A Unifying Bayesian Perspective on Structured Additive Regression and Mixed Models

Thomas Kneib

Institut für Statistik
Ludwig-Maximilians-Universität München

SFB 386

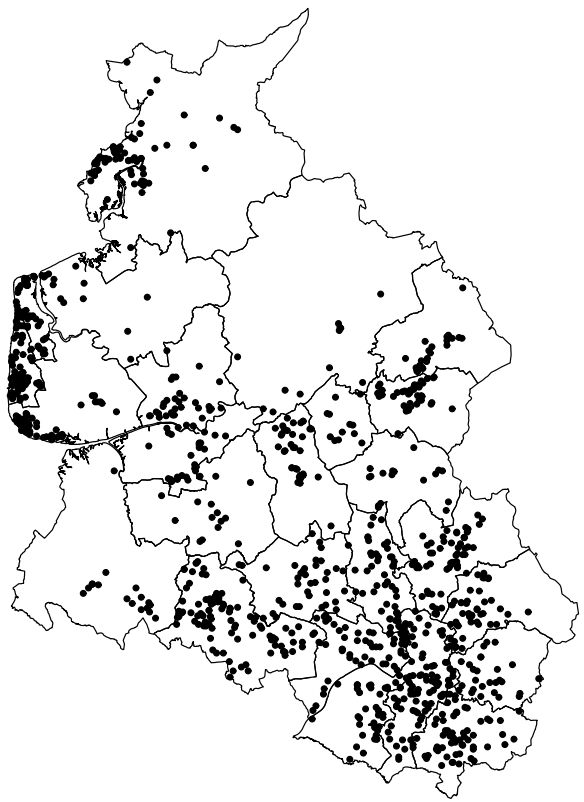
7.11.2006

Leukämie in Nordwest-England

- Überlebenszeiten Erwachsener nach der Diagnose akuter myeloischer Leukämie.
- 1.043 Fälle, die zwischen 1982 und 1998 diagnostiziert wurden.
- Circa 16 % Rechtszensierung.
- Stetige und kategoriale Kovariablen:

  age Alter zum Zeitpunkt der Diagnose,
  wbc Anzahl weißer Blutkörperchen,
  sex Geschlecht,
  tpi Townsend Index.

- Geographische Information in verschiedenen Auflösungen.



- Klassisches Cox-Modell:

$$\lambda(t;u) = \lambda_0(t)\exp(u'\gamma).$$

- Die Baseline-Hazardrate $\lambda_0(t)$ ist eine beliebige, unspezifizierte Funktion (Störparameter).
- Schätzung von γ basierend auf der partiellen Likelihood.
- Fragen / Probleme:
  - Simultane Schätzung der Baseline-Hazardrate und der Kovariableneffekte.
  - Flexible Modellierung von Kovariableneffekten (nichtlineare Effekte, Interaktionen).
  - Berücksichtigung der räumlichen Korrelation.
  - Nicht-proportionale Hazardraten / zeitvariierende Effekte.

$\Rightarrow$ Strukturiert additive Regressionsmodelle.

- Ersetze den linearen Prädiktor durch den flexiblen semiparametrischen Prädiktor

$$\lambda(t;\cdot) = \lambda_0(t)\exp[f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s_i) + \gamma_1 sex]$$

und ergänze die Baseline-Hazardrate

$$\lambda(t;\cdot) = \exp[g_0(t) + f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s_i) + \gamma_1 sex].$$

- Dabei bezeichnen
  - $g_0(t) = \log(\lambda_0(t))$ die logarithmierte Baseline-Hazardrate,
  - f_1, f_2, f_3 nonparametrische Funktionen des Alters, der Anzahl weißer Blutkörperchen und des Townsend Index, und
  - f_{spat} eine räumliche Funktion.

Strukturiert additive Regression

- Regressionsmodelle mit strukturiert additivem Prädiktor der Form

$$\eta_{it} = f_1(z_{it1}) + \ldots + f_p(z_{itp}) + u'_{it}\gamma.$$

- Dabei sind
  - $f_1, \ldots, f_p$ Funktionen verschiedenen Typs,
  - $z_1, \ldots, z_p$ generische Kovariablen.

- Beispiele:

$f(v) = f(x)$	$v = x$	nonparametrische Funktion einer metrischen Kovariablen,
$f(v) = f_{spat}(s)$	$v = s$	räumlicher Effekt,
$f(v) = g(x)u$	$v = (x, u)$	Effekt mit variierenden Koeffizienten,
$f(v) = f(x_1, x_2)$	$v = (x_1, x_2)$	Interaktionsoberfläche.

- Alle Effekte können als Produkt einer Designmatrix X_j und eines Vektors von Regressionsparametern β_j beschrieben werden:

$$f_j = X_j \beta_j.$$

- Bayesianischer Ansatz: Priori-Verteilung für β_j.

- Allgemeine Form:

$$p(\beta_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right)$$

  wobei K_j eine Strafmatrix ist und τ_j^2 ein Glättungsparameter.

