeinen Fall q gilt $l = p^{1/q}$. Zum Beispiel bei q=5 steigt l auf 0,63 im Vergleich zu l=0,1 für q=1. Da nichtparametrische Modelle ziemlich komplex sind, erhöht sich der computationale Ressourcenverbrauch bei höheren Dimensionen deutlich. Es ist schon ein sehr großer Stichprobenumfang nötig um eine Funktion $f(x_1, x_2, ..., x_j)$ mit ähnlicher Genauigkeit schätzen zu können wie $f(x_1)$. Denn der Stichprobenumfang, welcher nötig ist, um ungefähr den identischen MSE in höheren Dimensionen zu erreichen, steigt exponentiell mit der Dimension q an (Wasserman, 2006, vgl. S. 58). Bei Funktionen mit mehr als 4 Einflussgrößen ist die Modellierung meist zu zeitintensiv. Deshalb wird bei einem GAM die allgemeinere Struktur $f(x_1, x_2, ..., x_j)$ durch eine additive Struktur $f(x_1, x_2, ..., x_j) = f_1(x_1) + f_2(x_2) + ... + f_j(x_j)$ ersetzt. Dies ist zwar nicht mehr so flexibel, aber dafür ist die Schätzung effizienter möglich, da der Dimensionsfluch vermieden wird (Kauermann, 2006, s. S. 144). Außerdem lässt sich die additive Struktur besser interpretieren. Man kann sich in einer Grafik den additiven Einfluss von einzelnen Kovariablen auf die Zielgröße ansehen. Im folgenden Kapitel wird eine allgemeine Form der Spezifikation der unbekannten Funktion f() formal dargestellt.

2.1 Konstruktion von GAMs als penalisierte GLMs

Damit die Funktion f() sinnvoll geschätzt werden kann, wird diese in eine lineare Form gebracht. Basierend auf dem Ansatz der Gleichung 1 werden die nichtparametrischen Funktionen wie folgt definiert: (Wood, 2006, s. S. 163-164)

$$f_j(x_j) = \sum_{i=1}^{q_j} \beta_{ji} b_{ji}(x_j) \tag{3}$$

$b_{ji}(x_j)$ stellt die i-te Basisfunktion der j-ten glatten Funktion dar. Damit ist es eine alternative Darstellungsform von Vektoren. Es wurde gezeigt, dass es sich bei einem Splineraum um einen d-dimensionalen Vektorraum handelt. (Fahrmeir et al., 2009, vgl. S. 298) d ist die Anzahl aller Variablen des Modells, d.h. die Zielgröße und die Einflussgrößen. Deshalb ist die Bezeichnung als Basis im Sinne der linearen Algebra gerechtfertigt. Die Basisfunktion bestimmt den Raum der zulässigen Funktionen. q_j ist dabei die maximale Anzahl der Basisfunktionen der j-ten glatten Funktion und β_{ji} enspricht den zu schätzenden Koeffizienten ohne Penalisierung. Im folgenden wird die ganz einfache Basis des Polynomspline exemplarisch definiert (Fahrmeir et al., 2009, s. S. 295):

> Eine Funktion $f : [a, b] \longrightarrow \mathbb{R}$ heißt Polynomspline vom Grad l ≥ 0 mit m Knoten $a = k_1 < \ldots < k_m = b$, falls sie die folgenden Bedingungen erfüllt:
>
> 1. f(x) ist (l-1)-mal stetig differenzierbar. Für l=1 entspricht dies der Forderung, dass f(x) stetig ist, für l=0 werden keine Glattheitsanforderungen an f(x) gestellt.
> 2. f(x) ist auf den durch die Knoten gebildeten Intervallen $[k_a, k_{a+1})$ ein Polynom vom Grad l.

Für den Fall von l=3 handelt es sich um Polynom 3. Grades mit Knotenpunkten. Bei der Wahl der Knoten ist insbesondere auf die Anzahl und die Position zu achten. (Fahrmeir et al., 2009, vgl. 301-302) Die Lage der Knoten kann in gleichmäßigen Abständen, nach Quantilen oder gezielt gesetzt werden. Je höher die Knotenanzahl ist, desto rauer und je weniger Knoten gesetzt werden, desto glatter wird die geschätzte Kurve. Ein Polynomspline wird nun an einer Graphik von Längsschnittdaten aus dem sozioökonomischen Panel zwischen 1986 und 2007 mit 33841 Personen veranschaulicht:

In der Abbildung 1 ist auf der x-Achse das Alter in Jahren und auf der y-Achse sieht man den Zufriedenheitsscore auf einer ordinalen Skala von 0 bis 10. Die gestrichelte Kurve ist ein Modell mit einem linear, gemischtes Modell mit kubischen Polynomen und die durchgezogene Kurve wurde durch ein semiparametrisches, linear, gemischtes Modell mit einem kubischen Polynomspline gefittet. Bei einem gemischten Modell variert der Einfluss der Kovariablen auf die Zielgröße gruppenspezifisch, d.h. die Beobachtung hängt von der zugehörigen Gruppe des gezogenen Individuums ab. Man kann deutlich anhand der Grafik erkennen, dass der semiparametrische Ansatz eine midlife crisis um 50 Jahre identifiziert, während sich dies bei einem rein parametrischen Ansatz nicht erkennen lässt (Wiencierz et al., 2011, vgl. S. 1729-1730).

Nach dem vorherigen Beispiel wird nun eine beliebige, gegebene Basisfunktion angenommen. Damit das GAM identifizierbar bleibt muss jede Glättungsfunktion eine Zentrierungsrestriktion enthalten. Mit der Bedingung von zentrierten Modellmatrizen kann die Gleichung auch explizit als GLM formuliert werden:

$$\text{Identifizierbarkeitsbedingung: } 1^T \tilde{X}_j \tilde{\beta}_j = 0 \tag{4}$$

$$\tilde{X}_j = b_{ji}(x_{ji}) \tag{5}$$

$$\tilde{\beta}_j = \left[\beta_{j1}, \ldots, \beta_{jq_j}\right]^T \tag{6}$$

$$\text{Reparametrisierte Linkfunktion: } g(\mu_k) = X_k \beta;\ k = 1, \ldots, n \tag{7}$$

