Bayesianische Regularisierung in semiparametrischen Regressionsmodellen

Thomas Kneib

Institut für Statistik und Ökonometrie
Georg-August-Universität Göttingen

Institut für Statistik
Ludwig-Maximilians-Universität München

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

31.10.2008

Übersicht

- Regularisierung in additiven gemischten Modellen
- Inferenz basierend auf gemischten Modellen
- Modellierung der Dynamik der Lebenszufriedenheit
- Regularisierung in hochdimensionalen geoadditiven Regressionsmodellen
- Inferenz basierend auf Markov Chain Monte Carlo Simulationsverfahren
- Münchner Mietspiegel

Regularisierung

- Allgemeines Ziel der Regularisierung von Schätzproblemen: Schätzung „gutartiger“ machen, als sie eigentlich ist.
- Beispiele:
  - Penalisierte Schätzung von Spline-Koeffizienten: Regularisierung durch Norm im Funktionenraum.
  - Räumliche Glättung: Regularisierung durch Annahmen zum Einfluss zwischen benachbarten Regionen.
  - Individuenspezifische Effekte in gemischten Modellen: Regularisierung durch Verteilungsannahme für „zufällige“ Effekte.
- Verschiedene Ziele: Glattheitseigenschaften bzw. sparsame Modellstruktur.
- Verschiedene Vorgehensweisen: Strafterme bzw. Verteilungsannahmen.

- Vereinheitlichende Bayesianische Sichtweise:

  Regularisierung entspricht der Annahme einer informativen Priori-Verteilung.

Gemischte Modelle

- Einfachstes Beispiel: Random Intercept Modell für Longitudinaldaten

$$y_{it} = x_{it}'\beta + b_i + \varepsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i,$$

mit Verteilungsannahme

$$b_i \text{ u.i.v. } N(0, \tau^2).$$

- Klassische Sichtweise: Die Individuen sind eine Zufallsstichprobe aus der Grundgesamtheit.

- Dies spiegelt sich in der Verteilungsannahme für die individuenspezifischen Parameter b_i wider.

- Schätzung basierend auf der gemeinsamen Likelihood

$$p(y, b) = p(y|b)p(b) \rightarrow \max_{\beta, b}$$

- Äquivalente Formulierung über ein penalisiertes Kleinste-Quadrate-Kriterium:

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

- Lösung:

$$\begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda I \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

  mit $\lambda = \sigma^2/\tau^2$.

- Entspricht dem penalisierten KQ-Schätzer mit Ridge-Strafterm für die zufälligen Effekte.

- Bayesianische Perspektive: Die Verteilungsannahme für die zufälligen Effekte entspricht einer Priori-Verteilung.

- Bei nichtinformativer Priori für die festen Effekte $p(\beta) \propto \text{const}$ ergibt sich die Posteriori-Verteilung

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

d.h. die Posteriori ist äquivalent zur gemeinsamen Likelihood und zum penalisierten KQ-Kriterium.

- Allgemeiner gilt:
  - Penalisierte Log-Likelihood

$$l_{\text{pen}}(\beta) = l(\beta) - \lambda \, \text{pen}(\beta).$$

  - Posteriori

$$p(\beta|y) = p(\beta|y)p(\beta).$$

- Man erhält also die Beziehung

Penalisierung = logarithmierte Priori-Verteilung.

Dynamik der Lebenszufriedenheit I

- Beispiel: Analyse der Lebenszufriedenheit basierend auf Daten aus dem Sozio-Oekonomischen Panel.
- Projekt des Statistischen Beratungslabors in Kooperation mit Christoph Wunder, Universität Bamberg.
- Daten aus den Haushaltsstichproben A (Westdeutsche) und C (Ostdeutsche).
- Daten aus dem Zeitraum 1992 – 2006.
- Ursprünglich 88.749 Beobachtungen zu 12.250 Personen (durchschnittlich 7.24 Messwiederholungen).
- Gesamtdatensatz nicht numerisch handhabbar

  $\Rightarrow$ Stichprobe von 1.500 Personen mit 10.562 Beobachtungen.

- Zielgröße: Allgemeine Lebenszufriedenheit gemessen durch den elfstufigen Score zur Frage