- Verbindung zu penalisierter ML-Schätzung:

$$Pen(\beta_j) = \log\left[p(\beta_j|\tau_j^2)\right] = -\frac{1}{2}\lambda_j \beta_j' K_j \beta_j, \qquad \lambda_j = 1/\tau_j^2.$$

Mixed Model Repräsentation von P-Splines

- Das allgemeine lineare gemischte Modell besitzt die Form

$$y = U\gamma + Zb + \varepsilon,$$

mit festen Effekten γ und den Verteilungsannahmen

$$\begin{bmatrix} \varepsilon \\ b \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I & 0 \\ 0 & Q \end{bmatrix} \right).$$

- ML-Schätzung basierend auf der gemeinsamen Likelihood

$$p(y, b) \propto p(y|b)p(b) \rightarrow \max_{\gamma, b}$$

führt zu dem penalisierten KQ-Kriterium

$$(y - U\gamma - Zb)'(y - U\gamma - Zb) + \sigma^2 b' Q^{-1} b \rightarrow \min_{\gamma, b}.$$

- Einfachstes nonparametrisches Regressionsmodell:

$$y_i = f(x_i) + \varepsilon_i \qquad \varepsilon_i \text{ i.i.d. } N(0, \sigma^2).$$

- Basisfunktionenansatz:

$$y_i = \sum_{j=1}^{d} \beta_j B_j(x_i) + \varepsilon_i.$$

- In Matrixschreibweise:

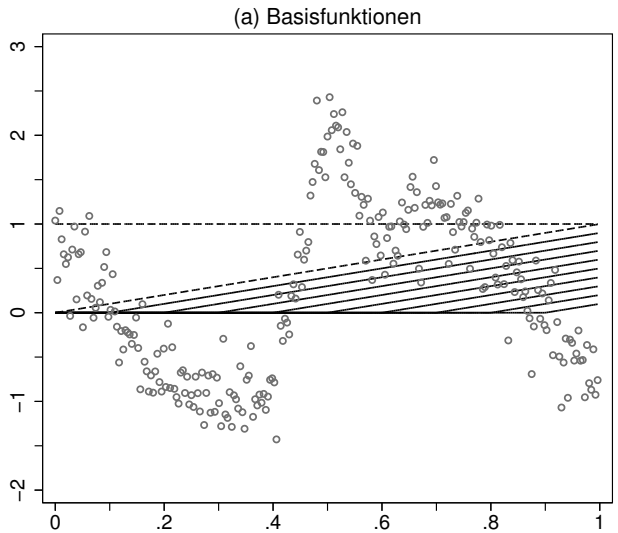
$$y = X\beta + \varepsilon$$

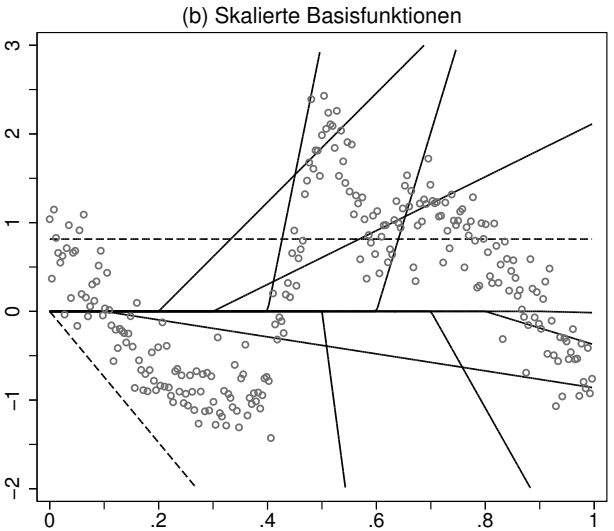
mit

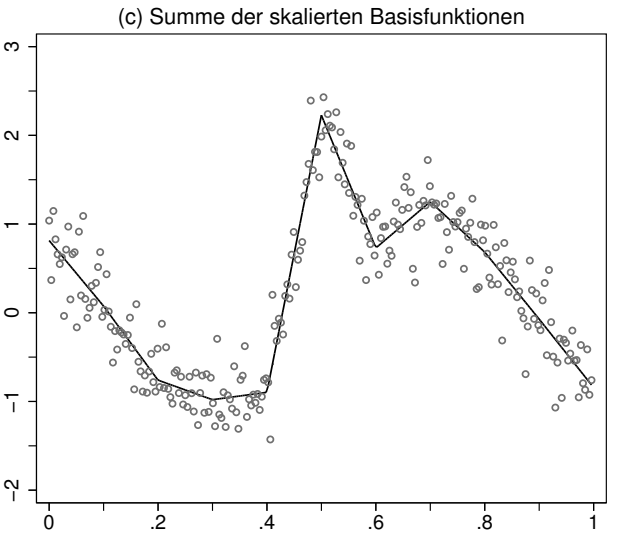
$$X[i, j] = B_j(x_i) \qquad \text{und} \qquad \beta = (\beta_1, \ldots, \beta_d)'.$$

- Truncated Power Series Basis:

$$f(x) = \beta_1 + \beta_2 x + \ldots + \beta_{l+1} x^l + \beta_{l+2}(x - \kappa_2)_+^l + \ldots + \beta_d(x - \kappa_{m-1})_+^l$$







- Üblicher Strafterm zur TP-Basis:

$$\lambda \sum_{j=l+2}^{d} \beta_j^2.$$

- Alternative Darstellung des Modells:

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

mit

$$U = \begin{pmatrix} 1 & x_1 & \dots & x_1^l \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^l \end{pmatrix}, \quad Z = \begin{pmatrix} (x_1 - \kappa_2)_+^l & \dots & (x_1 - \kappa_{m-1})_+^l \\ \vdots & & \vdots \\ (x_n - \kappa_2)_+^l & \dots & (x_n - \kappa_{m-1})_+^l \end{pmatrix}$$

und

$$\gamma = (\beta_1, \dots, \beta_{l+1})', \quad b = (\beta_{l+2}, \dots, \beta_d)'.$$