$$X = [X^* : X_1 : X_2 : \cdots] \tag{8}$$

$$\beta^T = \left[\gamma^T, \beta_1^T, \beta_2^T, \cdots\right] \tag{9}$$

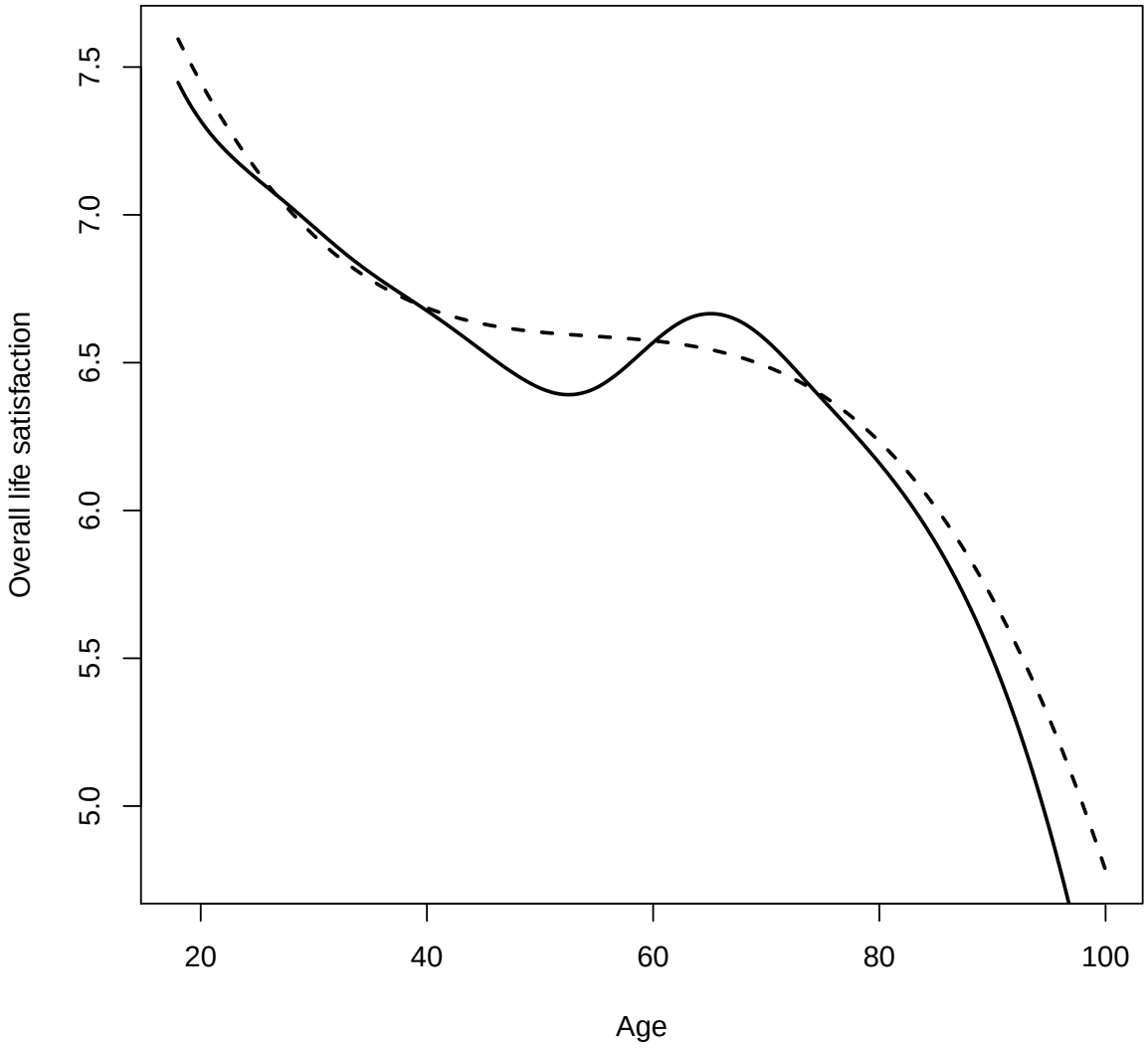

Abbildung 1: Beispiel Polynomspline

Die Notation $X = [X^* : X_1 : X_2 : \cdots]$ bedeutet, dass die Designmatrizen jeder Glättungsfunktion zu einer großen Matrix zusammengefasst wurden. Davon lässt sich dann die Likelihoodfunktion $l(\beta)$ aufstellen und mit Hilfe des IRLS-Verfahren und dem zugehörigen Fisher-Scoring Algorithmus lösen. Die sich daraus ergebende Schätzung hat den Nachteil, dass die Basisdimension vollständig ausgenutzt wird, was zum Overfitting führt. Deshalb wird ein Penalisierungsterm in die Schätzung mit einfließen, welcher zu starke Anpassungen bestraft. Normalerweise wird ein quadratische Bestrafung im Sinne von

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_j \lambda_j \beta_j^T S_j \beta_j;\ j = 1, \ldots, J \tag{10}$$

verwendet. $\beta_j^T S_j \beta_j$ ist der Penalisationsterm mit bekannter Matrix S_j. λ_j ist das Gewicht der Penalisierung. Je größer λ_j desto glatter wird die geschätzte Kurve. Dieser Parameter kann entweder festgelegt als auch geschätzt werden. Darauf wird in Kapitel 2.4 eingegangen.

Zum Abschluss dieses Kapitels wird noch auf den Dispersionsparameter ϕ der Exponentialfamilie aus dem Ansatz 1 eingegangen. Die Dispersion gibt an, wie stark die Streuung der Daten ist, denn es gilt $b(\theta) = \phi v(\mu)$ wobei $v(\mu)$ die Varianzfunktion in Abhängigkeit des Erwartungswerts ist. Um den Dispersionparameter in einem GAM zu schätzen, ist der Schätzer des Dispersionsparameters im GLM zu erweitern. Es sei die Hat-Matrix A gegeben, d.h. es gilt $\hat{\mu} = Ay$. Der Dispersionsparameter kann mit $\hat{\phi} = \dfrac{\sum_{k=1}^{n} V(\mu_k)(y_k - \mu_k)^2}{n - tr(A)}$ geschätzt werden (Wood, 2006, s. S. 166-168). $tr(A)$ sind dabei die effektiven Freiheitsgrade, d.h. die Anzahl der zu schätzenden Parameter abzüglich der Restriktionen. Im GAM kann gezeigt werden, dass gilt

$$A = (X^T W X + S)^{-1} X^T W X \tag{11}$$

2.2 Algorithmus zur Schätzung von GAMs

Ähnlich wie beim GLM gibt es auch beim GAM einen Algorithmus zur Maximierung der penalisierten Likelihood wie in Gleichung 10. Dieses Verfahren heißt P-IRLS (Penalized iterative reweighted least squares) (Wood, 2006, s. S. 165-166). Zuerst wird angenommen, dass λ_j bekannt sind. Damit kann man die Funktion in bekannte und nicht bekannte Teile neu formulieren mit $S = \sum_j \lambda_j S_j$. Nun kann man mit dem Quasi-Likelihood-Ansatz folgende Score-Funktion aufstellen:

$$S_p = \frac{\partial l_p}{\partial \beta_j} = \frac{\partial l}{\partial \beta_j} - [S\beta]_j = \frac{1}{\phi}\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} - [S\beta]_j = 0 \tag{12}$$

Zur Notation: Mit $[\cdots]_j$ ist die j-te Zeile eines Vektors gemeint. Man kann zeigen, dass durch geeignete Umformungen und bei bekannter Varianz diese Gleichung genau dem penalisieren, nichtlinearen, kleinsten Quadrate Schätzers entspricht. Diese Gleichung kann folgendermaßen abgeschätzt werden:

$$S_p(\beta) \approx \left\| \sqrt{W^{[k]}} \left(z^{[k]} - X\beta \right) \right\|^2 + \beta^T S \beta \tag{13}$$

$$w_i^{[k]} = \frac{1}{V(\mu_i^{[k]}) g'(\mu_i^{[k]})^2} \tag{14}$$

$$z_i^{[k]} = g(\mu_i^{[k]}) \left(y_i - \mu_i^{[k]} \right) + X_i \hat{\beta}^{[k]} \tag{15}$$

$$i \equiv \text{ Beobachtungen } = 1, \ldots, n$$

Dabei ist $g()$ Link-Funktion des Modells, $z^{[k]}$ ist ein Vektor von Pseudodaten und $W^{[k]}$ ist eine Diagonalmatrix mit den Elementen $w_i^{[k]}$. Die Lösung eines GLM's kann ebenfalls alternativ als gewichtete

KQ-Schätzung mit Pseudodaten aufgefasst werden. Mit diesen gegebenen Formeln läuft der Algorithmus in diesen Schritten ab:

1. Bestimme einen Startwert für $\hat{\beta}^{[0]}$
2. Bestimme unter, gegegebenen $\hat{\beta}^{[k]}$, die Pseudodaten $z_i^{[k]}$ sowie die Gewichte $w_i^{[k]}$.
3. Minimiere die Gleichung 13, um $\hat{\beta}^{[k+1]}$ zu erhalten. Erhöhe k um 1.
4. Prüfe ob ein gegebenes Abbruchkriterium erfüllt ist, z. B. stope wenn $\frac{\left\|\hat{\beta}^{[k+1]} - \hat{\beta}^{[k]}\right\|}{\left\|\hat{\beta^{[k]}}\right\|} < \varepsilon$.
5. Falls dies nicht erfüllt ist, wiederhole den Algorithmus ab dem 2. Schritt.

2.3 Thin Plate Regression Splines

Unter einem Spline versteht man ein glatte, stückweise zusammengesetzte Funktion. Glättungssplines sind eine Methode, die unbekannte Funktion f() zu schätzen. Zunächst wird eine Begründung gegeben, warum dieser Ansatz sinnvoll ist. Eine gute Begründung hierfür liefert u. a. Simon Wood (Wood, 2006, vgl. S. 142-144): Gegegeben seien $\{x_i, y_i | i = 1, \ldots n\}$ mit $x_i < x_{i+1}$. Wenn man diese Punkte interpolieren will, ist der glatteste Ansatz ein natürlicher, kubischer Spline $r(x)$. Um diesen Ansatz zu verstehen ist zunächst der Begriff der absoluten Stetigkeit zu definieren (Elstrodt, 2009, s. S. 303):

Eine messbare Funktion $F : [a, b] \to \mathbb{K}$ heißt *absolut stetig*, wenn zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass

$$\sum_{k=1}^{n} |F(\beta_k) - F(\alpha_k)| < \varepsilon \tag{16}$$

für alle $a \leq \alpha_1 < \beta_1 \leq \alpha_2 < \beta_2 \leq \ldots \leq \alpha_n < \beta_n < b$ mit $\sum_{k=1}^{n} (\beta_k - \alpha_k) < \delta$