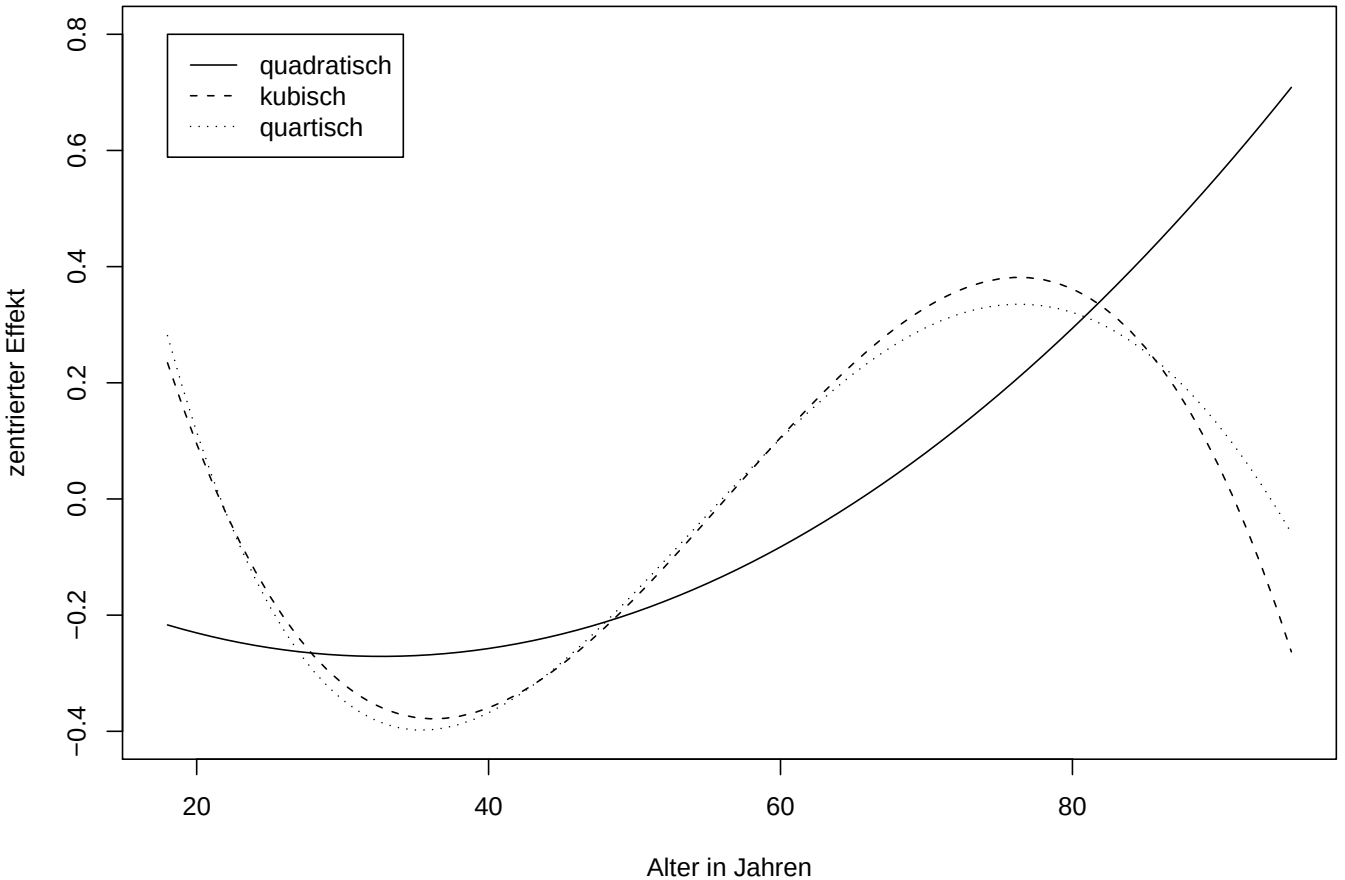
  „Wie zufrieden sind Sie gegenwärtig, alles in allem, mit Ihrem Leben?“

  (0 = „ganz und gar unzufrieden“, 10 = „ganz und gar zufrieden“)

- Kovariablen zu Geschlecht, Bildung, Haushaltseinkommen, Erwerbsstatus, Gesundheit, etc.

- Im Folgenden Fokus auf die Dynamik der Lebenszufriedenheit (Einfluss des Lebensalters).

- Üblicher Ansatz: Modelliere den Einfluss des Alters durch Polynome niedrigen Grades (insbesondere quadratisch).