- Dann lässt sich auch der Strafterm umschreiben zu

$$\lambda \sum_{j=l+2}^{d} \beta_j^2 = \lambda b'b = \sigma^2 b' Q^{-1} b$$

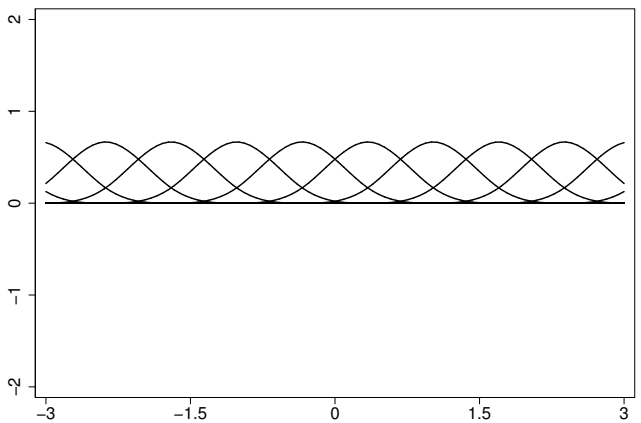
mit $Q = \tau^2 I$ und $\lambda = \sigma^2/\tau^2$.

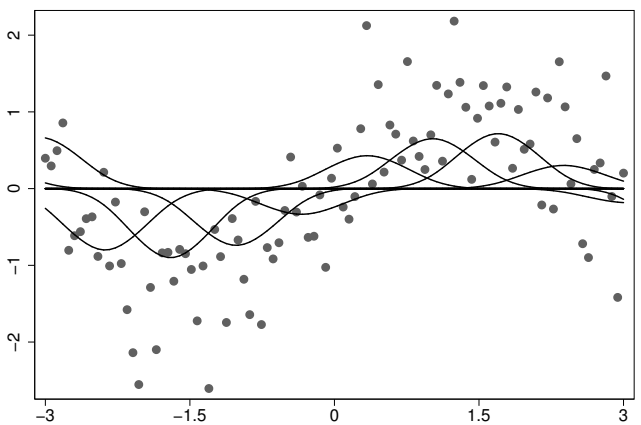
- Damit ergibt sich die folgende Interpretation:
  - Die Koeffizienten der ersten $l+1$ Basisfunktionen entsprechen festen Effekten in einem gemischten Modell.
  - Die Koeffizienten der trunkierten Polynome entsprechen zufälligen Effekten mit Varianz τ^2.

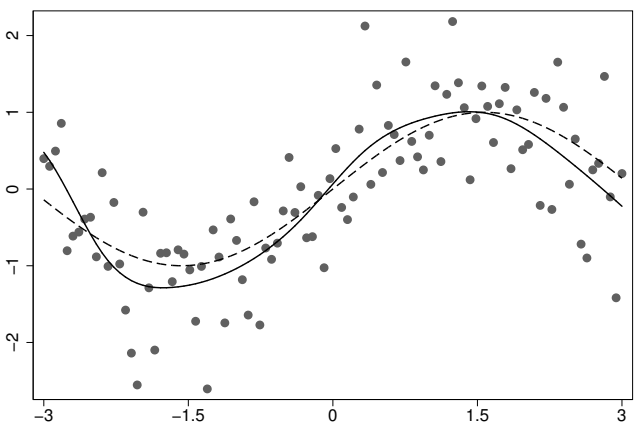
- Aus frequentistischer Perspektive ergibt sich trotz der deterministischen Formulierung des nonparametrischen Regressionsmodells eine formale Äquivalenz zu einem (stochastischen) gemischten Modell.

- Bayesianisch betrachtet ergibt sich kein Widerspruch. Die Reparametrisierung ergibt lediglich eine andere Form der Priori-Vereilung.

- Wie lässt sich dieser Ansatz auf allgemeine Penalisierungsansätze übertragen?
- Im Folgenden: Exemplarische Überlegungen für B(asic)-Splines.
- Numerisch stabilere Basis zur Darstellung von Polynom-Splines.
- Schätzung kubischer B-Splines:







- Geeignete Strafterme zur Regularisierung:

$$Pen(\beta|\tau^2) = \frac{1}{\tau^2}\sum_{j=2}^{d}(\beta_j - \beta_{j-1})^2 \qquad \text{erste Differenzen}$$

$$Pen(\beta|\tau^2) = \frac{1}{\tau^2}\sum_{j=3}^{d}(\beta_j - 2\beta_{j-1} + \beta_{j-2})^2 \qquad \text{zweite Differenzen}$$

- τ^2 wirkt als Glättungsparameter.

- Darstellbar als quadratische Formen:

$$Pen(\beta|\tau^2) = \frac{1}{\tau^2}\beta' K \beta$$

mit Strafmatrix

$$K = D_1' D_1 \qquad \text{bzw.} \qquad K = D_2' D_2.$$

- Die Strafmatrizen haben nicht vollen Rang:

$$\begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 1 & \end{pmatrix} \qquad \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}$$

$$\text{rang}(K) = d - 1 \qquad \text{rang}(K) = d - 2$$

$\Rightarrow$ Es existiert ein nichttrivialer Nullraum der durch K definierten linearen Abbildung ("Was wird nicht durch K bestraft?").

- Basis dieses Nullraums für k-te Differenzen:

$$\begin{pmatrix} 1 & 1 & \ldots & 1^{k-1} \\ 1 & 2 & \ldots & 2^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & d & \ldots & d^{k-1} \end{pmatrix}$$

- Penalisierte KQ-Schätzung:

$$(y - X\beta)'(y - X\beta) + \frac{\sigma^2}{\tau^2}\beta' K \beta \rightarrow \min_{\beta}$$

- Vergleich mit gemischtem Modell:

$$y = U\gamma + Zb + \varepsilon, \qquad b \sim N(0, Q)$$

$$(y - U\gamma - Zb)'(y - U\gamma - Zb) + \sigma^2 b' Q^{-1} b \rightarrow \min_{\gamma, b}$$

$\Rightarrow$ β entspricht einem zufälligen Effekt mit Erwartungswert 0 und Präzisionsmatrix $\frac{1}{\tau^2}K$.