Dies bedeutet, der Abstand zwischen zwei Funktionswerten darf nicht sehr stark von den Abständen der Funktionsargumente abweichen. Der natürliche, kubische Spline ist bis zur 2. Ableitung stetig $r''(x_1) = r''(x_n) = 0$ mit $r(x_i) = y_i$. Der natürliche, kubische Spline wird für jedes Punktpaar $[x_i, x_{i+1}]$ aus kubischen Polynomen zusammengesetzt. Von allen stetigen Funktionen auf $[x_1, x_n]$, die absolut stetig in der ersten Ableitung sind, welche die Punkte $\{x_i, y_i\}$ interpolieren minimiert $g(x)$ folgenden Ausdruck:

$$\min \left\{ \|y - f\|^2 + \lambda \int_{x_1}^{x_n} f''(x)^2 dx \right\} \tag{17}$$

$$\max \left| \tilde{f} - g \right| \leq \frac{5}{384} \max \left\{ (x_{i+1} - x_i)^4 \right\} \max \left| \tilde{f}'''' \right| \tag{18}$$

Es ist gezeigt worden, dass das Resultat der zweiten Gleichung im Bezug auf die Glattheit optimal ist. Durch zusätzliche Penalisierung kann dann ein kubischer Glättungsspline konstruiert werden. Neben diesen theoretischen Eigenschaften gab es ebenfalls Simulationen mit realen Datensätzen, bei denen Glättungssplines einen geringeren MSE aufwiesen als vergleichbare Lokalisierungstechniken (Kernel Regression). (Aydin, 2007, vgl. S. 1) Unter anderem aufgrund dieser Eigenschaften ist es sinnvoll Glättungssplines für die Schätzung zu verwenden, da diese glatte Funktionen sind und die Daten sehr genau interpolieren können.

In diesem Abschnitt wird die Variante des Thin Plate Regression Splines (TPRS) formal dargestellt. Zunächst wird die Theorie hinter Thin Plate Splines erklärt und im Anschluss daran wie man diesen Ansatz praktisch approximieren kann. Das Grundmodell des Thin Plate Splines kann formal in dieser Form beschrieben werden:

$$y_k = g(x_k) + \varepsilon_k; \; k = 1, \ldots, n \tag{19}$$

n ist die Anzahl der Beobachtungen und x_i ist ein d-dimensionaler Vektor. Die Funktion $\hat{f}$ wird über folgende Gleichung gefunden:

$$\min \left\{ \|y - f\|^2 + \lambda J_{md}(f) \right\} \tag{20}$$

y ist wie in in der Gleichung 19 der n-dimensionale Vektor, welcher den Beobachtungen der Zielvariablen entspricht und die Funktion ist $f = (f(x_1), f(x_2), \ldots, f(x_n))^T$. J_{md} misst die Rauheit der Funktion f, dass heißt wenn dieser Wert groß ist, wird die geschätzte Kurve flacher und umgekehrt. Die Index d ist die Anzahl der Basisfunktionen und m bestimmt die höhe des Ableitungsgrads von J_{md}. λ ist ein Gewichtungsfaktor. Bei diesem Ansatz wird im ersten Term die Anpassung der Kurve berücksichtigt und andererseits im zweiten Term glättet die Kurve, damit diese nicht nur die Daten interpoliert.

Die allgemeine Form von J_{md} ist kompliziert darzustellen, deswegen wird hier auf ein Beispiel für m=d=2 verwiesen:

$$J_{22} = \int\int \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2 \tag{21}$$

m bestimmt den Rauheitsgrad, d.h. bis zur welcher Ableitung die Penalisierung berechnet wird. Dabei werden alle möglichen Kombinationen der Kovariablen mit Dimension d quadratisch zusammengezählt und danach integriert. m ist unter folgender Restriktion wählbar: 2m > d. Unter diesen Bedingungen lässt sich zeigen, dass sich $\hat{f}(x)$ folgendermaßen darstellen lässt:

$$\hat{f}(x) = \sum_{k=1}^{n} \delta_k \eta_{md}(\|x - x_k\|) + \sum_{i=1}^{M} \alpha_i \phi_i(x) \tag{22}$$

$$M = \binom{m+d-1}{d} \quad \eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}\, r^{2m-d} \log(r); & \text{d gerade} \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}\, r^{2m-d}; & \text{d ungerade} \end{cases} \tag{23}$$

Die Gammafunktion lässt sich auch als Fakultätsfunktion darstellen: $\Gamma(m+1) = m! = m\Gamma(m)$ mit $m \in \mathbb{N}$ sowie $\Gamma(1) = 1$. ϕ_i sind linear unabhängige Polynome vom Grad kleiner als m. Diese spannen den Nullraum von J_{md} auf und sind damit vollständig glatt. Beispielsweise für m=d=2 gilt $\phi_1(x) = 1;\ \phi_2(x) = x_1;\ \phi_3(x) = x_2$. δ und α sind Vektoren von zu schätzenden Koeffizienten. δ unterliegt der linearen Restriktion $T^T\delta = 0$ mit $T_{ij} = \phi_j(x_i)$.
Die bisher dargestellte Schätzung mit Hilfe eines Thin Plate Splines hat auch ihre Nachteile: Die Algorithmen sind in dieser Form sehr rechenintensiv. Deswegen verwendet man in der Praxis eine Approximation mit Hilfe von Thin Plate Regression Splines: Bei dieser Approximation werden die zu schätzenden Matrizen, $F_{ij} \equiv \eta_{md}(\|x_i - x_j\|)$ mit Hilfe der Spektralzerlegung (auch als QR-Zerlegung bekannt) aufgeteilt. Nach diesem Verfahren sind die restlichen Berechnungsschritte deutlich schneller durchführbar. Eine Spektralzerlegung basiert auf der Bestimmung der Eigenwerte einer Matrix. (Harville, 1997, vgl. S. 516, 518, 550). Eigenwerte sind λ, mit einer nxn Matrix C und einem zugehörigen Eigenvektor z, welche diese Gleichung erfüllen:

$$Cz = \lambda z \tag{24}$$

Die Lösung dieser Gleichung sind gleichzeitig die Nullstellen des sogenannten charakteristischen Polynoms.

$$p(\lambda) = (-1)^n (\lambda - \lambda_1)^{\gamma_1} \ldots (\lambda - \lambda_{n-1})^{\gamma_{n-1}} (\lambda_n)^{\gamma_n} = 0 \tag{25}$$

γ gibt die algebraische Multiplizität an, dass heißt die Vielfachheit mit jener der Eigenwert im charakterischen Polynom auftritt. Nimmt man eine symmetrische ($D^T = D$), positv definite ($D_{ij} > 0\ \forall i, j$) Matrix mit Dimension nxn an, dann gilt:

$$D = Q \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix} Q^T \tag{26}$$

Dabei ist Q eine nxn orthogonalisierte Matrix und F eine Diagonalmatrix mit Eigenwerten als Einträge. Orthogonale Matrizen sind quadratische, reelle Matrizen und weisen unter anderem folgende Eigenschaften auf: $Q^TQ = E$ und $Q^T = Q^{-1}$ (Harville, 1997, s. S. 84). Durch Thin Plate Regression Splines kann eine niedrige Rangapproximation des Thin Plate Splines dargestellt werden. Durch einen effizienten Algorithmus von Lanczos (Demmel, 1997, vgl. S. 366 - 382), zur Eigenwertzerlegung von Matrizen, kann die Komplexität des Algorithmus in O-Notation von $O(n^3)$ zu $O(kn^2)$ reduziert werden. k bezeichnet in diesem Falle den zu approximierenden Rang. Die O-Notation gibt die Komplexitätsklasse (linear, quadratisch, exponentiell) an, d.h. wie stark sich die Komplexität einer Berechnung in Abhängigkeit von Einflussgrößen verändert. Diese Approximation ist gleichzeitig relativ nahe an den Optimalitätseigenschaften des Thin Plate Splines dran, so dass diese Methode ebenfalls praktisch anwendbar ist. Zusammenfassend hat die Thin Plate Regression Spline Basis folgende Eigenschaften (Wood, 2006, s. S. 150-156):

- Es ist nicht erforderlich Knotenpunkte festzulegen → Modell enthält weniger Subjektivität
- Es können Funktionen mit mehreren Kovariablen geschätzt werden.
- Flexiblere Modellierung durch Wahl eines Ableitungsgrades bei der Messung der Funktionsvolatilität
- Rotationsinvarianz (isotropisch), d.h. unabhängig von der Richtung im Datenkörper werden die zu schätzenden Koeffizienten gleich gewichtet. Dies ist besonders für Interaktionen von räumlichen Daten mit identischen Maßeinheiten sinnvoll. (Wood, 2006, vgl. S. 224)
- Approximation als Thin Plate Regression Splines verfügbar, welche im Wesentlichen die positiven Eigenschaften von Thin Plate Splines erhält und effizienter berechenbar ist
- Keine Invarianz gegenüber Skalentransformationen von Variablen
- Für große Datensätze nicht gut geeignet, da TPRS trotz der Approximation rechenintensiv ist (Wood, 2006, vgl. S. 244)