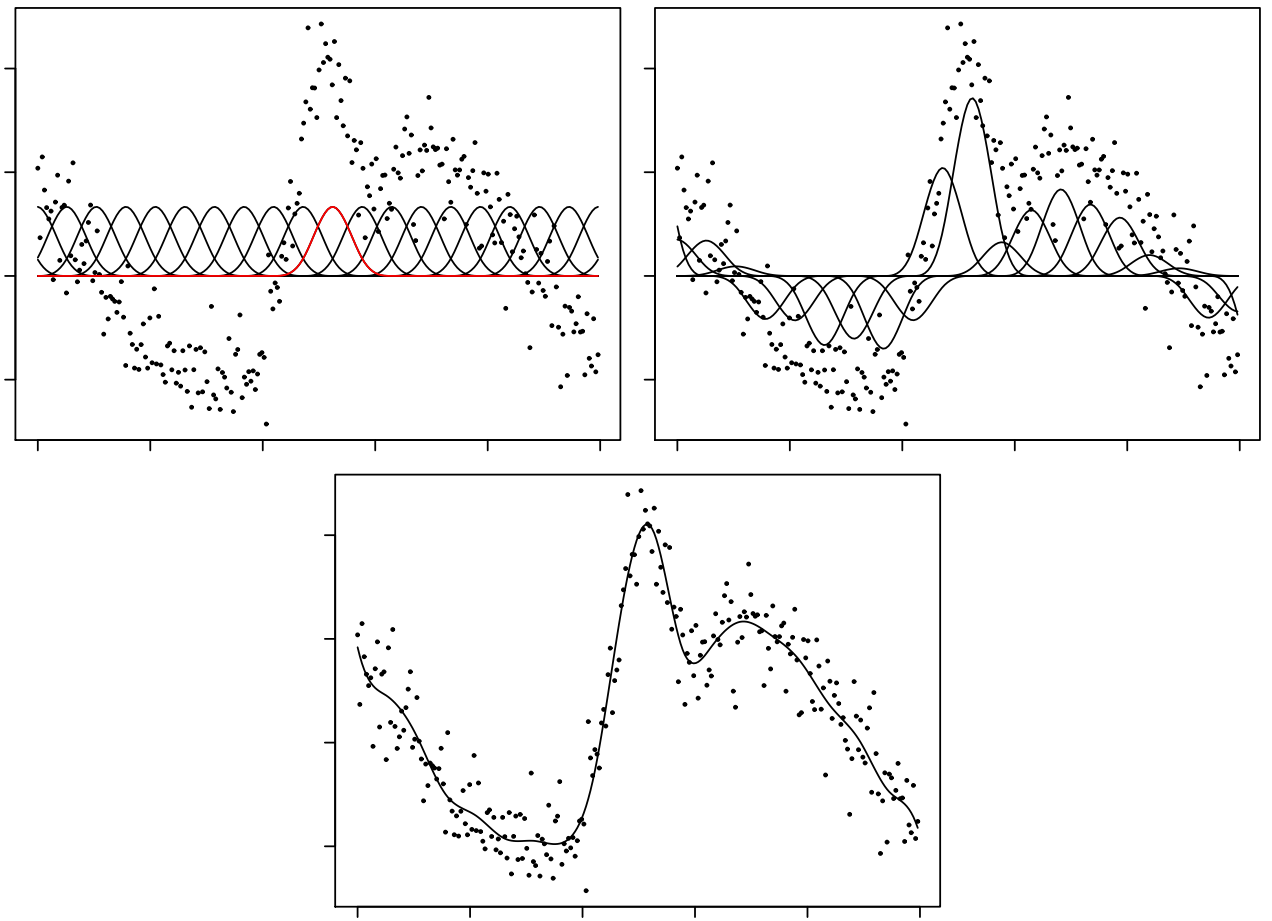
- Frage: Wie stark ist der Einfluss des polynomialen Modells auf das Ergebnis.

  $\Rightarrow$ Flexible, nichtparametrische Modellierung des Einflusses als $f(x)$.