- Problem: K hat nicht vollen Rang.
- Die Verteilung von β ist teilweise uneigentlich (nicht normierbar).

- Allgemein gilt für einen teilweise uneigentlich multivariat normalverteilten Zufallsvektor β:

  β lässt sich zerlegen in

$$\beta = \tilde{U}\gamma + \tilde{Z}b,$$

  so dass

$$\underbrace{p(\beta)}_{\text{teilweise uneigentlich}} = \underbrace{p(\gamma)}_{\text{uneigentlich}} \cdot \underbrace{p(b)}_{\text{eigentlich}} \propto p(b)$$

  wobei

$$\begin{aligned} \dim(\gamma) &= \dim(\beta) - \text{rang}(K), \\ \dim(b) &= \text{rang}(K). \end{aligned}$$

- γ = deterministischer Anteil von β,

  b = stochastischer Anteil von β.

- Anforderungen an die Zerlegung:
  - $(\tilde{U} : \tilde{Z})$ hat vollen Rang, so dass die Abbildung zwischen β und $(\gamma', b')'$ eineindeutig ist.
  - $\tilde{U}$ und $\tilde{Z}$ sind orthogonal, d.h. $\tilde{U}'\tilde{Z} = 0$.
  - $\tilde{U}'K\tilde{U} = 0$, so dass γ nicht von K penalisiert wird.
  - $\tilde{Z}'K\tilde{Z} = I$, so dass die Verteilung von b eigentlich ist.

- Dann folgt

$$p(\gamma) \propto const \qquad \text{und} \qquad b \sim N(0, \tau^2 I).$$

- γ entspricht einem Vektor fester Effekte.

- b entspricht einem Vektor von i.i.d. zufälligen Effekten.

- Konstruktion von $\tilde{U}$ und $\tilde{Z}$:
  - $\tilde{U}$ enthält eine Basis des Nullraums von K.
  - $\tilde{Z}$ enthält eine Basis des orthogonalen Komplements dieses Nullraums.
  - Bestimmung über die Eigenwertzerlegung von K:

$$K = \Gamma \Omega_+ \Gamma' = LL'$$

Ω_+ = Matrix der (positiven) Eigenwerte,
Γ = Matrix der zugehörigen Eigenvektoren,
L = Eigenvektoren skaliert mit den Eigenwerten.

$\Rightarrow$ Setze $\tilde{Z} = L(L'L)^{-1}$.

- Es existieren alternative Definitionen von $\tilde{Z}$, die die Berechnung der Eigenwerte vermeiden.

- Beispiel: $\tilde{Z} = D'(DD')^{-1}$.

- Einsetzen der Zerlegung in die Modellgleichung ergibt

$$\begin{aligned} y &= X\beta + \varepsilon \\ &= X(\tilde{U}\gamma + \tilde{Z}b) + \varepsilon \\ &= U\gamma + Zb + \varepsilon. \end{aligned}$$

$\Rightarrow$ Darstellung eines P-Splines als gemischtes Modell.

- Vorteil: Zusätzlich zu den Regressionsparametern können auch die Varianz- bzw. Glättungsparameter geschätzt werden!

Übertragung auf weitere Modellterme

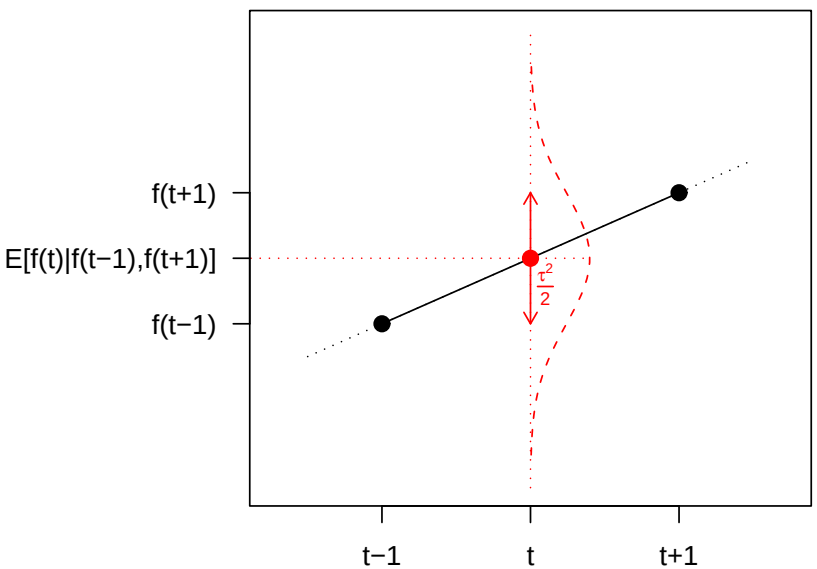
- Das vorgestellte Prinzip lässt sich allgemein auf Regressionsansätze übertragen, die auf quadratischen Straftermen beruhen.