2.4 Schätzung des Glättungsparameters

Wie bereits in den vorherigen Kapiteln erwähnt, bestimmt der Glättungsparameter λ wie stark die nichtparametrische Schätzung penalisiert wird. Es gibt einen Konflikt zwischen Glattheit der geschätzten Kurve und der Datentreue. Eine glatte Kurve hat zwar eine geringe Variabilität in der Schätzung, dafür aber eine hohe Verzerrung gegenüber den Daten. Bei einer sehr gut angepassten Kurve ist zwar der Bias gering, dafür muss man aber eine hohe Variabilität der Schätzung in Kauf nehmen. Falls $\lambda \to 0$ dann ergibt sich ein, bei genügend großer maximale Basisdimension, ein nahezu perfekter Fit. Wenn hingegen $\lambda \to \infty$ nähert sich die Lösung einer Geraden an, denn dies ist die glatteste Kurve, unabhängig von der wählbaren Funktion f in der Gleichung 20. (Sprent and Smeeton, 2001, vgl. S. 311)

Gegeben der Skalenparameter einer exponentialverteilten Zielvariable sei unbekannt. Dann lässt sich λ durch ein Kreuzvalidierungsverfahren schätzen. Die Idee des Kreuzvalidierungsverfahrens besteht darin, jeweils ein einzelnes Datum aus dem Datensatz zu entfernen und mit den restlichen Informationen genau jenes zu schätzen. Dies wird solange durchgeführt, bis alle Daten prognostiziert wurden. Danach werden die quadratischen Abweichungen zu den Originaldaten aufsummiert: $\sum_{k=1}^{n} \left(y_k - \hat{\mu}_k^{[-k]} \right)^2$ Zudem konnte gezeigt werden, dass Akaikes Informationskriterium und Kreuzvalidierungsverfahren asymptotisch äquivalent sind. (Wood, 2006, vgl. S. 170 aus Stone, M. (1977). An asymptotic equivalence of choise of model by cross-validation and akaikes criterion)

Allerdings hat die gewöhnliche Kreuzvalidierung 2 Nachteile (Wood, 2006, vgl. S. 171-174): Einerseits ist diese Methode ziemlich computerintensiv, da bei einem GAM viele Glättungsparameter geschätzt werden müssen. Des Weiteren ist die gewöhnliche Kreuzvalidierung nicht invariant gegenüber orthogonalen Transformationen der Daten (Ziel- und Einflussgrößen). Das Problem dabei ist, dass die Schätzungen der Parameter, die effektive Anzahl an Freiheitsgraden und der erwartete Prognosefehler invariant bezüglich Rotationen von $y - X\hat{\beta}$ mit irgendeiner orthogonalen Matrix Q sind, aber gleichzeitig sind dies die Diagonalelemente von A nicht. Die gewöhnliche Kreuzvalidierung ist also abhängig von den Diagonalelementen von der Hat-Matrix A. Die Lösung dieser Probleme führt zur Generalisierten Kreuzvalidierung auch GCV genannt. Wie in Kapitel 2.1 sei nun A wieder die Hat-Matrix. Die Idee ist, eine Kreuzvalidierung von einer Rotation des der ursprünglichen Gleichung durchzuführen, welche invariant gegenüber orthogonalen Transformationen ist. A_Q ist die transformierte Hat-Matrix A und B sei eine Matrixzerlegung der Form, dass gilt $BB^T = A$ so gilt:

$$A_Q = QBB^TQ^T; BB^T = A \tag{27}$$

Falls nun jede Zeile von QB die gleiche euklidische Länge hat, gilt $A_{ii} = \frac{tr(A_Q)}{n}$ da $tr(A_Q) = tr(QAQ^T) = tr(AQQ^T) = tr(A)$. Die letzte Gleichungs lässt sich ableiten, da für alle Matrizen mxn C und nxm D folgendes gilt: $tr(CD) = tr(DC)$ (Harville, 1997, s. S. 50). Die Matrix Q lässt sich durch iterative Givens Rotation erzeugen (Golub and Loan, 1996, vgl. S. 215, 216):

$$G(i,k,\theta) = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \tag{28}$$

$$c = cos(\theta) \text{ und } s = sin(\theta) \tag{29}$$

Dabei ist die i-te Zeile mit den Einträgen c, s und die k-te Zeile mit den Einträgen $-s$, c. Givens Rotationen sind orthogonal und rotieren z. B. $x \in \mathbb{R}^n;\ y = G(i,k,\theta)^T x$ gegen den Uhrzeigensinn um den Winkel θ. Unter diesen Annahmen gilt:

$$y_j = \begin{cases} cx_i - sx_k, & j = i \\ sx_i + cx_k, & j = k \\ x_j, & j \neq i,k \end{cases} \tag{30}$$

Die Rotation wird für zwei Zeilen so lange durchgeführt, bis die Länge der Zeilen identisch sind. Dieser Punkt wird erreicht, wenn man mit $\theta = 0$ beginnt und den Winkel sukzessive erhöht. Denn bei 90 Grad sind die Längen beider Zeilen vertauscht. Dies wiederholt man für alle weiteren Zeilen. Praktisch gesehen, muss diese Transformation nicht durchgeführt werden, denn das Ergebnis der Diagonaleinträge ist bereits vorher bekannt. Eine approximative Form des GCV Kriteriums lautet wie folgt:

$$V_g = \frac{n\ D(\hat{\beta})}{(n - tr(A))^2} \tag{31}$$

mit $D(\hat{\beta})$ als Devianz des Modells. Die Devianz ist ein Maß für die Anpassung des Modells an die Daten. Als Algorithmus zur Schätzung des Minimierungsproblems wird eine erweiterte Form des P-IRLS Verfahrens verwendet, siehe Kapitel 2.2. Zusätzlich zum bisherigen Ansatz wird die erste Ableitung der

Linkfunktion als auch die zweite Ableitung der Varianzfunktion benötigt. Für Details zum Algorithmus sei auf Simon Wood verwiesen. (Wood, 2006, vgl. S. 182 - 184).

O'Sullivan hat in Monte Carlo Simulationen von Binomial- und Poissonmodellen gezeigt, dass die theoretische Grundlage des GCV in der Praxis zu sinnvollen Ergebnissen führt. (O'Sullivan et al., 1986, vgl. S. 90 - 101) Es wurde die Effizienz des GCV gemessen, indem der gewichtete, mittlere quadratische Fehler (WMSE) mit dem geschätzten Parameter λ im Verhältnis zu dem minimalen WMSE gesetzt wurde. Bei den Binomialmodell lag die der Mittelwert der Effizienz bei 0.86, beim eindimensionalen Poissonmodell bei 0.82 und beim zweidimensionalen Poissonmodell sogar bei 0.93.

2.5 Schätzung von Konfidenzbändern

In diesem Kapitel geht es um die Konstruktion von Konfidenzbändern für die geschätzte Funktion. Im Unterschied zu einem Konfidenzintervall wird hier ein Bereich um die gesamte geschätzte Regressionskurve angegeben, welcher mit Wahrscheinlichkeit $1-\alpha$ die neu geschätzte Regressionskurve überdeckt. Bei einem Konfidenzintervall interessiert man sich im Gegensatz zum Konfidenzband nur für einzelne Parameter oder Schwankungen der Zielgröße für ein fest gegebenes x. Für die Interpretation ist es hilfreich nicht nur eine Punktschätzung für gegebene Kovariablen zu haben, sondern ebenfalls die Variabilität dieser Schätzung zu messen. Dazu benötigt man Konfidenzintervalle. Im vorangegangen Kapitel 2.4 wurde gezeigt, wie man den Glättunsparameter schätzen kann. Dieser Ansatzt lässt sich auch unter dem Gesichtspunkt betrachten, dass man hier Vorwissen verwendet. Man nimmt beispielsweise an, dass die wahre Funktion eine glatte Kurve und keine beliebig rauhe Kurve ist, welche nur die Daten interpoliert. GAMs sind ziemlich komplex und um die Schätzungen zu verbessern, bietet sich ein Bayes-Ansatz an. Zunächst wird der allgemeine Bayes Ansatz nochmal kurz erläutert: (Hothorn et al., 2011, s. S. 65-68)

$$\textbf{Priori Verteilung: } \text{Zufallsveriable } \vartheta \sim \mathbb{P}_\vartheta;\ f_\vartheta(\vartheta) \tag{32}$$

$$\textbf{Posteriori Verteilung: } \text{Zufallsveriable } X \sim \mathbb{P}_{X|\vartheta};\ X = x \text{ Realisation von X} \tag{33}$$

$$f_{\vartheta|X=x} = \frac{f_X\left(x|\vartheta\right) f_\vartheta\left(\vartheta\right)}{\int f_{X|\vartheta}(x|\vartheta) f_\vartheta(\vartheta) d\nu(\vartheta)}$$

$$\textbf{Kredibilitätsintervall: } \int_{[\underline{b},\bar{b}]} f_{\vartheta|x}(\vartheta|x) d\nu(\vartheta) = 1-\alpha \tag{34}$$