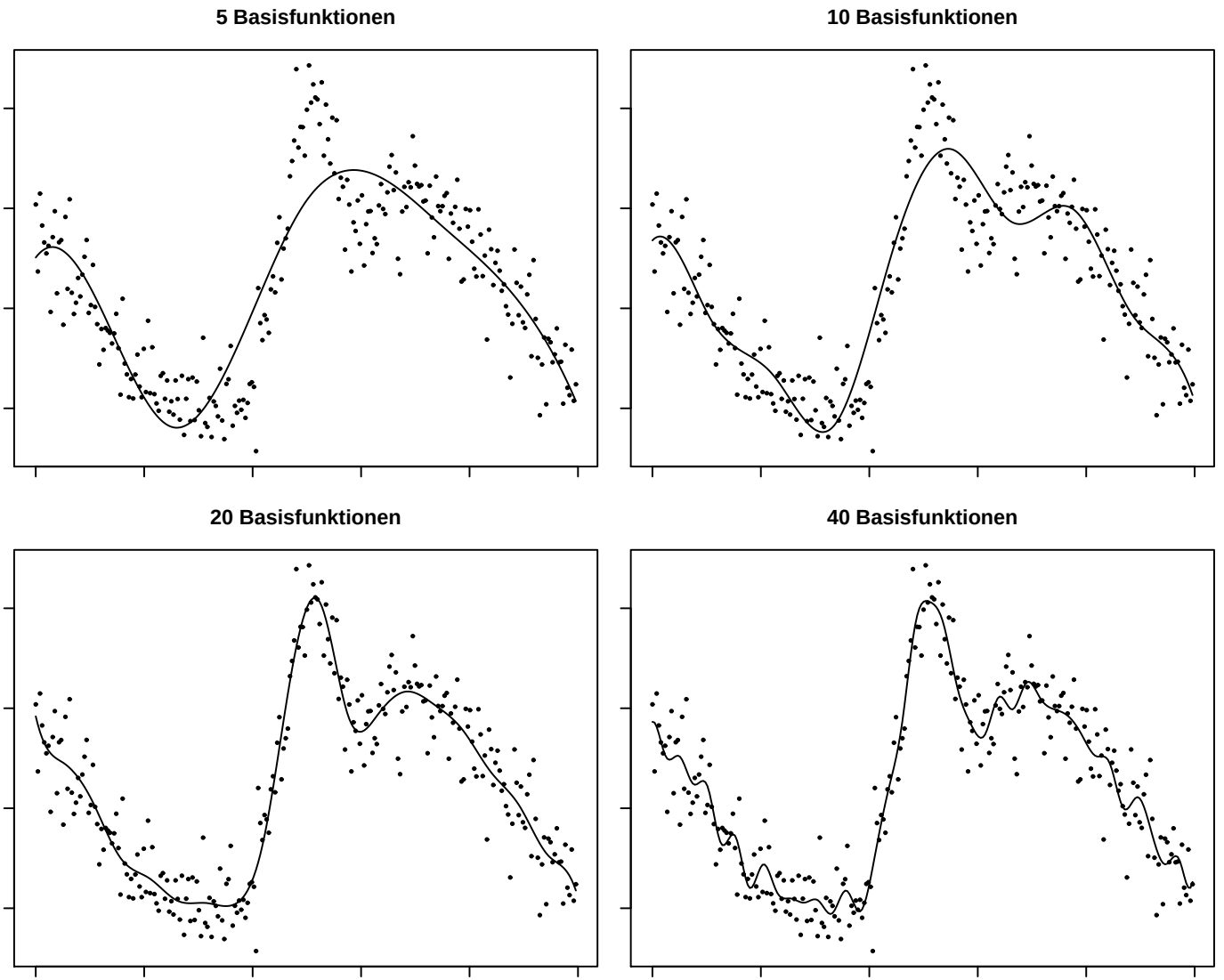
Penalisierte Splines

- Approximiere die nichtparametrisch zu schätzende Funktion $f(x)$ durch eine Linearkombination von B-Spline Basisfunktionen

$$f(x) = \sum_j \beta_j B_j(x)$$



- B-Spline Schätzungen für variierende Anzahlen von Basisfunktionen:



- Nicht regularisierte Schätzungen hängen stark von der Anzahl der Basisfunktionen ab.

  $\Rightarrow$ Ergänze die Likelihood um einen Regularisierungs-Term der raue Funktionsschätzungen bestraft.

- Beliebter Ansatz: Bestrafung der quadrierten zweiten Ableitung

$$\text{pen}(f) = \lambda \int (f''(x))^2 dx.$$

- Einfache Approximation für B-Splines: Differenzen-Strafterme, z.B. für erste Differenzen

$$\text{pen}(\beta) = \lambda \sum_j (\beta_j - \beta_{j-1})^2 = \lambda \beta' K \beta$$

- Der Glättungsparameter λ bestimmt den Einfluss der Regularisierung auf die Schätzung.

- Random Walks liefern eine äquivalente bayesianische Formulierung der Differenzen-Regularisierung, z.B. durch Random Walks erster Ordnung

$$\beta_j = \beta_{j-1} + u_j, \qquad u_j \sim N(0, \tau^2).$$

- Die gemeinsame Verteilung des Random Walks entspricht einer multivariaten Normalverteilung mit Dichte

$$p(\beta) \propto \exp\left(-\frac{1}{2\tau^2}\beta' K \beta\right).$$

- Die Präzisionsmatrix K besitzt nicht vollen Rang

  $\Rightarrow$ Teilweise uneigentliche Priori-Verteilung.

Inferenz basierend auf gemischten Modellen

- Allgemeine Struktur der betrachteten Modellklasse:
  - Beobachtungsmodell

$$y = X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p + \varepsilon.$$

  - Multivariate Gauß-Priori für die Regressionskoeffizienten

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

- Entspricht einem Regressionsmodell mit zufälligen Effekten

$$\beta_j \sim N(0, \tau_j^2 K_j^-).$$

- Problem: Die Präzisionsmatrix K_j ist i.A. nicht invertierbar.

  $\Rightarrow$ Reparametrisierung um zu einem gemischten Modell mit eigentlichen zufälligen Effekten zu gelangen.

- Inferenz basierend auf gemischten Modellen:

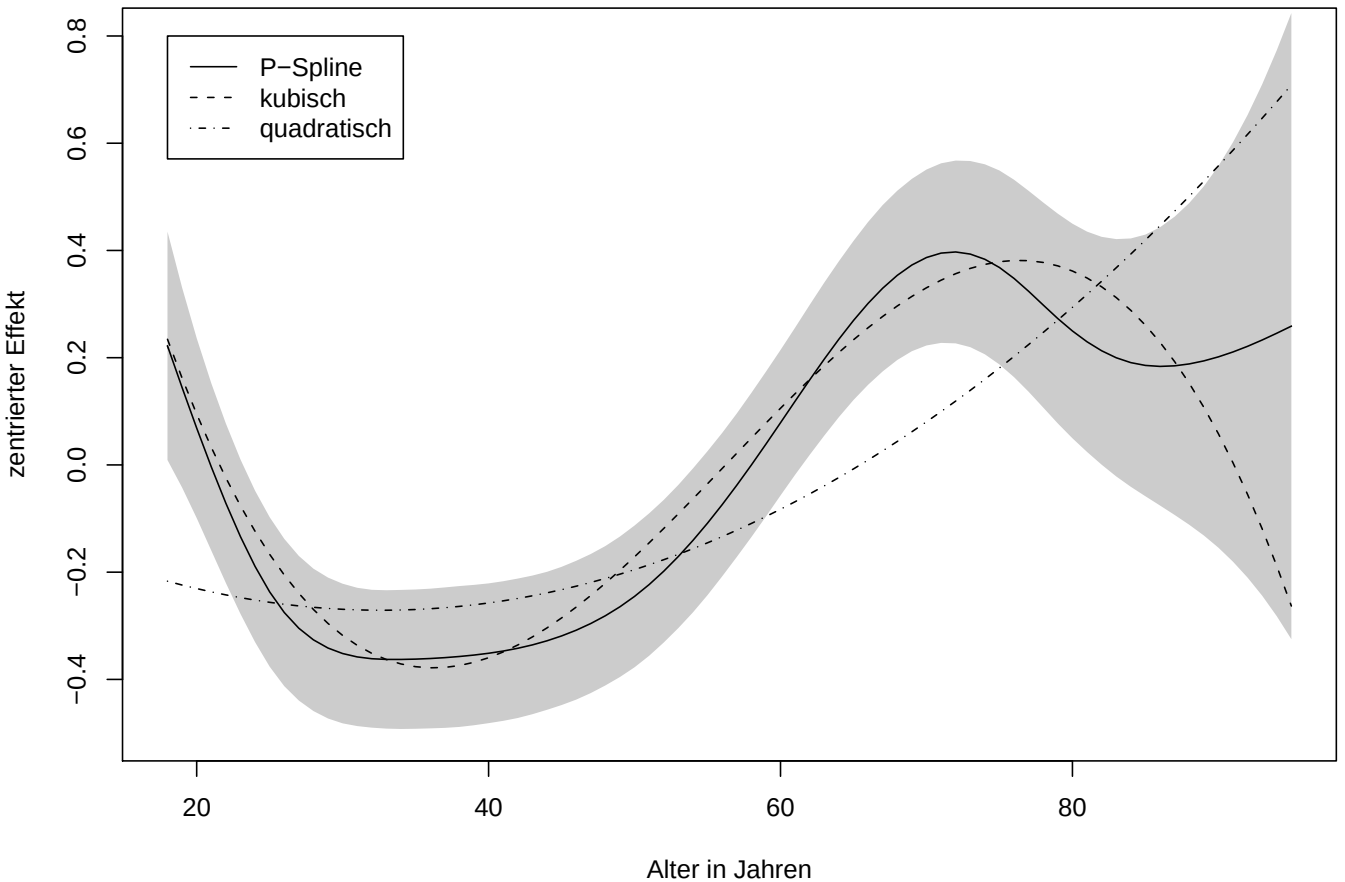
  – Penalisierte KQ-Schätzung für die Regressionskoeffizienten.

  – Restricted Maximum Likelihood-Schätzung für die Varianzparameter.

- Entspricht einem empirischem Bayes-Ansatz mit marginaler Likelihood-Schätzung für die Hyperparameter und Posteriori-Modus-Schätzung für die Regressionskoeffizienten.

Dynamik der Lebenszufriedenheit II

- Nichtparametrische Schätzung des Alterseffekts über P-Splines:

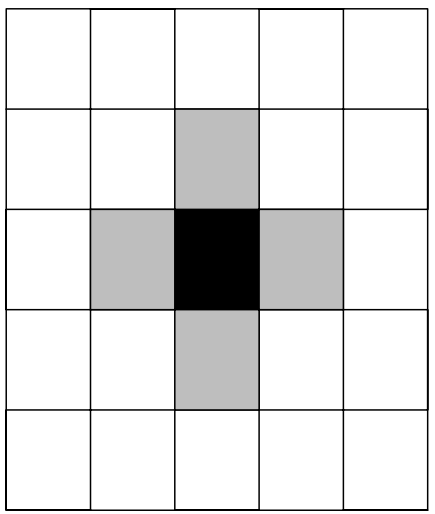


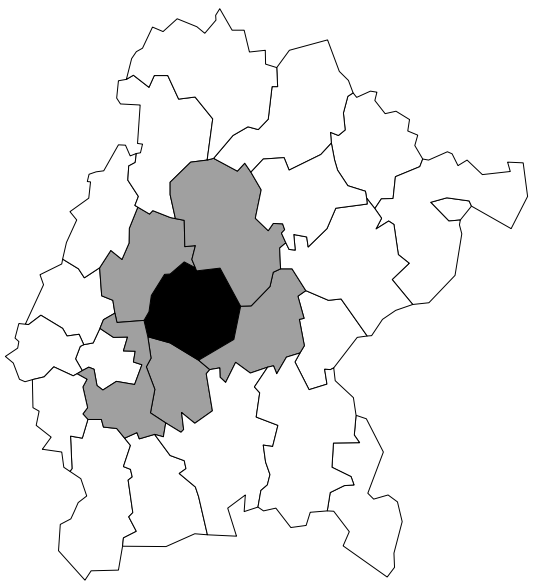
- Offene Fragen:
  - Vermutlich wird zusätzlich eine autoregressive Struktur für die Messfehler benötigt (methodisch einfach umsetzbar und z.B. im R-Paket mgcv bereits implementiert).
  - Numerisch effiziente Umsetzung, um den kompletten Datensatz analysieren zu können.
  - Modellwahl zur Unterscheidung zwischen kubischer und nichtparametrischer Modellierung (z.B. über AIC).