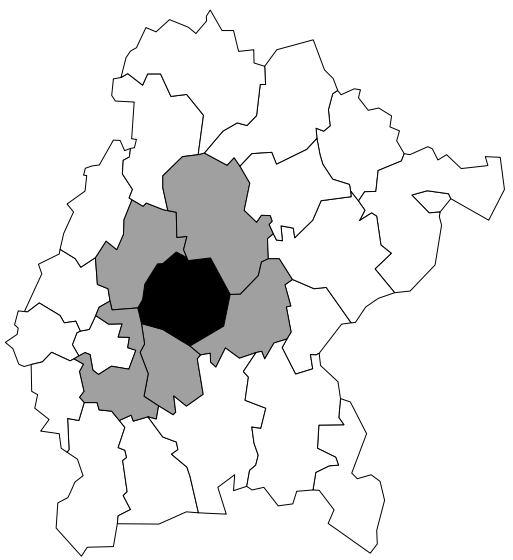
- Beispiele sind:
  - Markov-Zufallsfelder,
  - Interaktionsoberflächen basierend auf 2d P-Splines,
  - Flexible Saisonkomponenten,
  - Variierende Koeffizienten,
  - Glättungssplines,
  - (Stationäre Gauß-Felder),
  - (zufällige Effekte).

- Markov-Zufallsfelder: Modellierung räumlicher Effekte.
  - β_s gibt den räumlichen Effekt an einer Lokalisation $s \in \{1, \ldots, S\}$ an.
  - Einfachstes Markov-Zufallsfeld:

$$\beta_s | \beta_{s'}, s' \neq s, \tau^2 \sim N\left(\frac{1}{N_s}\sum_{s' \in \partial_s} \beta_{s'}, \frac{\tau^2}{N_s}\right),$$

δ_s = Menge der Nachbarn von s, $N_s = |\delta_s|$.





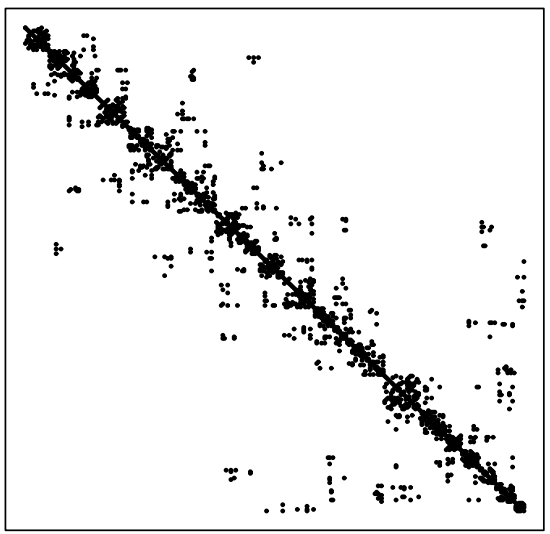
- Auch hier lässt sich ein penalisiertes KQ-Kriterium aufstellen:

$$(y - X\beta)'(y - X\beta) + \frac{\sigma^2}{\tau^2}\beta' K \beta \rightarrow \min_{\beta}.$$

- X ist eine Inzidenzmatrix, die die räumlichen Effekte mit den entsprechenden Beobachtungen verknüpft.
- K ist eine Nachbarschaftsmatrix:

$$\begin{aligned} k_{ss'} &= \begin{cases} -1 & \text{falls } s \text{ und } s' \text{ benachbart,} \\ 0 & \text{sonst,} \end{cases} \\ k_{ss} &= N_s. \end{aligned}$$

$$\Rightarrow \operatorname{rank}(K) = \dim(\beta) - 1.$$



- Der Nullraum von K ist eindimensional und besitzt die Basis $(1, \ldots, 1)'$.

Empirische Bayes-Inferenz

- Idee empirischer Bayes-Inferenz:
  - Unterscheide zwischen primär interessierenden Parametern und Hyperparametern.
  - Schätze die Hyperparameter vorab aus ihrer marginalen Posteriori-Verteilung.
  - Setze die Schätzer in die Posteriori ein und maximiere bezüglich der primär interessierenden Parameter (Posteriori-Modus-Schätzer).
- In strukturiert additiven Regressionsmodellen:
  - Regressionsparameter sind primär interessierende Parameter,
  - Varianzparameter sind Hyperparameter.
- Volle Bayes-Inferenz: Auch die Hyperparameter werden mit Prioris versehen und simultan mitgeschätzt.

- Posteriori in strukturiert additiven Regressionsmodellen:

$$p(\beta_1, \ldots, \beta_p|y) \propto L(y, \beta_1, \ldots, \beta_p) \prod_{j=1}^{p} p(\beta_j|\tau_j^2).$$

- Posteriori-Modus / penalisierte Maximum Likelihood-Schätzung

$$p(\beta_1, \ldots, \beta_p|y) \rightarrow \max_{\beta_1, \ldots, \beta_p} .$$

- Normierungskonstante der Posteriori muss zur Maximierung nicht bekannt sein.

- Logarithmierte Posteriori in Mixed Model Formulierung:

$$l_p(\gamma, b|y) = l(y, \gamma, b) - \frac{1}{2} b' Q^{-1} b.$$

- Newton-Raphson-Algorithmus bzw. Fisher-Scoring bei gegebenen Varianzen τ_j^2 anwendbar.

- Score-Funktion und Fisher-Information:

$$s_p(\gamma, b) = \begin{pmatrix} \dfrac{\partial l(y, \gamma, b)}{\partial \gamma} \\ \dfrac{\partial l(y, \gamma, b)}{\partial b} - Q^{-1} b \end{pmatrix}$$

$$F_p(\gamma, b) = \begin{pmatrix} \dfrac{\partial^2 l(y, \gamma, b)}{\partial \gamma \partial \gamma'} & \dfrac{\partial^2 l(y, \gamma, b)}{\partial \gamma \partial b'} \\ \dfrac{\partial^2 l(y, \gamma, b)}{\partial b \partial \gamma'} & \dfrac{\partial^2 l(y, \gamma, b)}{\partial b \partial b'} - Q^{-1} \end{pmatrix}.$$

- Marginale Likelihood-Schätzung der Varianzparameter:

$$L(Q) = \int L(\gamma, b, Q)p(b)d\gamma db \rightarrow \max_Q .$$

- Hochdimensionales Integral $\Rightarrow$ im Allgemeinen weder analytisch noch numerisch bestimmbar.