Zunächst definiert man eine a priori Verteilung, um mit dieser zusätzlichen Information eine a posteriori Dichte abzuleiten. Aus der Posteriori-Verteilung kann dann ein Kredibilitätsintervall abgeleitet werden. Falls für den Stichprobenumfang $n \to \infty$ gilt, dann konvergiert die Wahrscheinlichkeit den wahren Parameter in der Posteriori-Verteilung zu erhalten gegen 1. Im Falle des GAM wird folgende Priori-Verteilung definiert:

$$f_\beta(\beta) \propto \exp\left(-\frac{1}{2}\beta^T \sum S_i/\tau_i \beta\right) \tag{35}$$
$$\text{mit } \tau_i \equiv \text{ Dispersion}$$

Bei dieser Dichte geht man davon aus, dass jedes i des Rauheitsgrads $\beta^T S_i \beta$ eine unabhängige, Zufallsvariable aus der Exponentialfamilie ist. Zusätzlich geht die Annahme ein, dass ein glatteres Modell wahrscheinlicher als ein raues ist. Modelle mit gleicher Glättung haben identische Wahrscheinlichkeiten. Es lässt sich mit dem Satz von Bayes zeigen, dass die Posteriori Dichte der Koeffizienten β für Stichprobenumfang $n \to \infty$ multivariat normalverteilt ist (Wood, 2006, vgl. S. 185 - 190). Dabei wird $\tau_i = \lambda_i/\sigma^2$ gesetzt:

$$\beta|y \sim N\left(\hat{\beta}, \left(X^T W X + \sum \lambda_i S_i\right)^{-1} \sigma^2\right) \tag{36}$$

Mit diesen asymptotischen Eigenschaften kann die empirische, posteriori Verteilungsfunktion einer interessierenden Funktion $G(\beta)$ geschätzt werden. Dazu wird aus der multivariaten Normalverteilung, siehe Gleichung 36, die Menge $\{\beta_i^* : i = 1, \ldots, N\}$ gezogen. Diese werden dann folgendermaßen verwendet:

$$\hat{F}(g) = \frac{1}{N} \sum_{i=1}^{N} H\left(g - G(\beta^*)\right) \tag{37}$$

$$H(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{38}$$

H(g - x) ist die Heaviside Funktion, welche für Werte größer als g 1 ergibt und sonst 0. (Bracewell, 2000, vgl. S. 61 - 65). Mit festgelegten Quantilen von $\hat{F}(g)$ werden dann die bayesianischen Kredibilitätsintervalle konstruiert. Aus einigen Simulationen mit Konfidenzbändern, basierend auf unterschiedlichen Funktionen, ergab sich folgendes Ergebnis: Die Konfidenzintervalle für das komplette Modell sind zuverlässig, aber komponentenweise Intervalle können nur als grobe Orientierung gesehen werden. Dies liegt unter anderem daran, dass die Konfidenzintervalle von den Glättungsparametern abhängig sind.

3 Praktische Anwendung mit Statistiksoftware R

In diesem Abschnitt werden die vorhergehenden, theoretischen Methoden mit einer Simulation in der Statistiksoftware R dargestellt. Es wird dabei hauptsächlich das von Simon Wood geschriebene Packet *mgcv* Wood (2011) verwendet. Es soll an einem einfachen Beispiel der zusätzliche Erkenntnisgewinn von nichtparametrischen Methoden gegenüber parametrischen Modellen, wie z. B. GLM, gezeigt werden. Die simulierten Daten könnten bei einer Messung eines Störfalls, Erdbebens, Schallwellen, Herzfrequenz etc. auftreten. Die Zielvariable y als auch die Einflussgröße x werden als metrisch und stetig vorausgesetzt. Die x-Werte können als direkt aufeinanderfolgende Messzeitpunkte der Zielgröße y interpretiert werden. Man nehme an es gilt folgendes Modell:

$$y = f(x) + \varepsilon;\ \varepsilon \sim N(0,1) \tag{39}$$

$$f(x) = 1000 * \underbrace{\sin\left(\frac{\pi * x * 2}{180}\right)}_{\text{Zyklische Funktion}} * \underbrace{\exp\left(-\frac{1}{2}\frac{x^2}{100} - \ln\left(\sqrt{2\pi 100}\right)\right)}_{\sim N(0,100)} \tag{40}$$

f(x) ist der wahre Zusammenhang von x auf y und ε der standardnormalverteilte Fehler. Die wahre Funktion lässt sich im Intervall $[0, 800]$ in dieser Form graphisch darstellen:

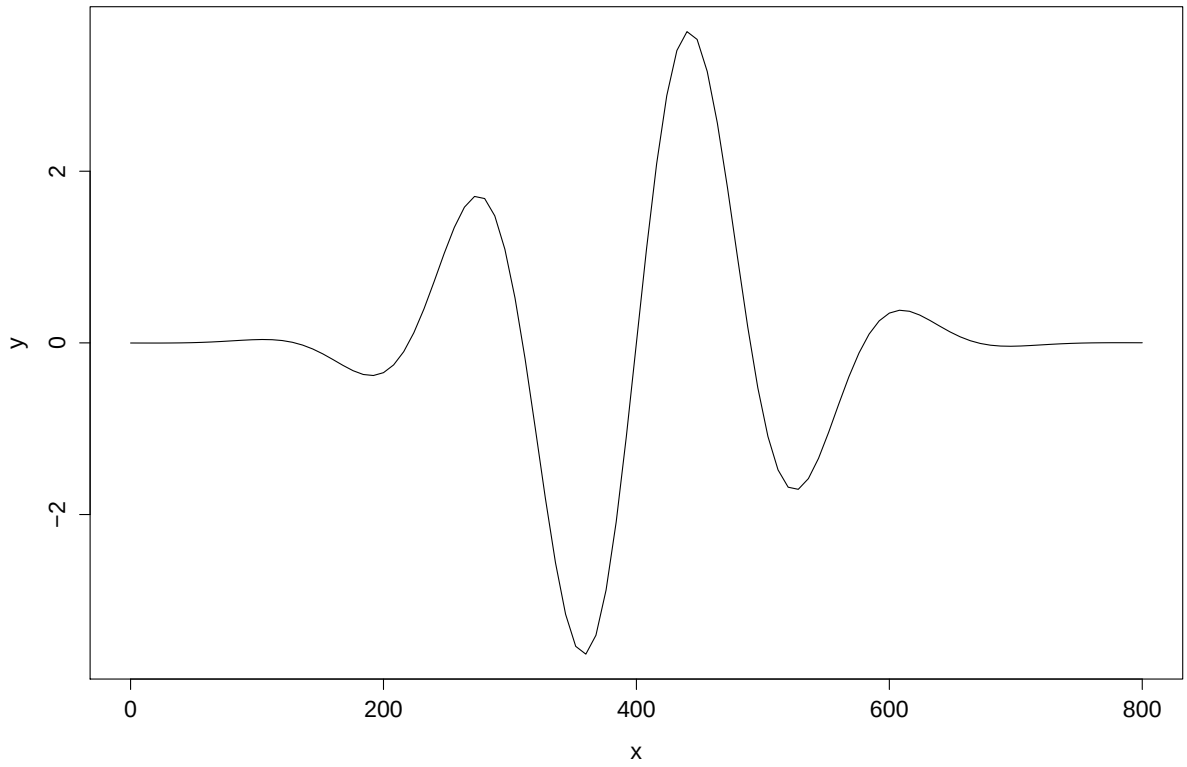

Abbildung 2: Funktionaler Zusammenhang der Simulation

Unter Hinzunahme des normalverteilten Fehlers ε ergibt sich folgende Datenlage:

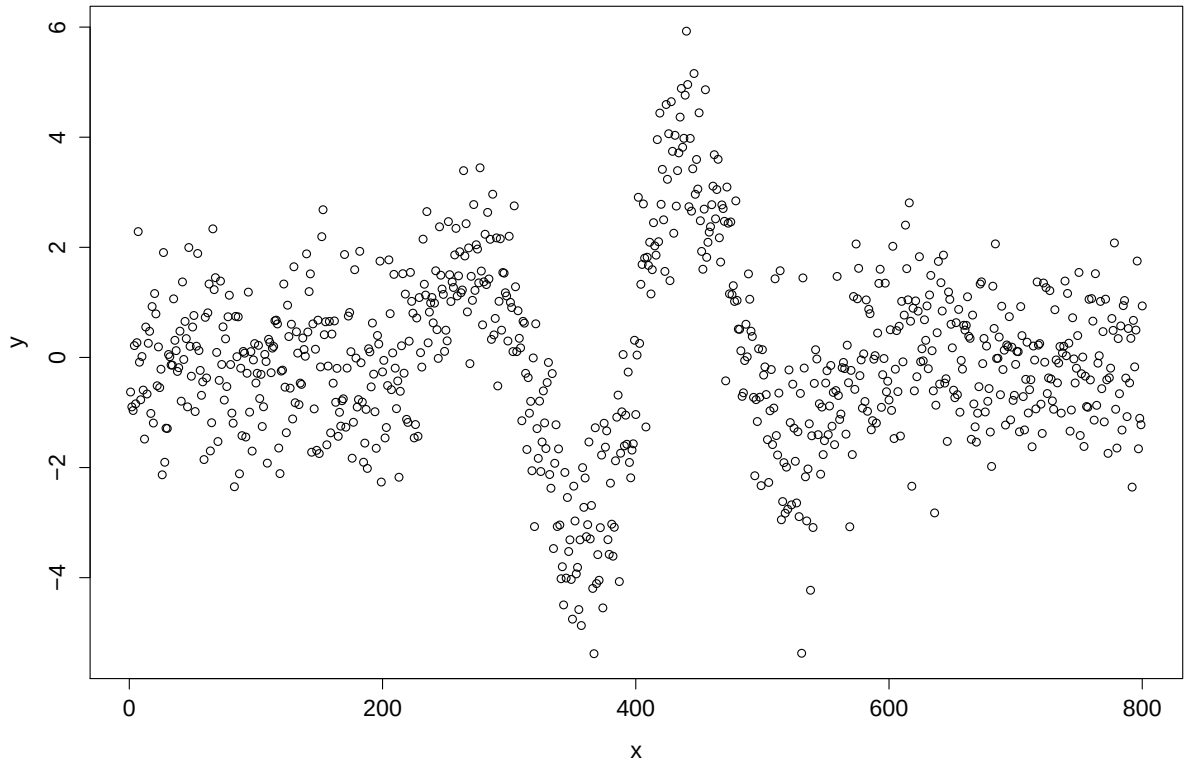

Abbildung 3: Daten aus der Simulation

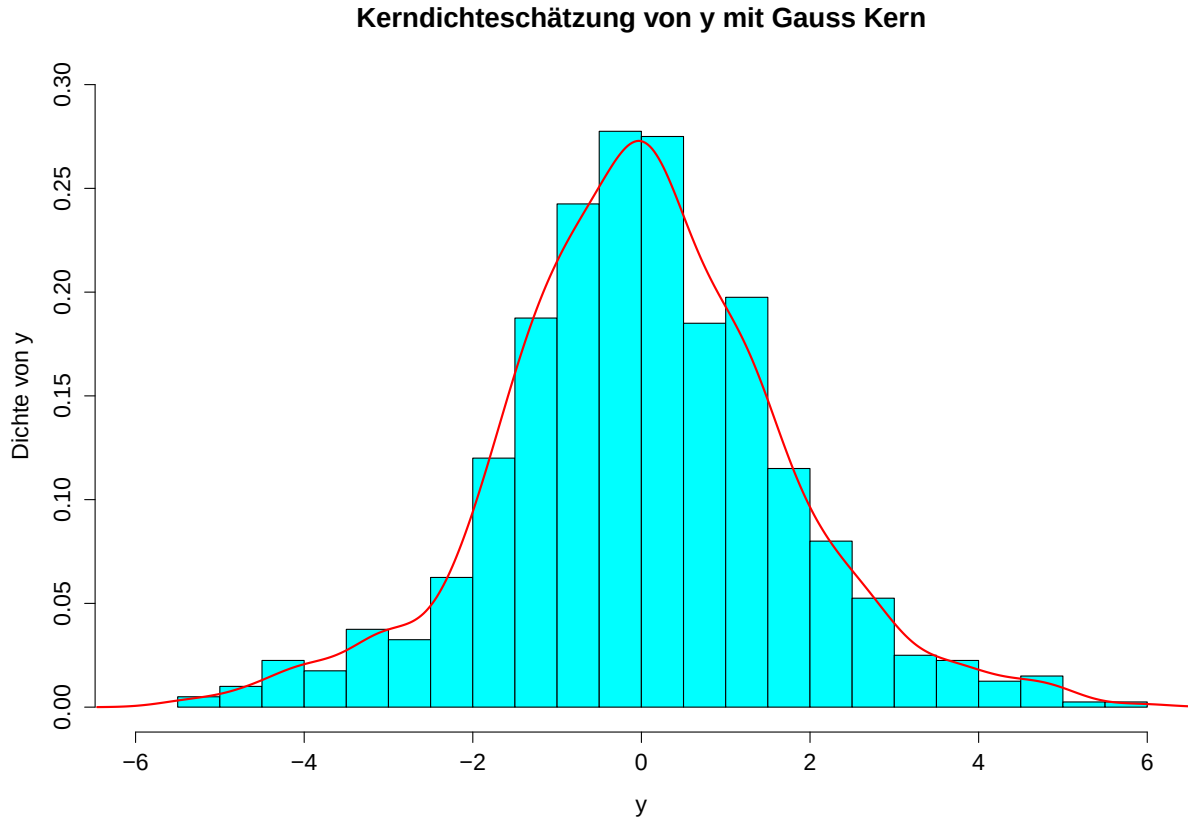

Abbildung 4: Explorative Analyse von y

Nach der Intuition sollte y wieder normalverteilt sein, da in der Simulation diese Werte nur transformiert worden sind. Das wahre Histogramm aus Abbildung 4, normiert zur Fläche 1, mit Gausskerndichteschätzung zeigt, dass es sich näherungsweise um eine Normalverteilung handelt. Zusätzlich dazu liefert der Vorzeichen-Test, dass Median und Mittelwert sich deutlich nicht signifikant voneinander unterscheiden. Somit lässt sich mit Irrtumswahrscheinlichkeit des beta-Fehlers davon ausgehen, dass die Verteilung symmetrisch ist.
Für die Daten aus der vorletzten Grafik 3 auf Seite 13 wird nun ein GAM geschätzt. Bei der Wahl der größtmöglichen Basis wird die gerundete Heuristik $n^{2/9} * 10$ in Abhängigkeit des Stichprobenumfangs (n=800) verwendet. (Wood, 2006, vgl. S. 170) Die Basisfunktion ist ein Thin Plate Regression Spline, welche in Kapitel 2.3 dargestellt sind. Zur Schätzung des Glättungsparameters wird, wie in Kapitel 2.4 beschrieben, GCV verwendet. Es wird ein Quasi-Likelihood-Ansatz zur Schätzung der Dispersion verwendet. In der Simulation gibt es keine Einschränkungen an die Zielvariable und es wird die Standardvoreinstellung (natürlicher Link der Normalverteilung mit konstanter Varianz) verwendet. Der Rauheitsgrad der Funktion wurde standardmäßig mit der 2. Ableitung bestimmt. Die Schätzung der Konfidenzbänder erfolgt analog zu Kapitel 2.5.