Räumliche Effekte für Regionendaten

- Ziel: Schätzung eines Regressionsparameters β_s für jede Region.
- Instabile Schätzung bei im Verhältnis zum Stichprobenumfang großer Regionenzahl.

  $\Rightarrow$ Regularisierte Schätzung, um räumlich glatte Effekte zu erhalten.
- Räumlich glatt: Effekte benachbarter Regionen sollten ähnlich sein.





- Strafterm basierend auf Differenzen benachbarter Regionen:

$$\text{pen}(\beta) = \lambda \sum_s \sum_{r \in N(s)} (\beta_s - \beta_r)^2$$

wobei $N(s)$ die Menge der Nachbarn der Region s bezeichnet.

- In stochastischer Formulierung ergibt sich eine Markov Zufallsfeld-Priori

$$\beta_s | \beta_r, r \in N(s) \sim N\left(\frac{1}{|N(s)|} \sum_{r \in N(s)} \beta_r, \; \frac{\tau^2}{|N(s)|} \right).$$

- Wieder erhält man eine multivariate Normalverteilung

$$p(\beta) \propto \exp\left(-\frac{1}{2\tau^2}\beta' K \beta\right)$$

wobei K eine Nachbarschaftsmatrix bezeichnet und

$$\text{pen}(\beta) = -\log(p(\beta)).$$

Hochdimensionale Kovariablenvektoren

- Regularisierung in Regressionsmodellen mit vielen Kovariablen: Bevorzuge sparsame Modelle in denen eine große Zahl von Koeffizienten nahe oder gleich Null ist.

- Beispiel Ridge-Regression:

$$LS_{\text{pen}}(\beta) = (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \rightarrow \min_{\beta}.$$

- Penalisierter KQ-Schätzer

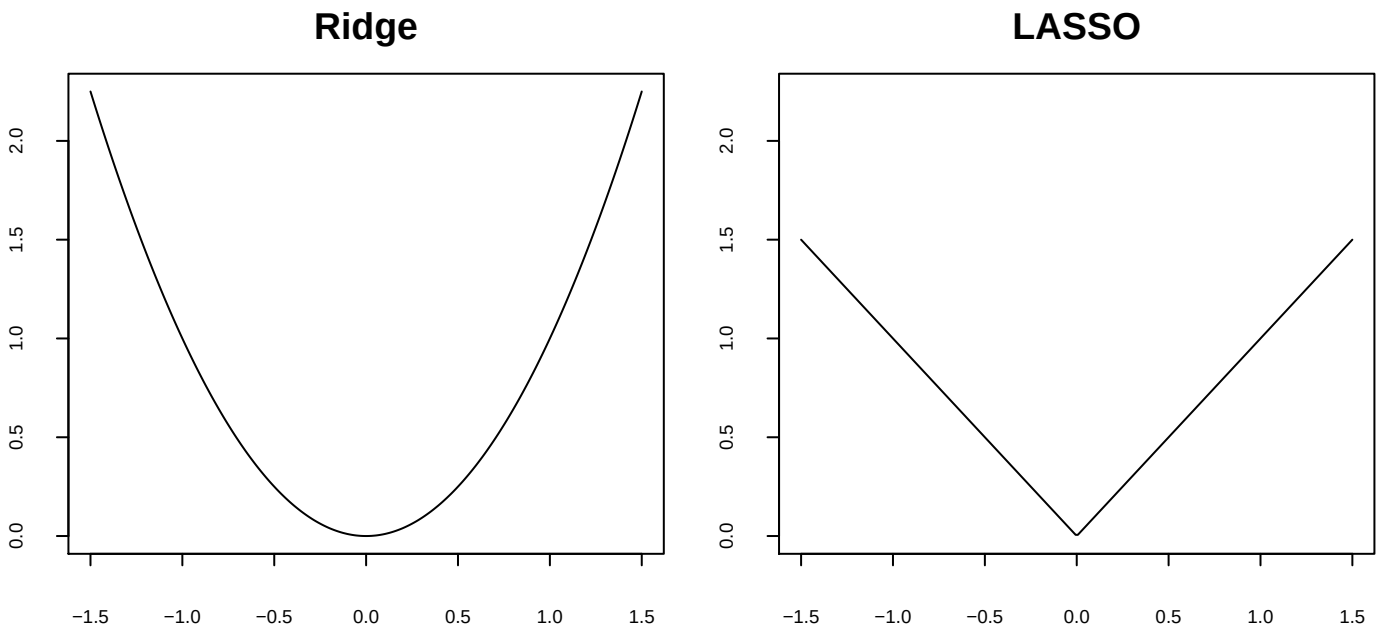
$$\hat{\beta} = (X'X + \lambda I)^{-1} X'y.$$

- Nachteil der Ridge-Regression: Die resultierenden Modelle sind nicht sparsam genug.

  $\Rightarrow$ Betrachte Strafterme mit Peak in der Null.