- Approximatives Vorgehen:
  - Laplace-Approximation an die Likelihood (quadratische Taylor-Entwicklung um den Modus).
  - Ergibt Normalverteilungs-Likelihood $\Rightarrow$ Integral explizit lösbar.
  - Die resultierende marginale Likelihood entspricht der Restricted Likelihood des approximierenden Normalverteilungs-Modells.

- Für Exponentialfamilien entspricht die Laplace-Approximation der Normalverteilungs-approximation der iterativ gewichteten KQ-Schätzung:

$$\tilde{y}|\gamma, b \stackrel{a}{\sim} N(U\gamma + Zb, W^{-1}).$$

  Dabei bezeichnen $\tilde{y}$ und W die üblichen Arbeitsbeobachtungen und -gewichte in GLMs.

- Für Überlebenszeitmodelle sind zusätzliche Approximationsschritte notwendig.

Software

- BayesX - Software für empirische und volle Bayes-Inferenz in strukturiert additiven Regressionsmodellen.

- Erhältlich unter

`http://www.stat.uni-muenchen.de/~bayesx`

- Modellkomponenten und Prioris:
  - P-Splines (metrische Kovariablen und Zeitskalen),
  - Random Walks (metrische Kovariablen und Zeitskalen),
  - Flexible Saisonkomponenten (Zeitskalen),
  - Markov-Zufallsfelder (räumliche Effekte diskreter Lokationsvariablen),
  - Stationäre Gauß-Felder (Kriging, räumliche Effekte stetiger Lokationsvariablen),
  - 2D P-Splines (Interaktionsoberflächen),
  - Zufällige Effekte (Random Intercept und Random Slope),
  - Variierende Koeffizienten (metrische und räumliche Effektmodifzierer).
- Univariate Zielvariablen:
  - Normalverteilung,
  - Bernoulli- und Binomialverteilung (Logit, Probit und Komplementäres Log-log-Modell),

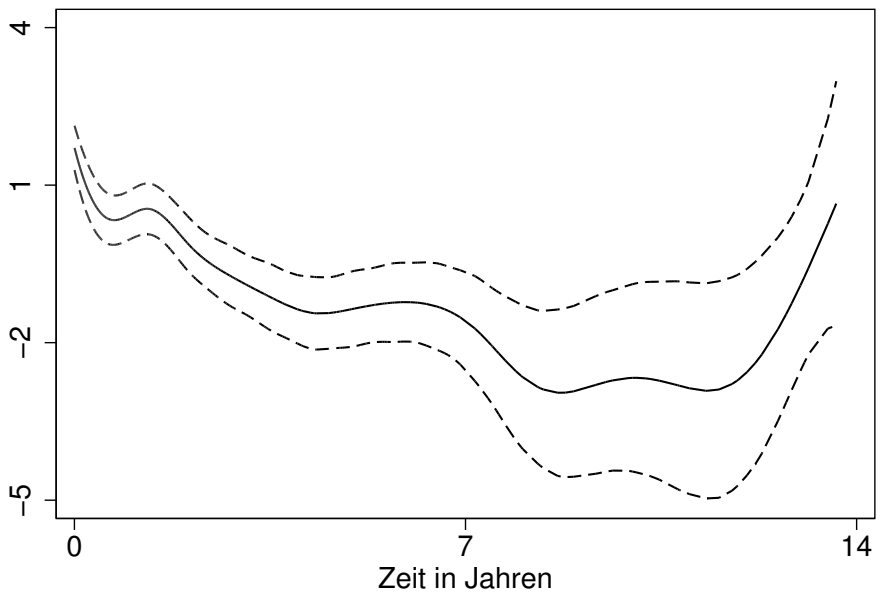
  - Poissonverteilung,
  - Gammaverteilung,
  - Negativ Binomial-Verteilung,
  - Zero-inflated Poisson-Verteilung.

- Kategoriale Zielvariablen:
  - Nominal-skaliert (Multinomiales Logit- und Probit-Modell, Effekte kategorienspezifischer und globaler Kovariablen),
  - Ordinal-skaliert (kumulative und sequentielle Modelle, Logit- und Probit-Link, globale und kategorienspezifische Effekte).