In der Abbildung 5 sieht man, dass die geschätzte Funktion im mittleren Bereich den realen Zusammenhang ziemlich genau trifft und nur im linken und rechten Randbereich $[0, 200]$; $[600, 800]$ weicht die Schätzung etwas ab. Der Glättungsterm ist zudem signifikant (p-value $< 2 * 10^{-16}$; es wird standardmäßig alpha=0.05 verwendet) und es werden 64 Prozent der Nulldevianz durch das Modell erklärt. Dies ist ein relativ guter Fit zu beurteilen. Bezüglich der Modelldiagnostik, siehe Abbildung 6, lässt sich folgendes aussagen: Das linke, obere Diagramm ist ein Normal QQ Plot. Dort werden die theoretischen Quantile der entsprechenden Normalverteilung mit den Quantilen der Residuen verglichen. Es ist nicht überraschend, dass diese ziemlich genau übereinstimmen, denn schließlich wurden die Residuen aus einer Standardnormalverteilung gezogen. In der Darstellung rechts daneben fällt eine Häufung des linearen Prädiktors um 0 herum auf. Ansonsten wird bestätigt,

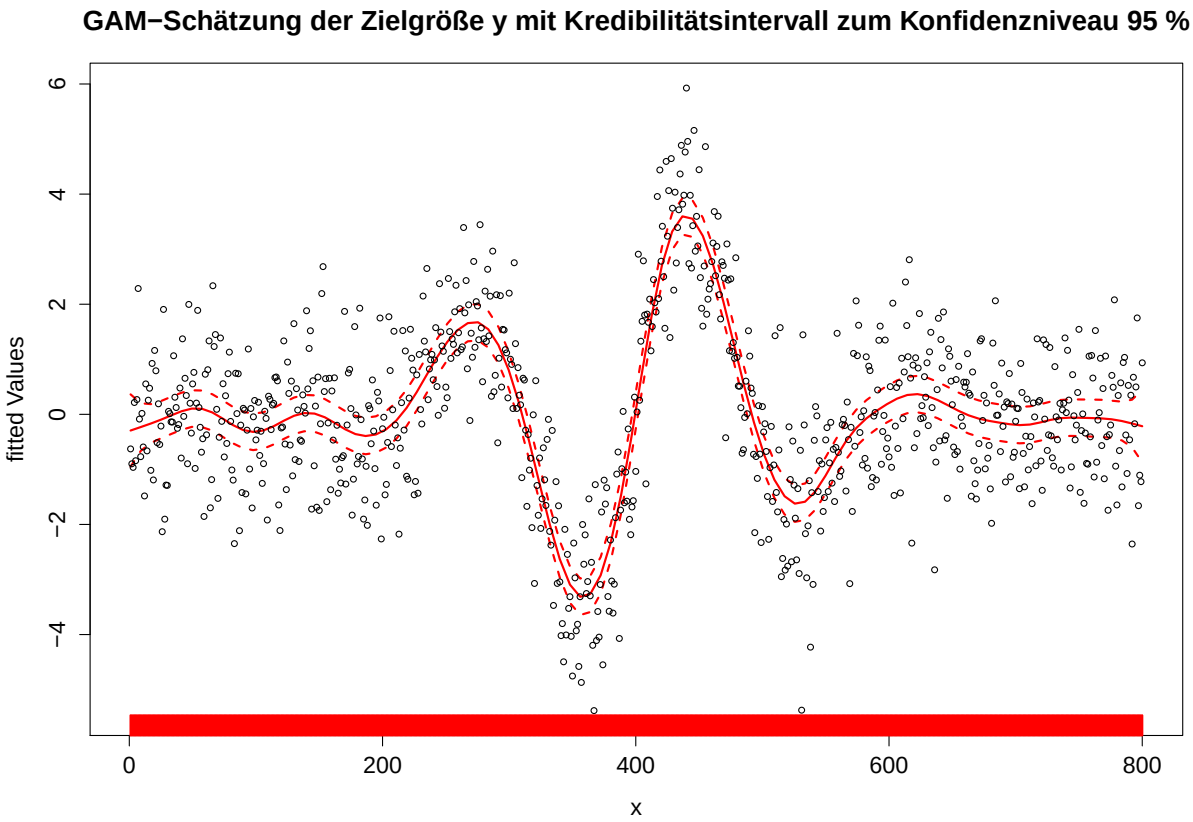

Abbildung 5: Beispiel GAM

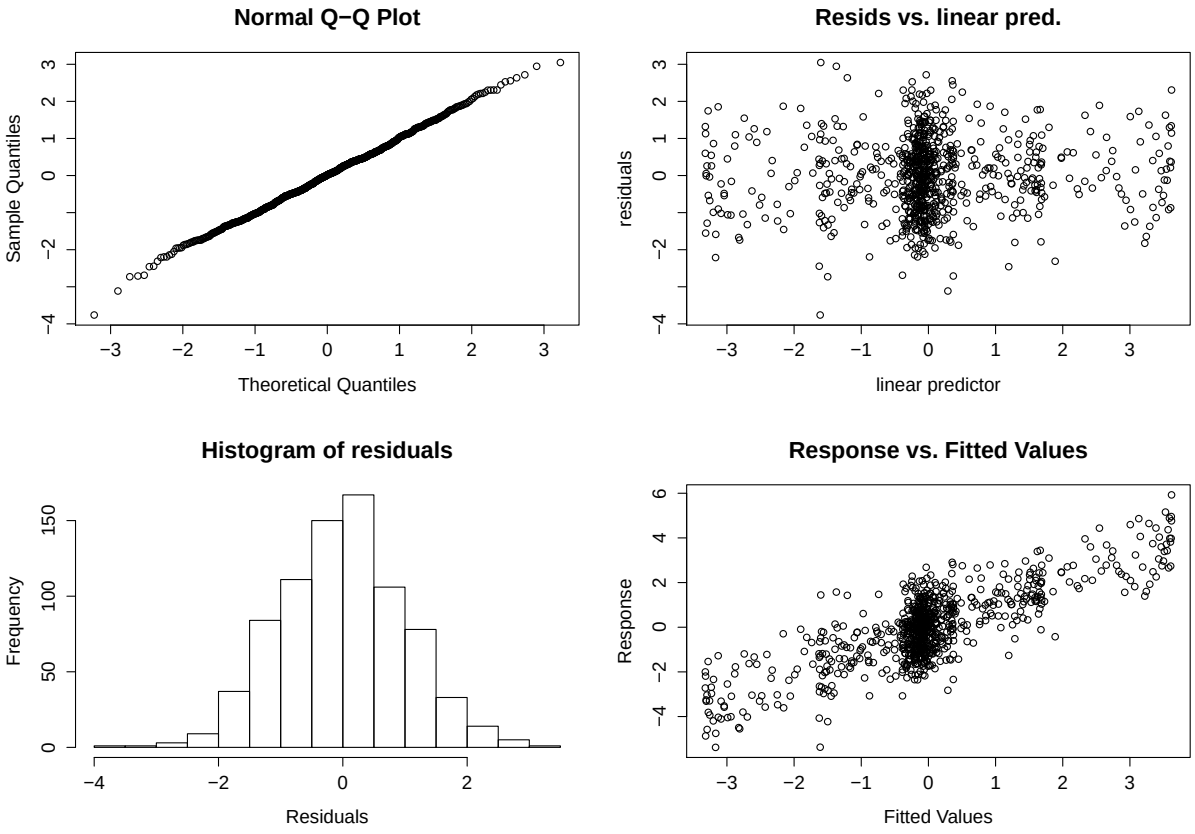

Abbildung 6: GAM Diagnostik

dass die Residuen eine konstante Varianz haben und kein linearer Zusammenhang erkennbar ist. Im Histogramm der Residuen aus der Abbildung 6 sieht man ebenfalls, dass die Dichte der Residuen normalverteilt ist. Aus der Graphik im rechten unteren Bereich lässt sich ableiten, dass die gefitteten Werte proportional zum Response ist. Genauso sollte dies im Idealfall sein. Innerhalb dieser Simulation ist somit gezeigt, falls die wahren Zusammenhänge mit den Modellannahmen übereinstimmt, dass dies erstens mit Regressionsdiagnostik erkannt werden kann und zweitens bei passenden Modellannahmen das Modell den wahren Zusammenhang gut erkennt.