- Beispiel LASSO: Ersetze den quadratischen Strafterm durch den Absolutbetrag:

$$KQ_{\text{pen}}(\beta) = (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \to \min_{\beta}.$$



- (Frequentistisches) LASSO ergibt eine sparsame Lösung in der einige Koeffizienten exakt Null sind (L_1-Geometrie).

- Äquivalenzbeziehungen zwischen Penalisierungen und Priori-Verteilungen:

Strafterm	**Priori-Dichte**	**Verteilung**
Ridge	$p(\beta_j) \propto \exp(-\lambda\beta_j^2)$	Gauß
LASSO	$p(\beta_j) \propto \exp(-\lambda\|\beta_j\|)$	Laplace
L_p	$p(\beta_j) \propto \exp(-\lambda\|\beta_j\|^p)$	Power Exponential

- In der bayesianischen Formulierung passt die Laplace-Priori nicht zu unserer Standardannahme einer multivariaten Gauß-Priori.

- Allgemeine Darstellung über Skalenmischungen von Normalverteilungen:

$$p(\beta_j|\lambda) = \int_0^\infty p(\beta_j|\tau_j^2)p(\tau_j^2|\lambda)d\tau_j^2$$

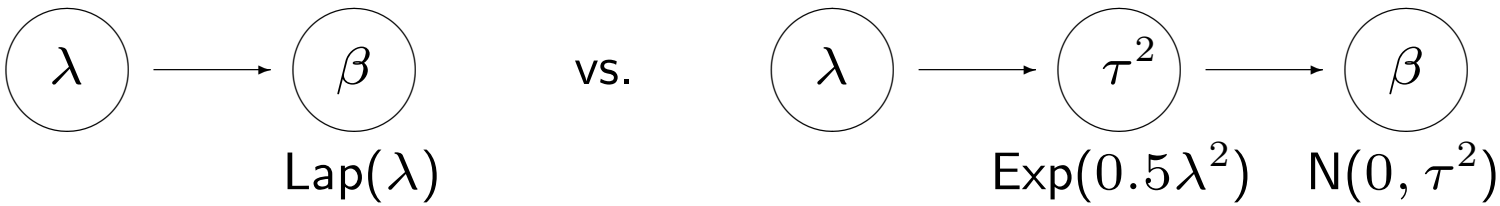
wobei

$$\beta_j|\tau_j^2 \sim N(0,\tau_j^2) \qquad \text{und} \qquad \tau_j^2|\lambda \sim p(\tau_j^2|\lambda).$$

- Speziell für die Laplace-Priori / LASSO:

$$\tau_j^2|\lambda \sim Exp\left(\frac{\lambda^2}{2}\right).$$

- Interpretation: Hierarchische Priori-Formulierung.



- Vorteil: Das Problem lässt sich auf die (einfacher zugängliche) Gauß-Formulierung zurückführen.

Markov Chain Monte Carlo Simulationsverfahren

- Beobachtungsmodell:

$$\eta = X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p.$$

- Gauß-Priori für die Regressionskoeffizienten:

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

- Geeignete Hyperprioris für τ_j^2, z.B.

$$\tau_j^2 \sim IG(a, b)$$

für penalisierte Splines und räumliche Glättung bzw.

$$\tau_j^2|\lambda \sim Exp\left(\frac{\lambda^2}{2}\right) \qquad \lambda^2 \sim Ga(a, b).$$

für LASSO-Regularisierung.

- Für normalverteilte Zielvariablen ergeben sich die vollständig bedingten Dichte in geschlossener Form und man erhält einen Gibbs-Sampler.

- Wesentliche Bestandteile:
  - Vollständig bedingte Verteilung für Regressionskoeffizienten β_j: $N(\mu_j, \Sigma_j)$ mit

$$\mu_j = \left(X_j' X_j + \frac{\sigma^2}{\tau_j^2} K_j \right)^{-1} X_j'(y - \eta + X_j \beta_j), \qquad \Sigma_j = \left(X_j' X_j + \frac{\sigma^2}{\tau_j^2} K_j \right)^{-1}$$

  - Vollständig bedingte Verteilung für Glättungsvarianzen: $\tau_j^2 \sim IG(a_j, b_j)$ mit

$$a_j = a + \frac{1}{2} \operatorname{rang}(K_j), \qquad b_j = b + \frac{1}{2} \beta_j' K_j \beta_j.$$

  - Für das bayesianische LASSO ergeben sich als vollständig bedingte Verteilungen inverse Gauß- und Gamma-Verteilungen.