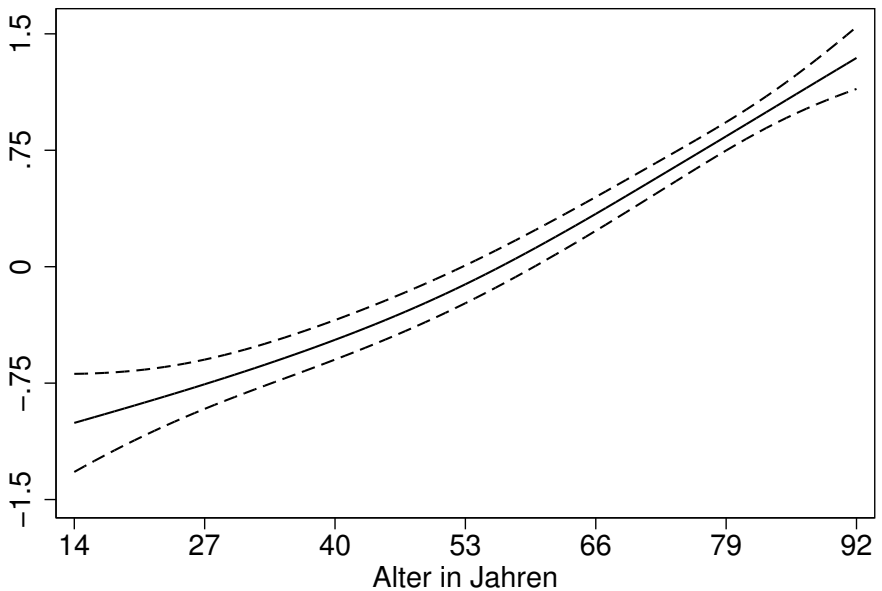
- Modelle der Überlebenszeitanalyse (Hazard-Regression, Schätzung der Baseline-Hazardrate, zeitvariierende Effekte, Rechts-, Links- und Intervallzensierung, Linkstrunkierung).

- Multi-State-Modelle (zeitstetige stochastische Prozesse mit diskretem Zustandsraum, Modellierung der Übergangsintensitäten).

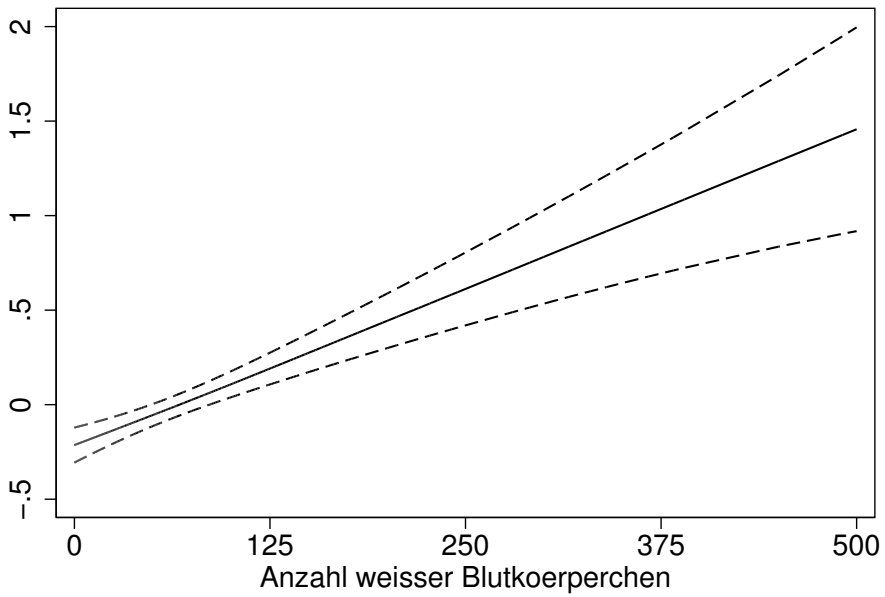
Anwendungsbeispiel: Leukämie in Nordwest-England



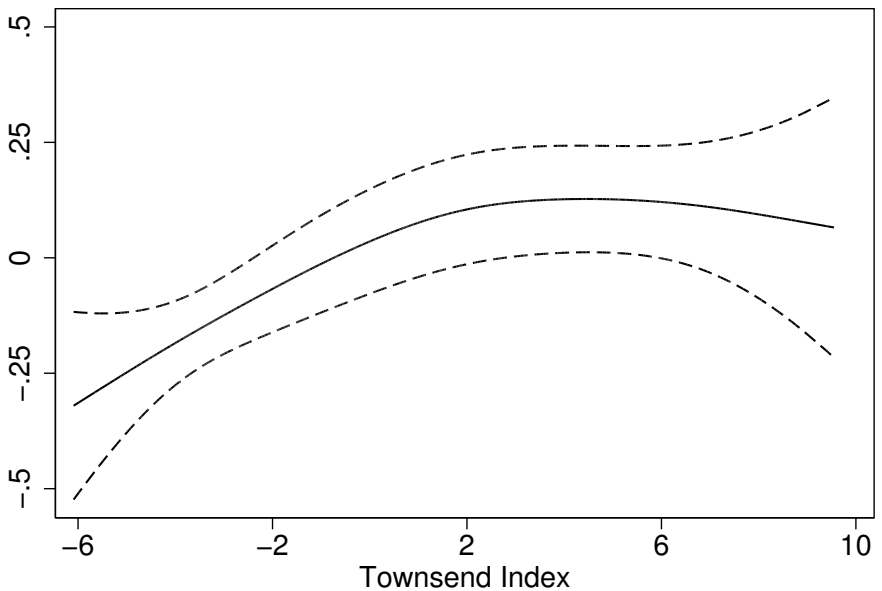
Logarithmierte Baseline Hazardrate.