Als Gegenbeispiel wird nun ein parametrisches Modell in Form eines GLM für das identische Modell gerechnet. Der wesentliche Unterschied zum GAM ist die Spezifikation des linearen Prädiktors: In diesem Fall wird ein kubisches Polynom angenommen: $y \sim x + x^2 + x^3$. Die restlichen Annahmen bleiben gleich. Es ergibt sich folgendes Bild:

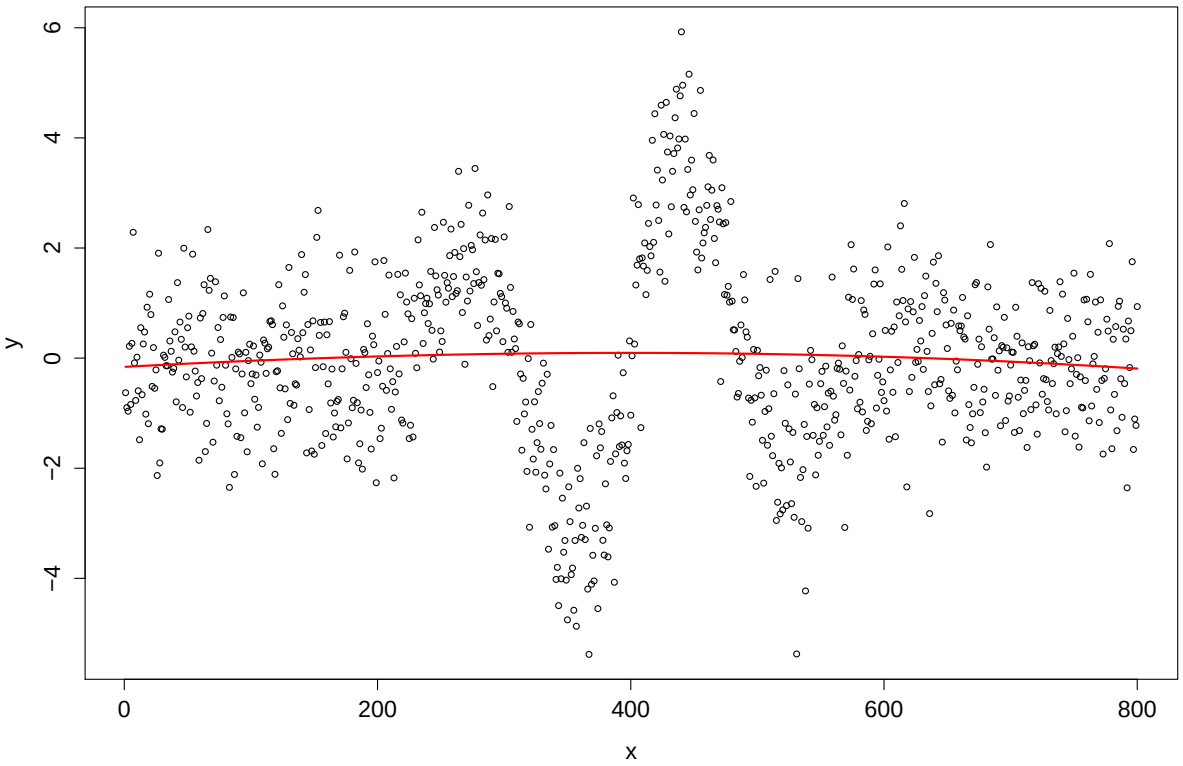

Abbildung 7: Gegenbeispiel GLM

Es ist eine leicht konkav gekrümmte Funktion, d.h. eine Art Parabel, zu sehen. Man erkennt deutlich, dass das parametrische Modell nur den langfristigen Trend um den Wert 0 erkennt. Die stückweisen Ausschläge kann dieses Modell nicht erfassen. Dieses Modell ist nicht signifikant von dem reinen Intercept-Modell ohne Kovariablen verschieden (p-Wert = 0.6052). Des Weiteren ist jeder Einflussterm ebenfalls nicht signifikant. Die Abweichungen der gefitteten Werte zu den Responsewerten ist ziemlich groß, denn das Modell erklärt nur 0,02 Prozent der Nulldevianz. Somit ist diese Alternative deutlich schlechter als das GAM.

4 Zusammenfassung und Fazit

In Kapitel 2 ist eine Einleitung zur Schätzung von GAMs erläutert. Das GAM lässt sich bei gegebenen Basisfunktion wieder in eine lineare Form bringen. Diese kann dann mit Hilfe von P-IRLS Verfahren berechnet werden. Im praktischen Teil wurde der Vorteil von GAMs bei nichtlinearen Zusammehängen deutlich. Bei komplexen Funktionen reicht die bloße Transformation des linearen Prädiktors im GLM nicht mehr aus, um einen guten Fit zu erreichen.

GAMs in der Regel flexibler, da weniger Annahmen benötigt werden. Somit ist eine bessere Anpassung an die Daten möglich. Nichtparametrische Modelle sind außerdem in gängiger Statistiksoftware automatisiert implementiert, so dass diese für eine breite Anzahl von Nutzern verfügbar ist. Durch immer leichter verfügbare Computerressourcen und innovative Schätzalgorithmen können GAMs in der Praxis effizient für immer größere Datensätze eingesetzt werden. Natürlich hat die Komplexität das GAM auch seine Nachteile: Bei nichtparametrischen Verfahren ist es meist wenig sinnvoll jeden einzelnen Parameter eines Splines innerhalb der geschätzten Funktion gesondert zu interpretieren; was bei einem parametrischen Modell möglich ist. Vielmehr betrachtet man in einer graphische Analyse die Funktion oder die Teile der Funktionen (Fahrmeir et al., 2009, vgl. S.295). Je allgemeiner ein Verfahren ist, desto komplizierter wird es dieses zu modellieren. Zudem können die Schätzer nicht mehr analytisch bestimmt werden, sondern müssen oft mit Approximationsverfahren der Numerik bestimmt werden. Dies benötigt zwangsläufig mehr computationale Resourcen als beispielsweise das lineare Modell. Des Weiteren ist die Modelldiagnostik schwieriger interpretierbar wie beim linearen Modell oder beim GLM. Ein weiterer wichtiger Punkt ist die Konvergenzgeschwindigkeit, d.h. wie schnell die Parameterschätzer in Abhängigkeit des Stichprobenumfangs gegen den wahren Wert konvergieren: Bei parametrischen Modellen ist dies für den typischen Fall $n^{-\frac{1}{2}}$. Im nichtparametrischen Fall, wenn man als Maß die Standarabweichung verwendet, liegt die Konvergenzgeschwindigkeit für gewöhnlich bei $n^{-r};\ 0 < r < \frac{1}{2}$. Je höher die Dimension der Kovariablen und je höher die zu schätzende Ableitung der zu schätzenden des Glättungsterms ist, desto langsamer ist die Konvergenzgeschwindigkeit. Je mehr Ableitungen einer Funktion existierten, desto tendenziell höher ist die Konvergenzgeschwindigkeit (Haerdle, 1994, s. S. 116,119)

Des Weiteren gibt es neben der hier vorgestellten Methodik eine Vielzahl von weiteren Möglichkeiten (z. B. Lokalisierungstechniken ohne globale Modellspezifikation), die unbekannte Funktion nichtparametrisch zu schätzen. Welche von diesen wirklich besser ist, kann allgemein kaum festgestellt werden, da die wahren Zusammenhänge in der Realität nicht bekannt sind.

5 Anhang

5.1 Latex-Code

5.2 R-Code

Literatur

Aydin, D. (2007). A comparison of the nonparametric regression models using smoothing spline and kernel regression, *World Academy of Science, Engineering and Technology* **36**.

Bracewell, R. (2000). *Fourier Transforms and its Applications*, 3. edn, The Mc Graw Hill Companies, printed in Singapore.

Demmel, J. (1997). *Applied Numerical Linear Algebra*, Society for Industrial and Applied Mathematics.

Elstrodt, J. (2009). *Mass und Integrationstheorie*, Springer Verlag Berlin, Heidelberg.

Fahrmeir, L., Kneib, T. and Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen*, 3. edn, Springer-Verlag.

Golub, G. and Loan, C. V. (1996). *Matrix Computations*, 3. edn, The Johns Hopkins University Press, Balitmore and London.

Haerdle, W. (1994). *Applied Nonparametric Regression*, Humboldt Universitaet zu Berlin.

Harville, D. (1997). *Matrixalgebra from a Statisticians Perspective*, Springer Verlag New York Inc.

Hothorn, T., Boeck, A. and Kobl, M. (2011). Wahrscheinlichkeitstheorie und inferenz 2, Internet: $http://www.statistik.lmu.de/institut/ag/biostat/vorlesungen/SS11/Statistik_IV/materialien.html$. Skript zur Vorlesung.

Kauermann, G. (2006). Nonparametric models and their estimation, *Allgemeines Statistisches Archiv* **90**(ISSN 0002-6018): 16.

O'Sullivan, F., Yandell, B. and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models, *Journal of the American Statistical Assoziation* **81**(393): 96–103.

Sprent, P. and Smeeton, N. (2001). *Applied nonparametric statistical methods*, Chapman u. Hall, CRC texts in statistical science series, United States of America.

Wasserman, L. (2006). *All of Nonparametric Statistics*, Springer Verlag, United States of America.

Wiencierz, A., Greven, S. and Kuechenhoff, H. (2011). Restricted likelihood ratio testing in linear mixed models with general error covariance structure, *Electronic Journal of Statistics* **5**: 1718–1734.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*, Chapman u. Hall.

Wood, S. (2011). *R-Package mgcv*.