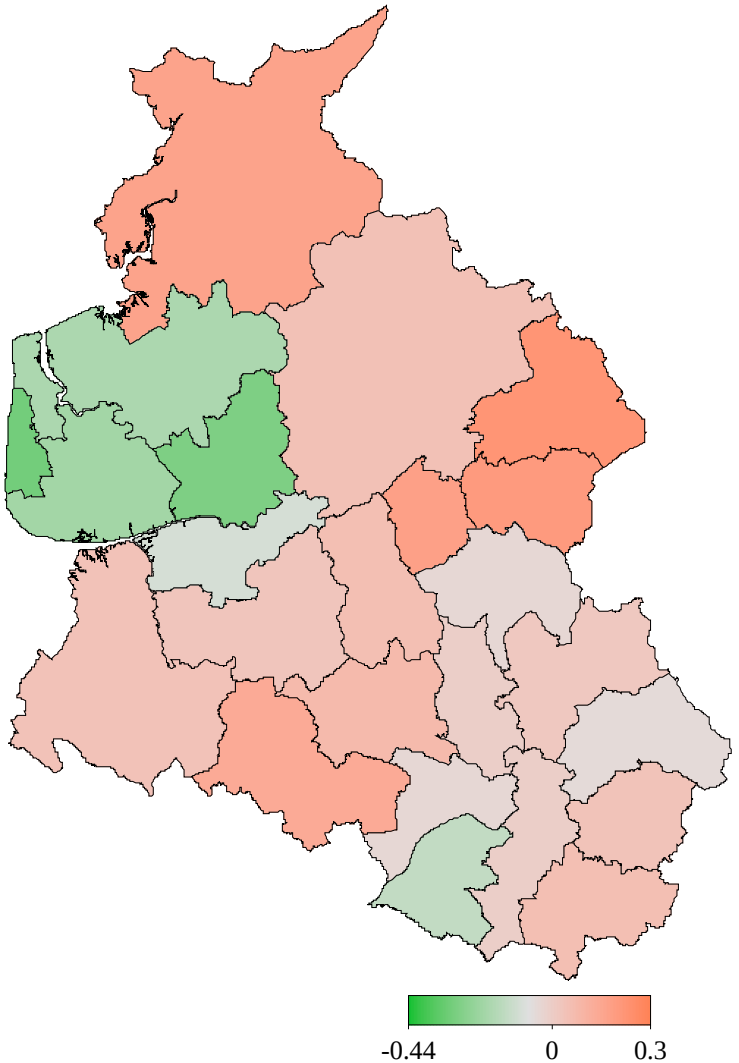
Alterseffekt.



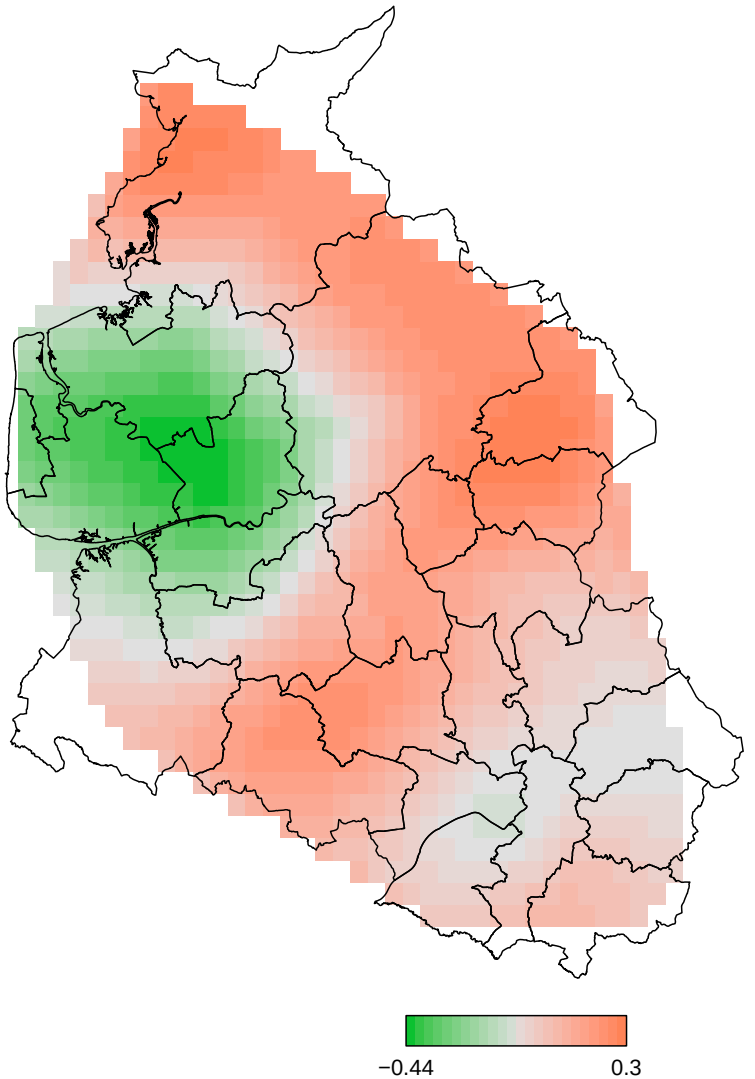
Effekt der Anzahl weißer Blutkörperchen.



Effekt des Townsend Index.



Analyse basierend auf Distrikten



Analyse basierend auf Koordinaten

Diskussion

- Penalisierungsansätze sind formal äquivalent zu Modellen mit zufälligen Effekten.
- Umgekehrt lassen sich beispielsweise stationäre Gauß-Felder als Basisfunktionenansätze interpretieren.
- Aus frequentistischer Perspektive problematisch: Was sind feste Parameter und was sind zufällige Effekte?
- Aus bayesianischer Sicht unproblematisch: Unterschiedliche Sichtweisen und Prioris für äquivalente Modellformulierungen.

- Empirische Bayes-Inferenz kann über Methodik für gemischte Modelle durchgeführt werden.

- In vielen Fällen ernst zu nehmende Alternative zum bayesianischen Standard MCMC.

- Anwendbar in einer Vielzahl von Situationen, auch in einer voll bayesianischen Modellformulierung (Håvard Rue), zur Modellwahl oder in Modellen mit adaptiven Varianzen (Göran Kauermann).