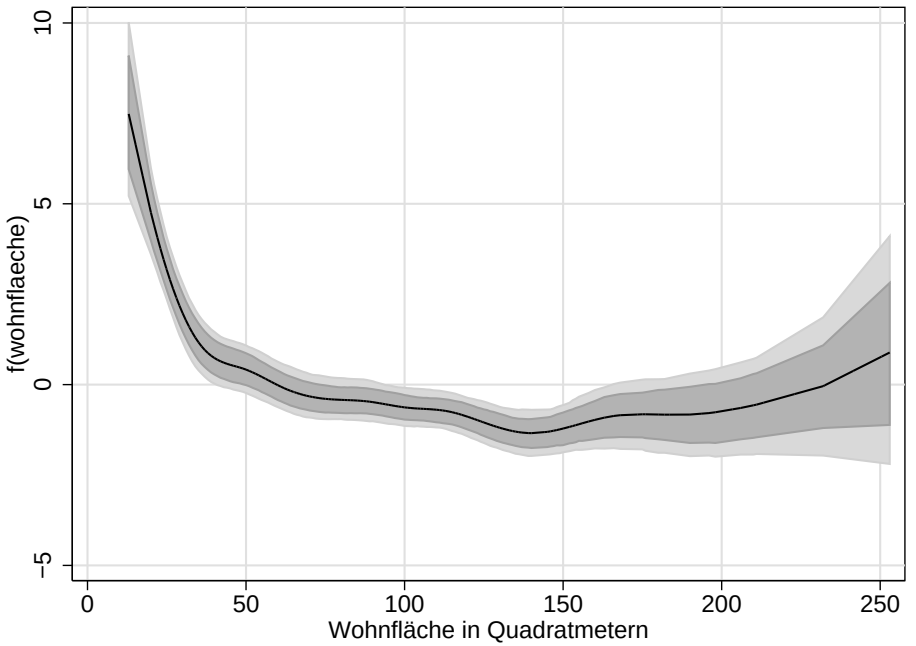
Münchner Mietspiegel

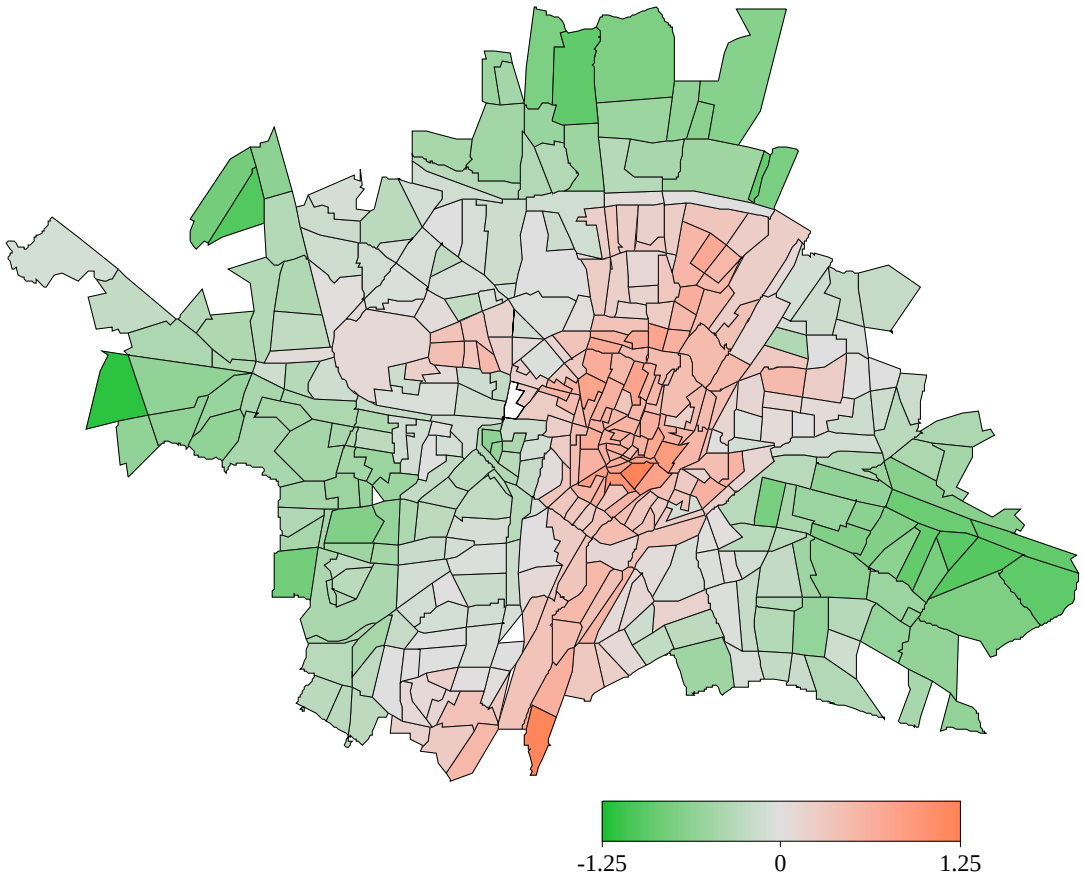
- Erstellt zur Bestimmung der ortsüblichen Vergleichsmiete.
- Ältere Mietspiegel beruhten auf Gruppierung und Mittelwertbildung.
- Zunehmende Verwendung von Regressionsmodellen mit Nettomiete pro Quadratmeter als Zielvariable (volle Ausnutzung der vorhandenen Information).
- Mögliches Regressionsmodell für die Nettomiete pro qm nm:

$$nm = f_1(wohnflaeche) + f_2(baujahr) + f_3(bezirksviertel) + x'\beta + \varepsilon.$$

wobei $x'\beta$ die Effekte eines potenziell hochdimensionalen Vektors von Kovariablen beinhaltet.

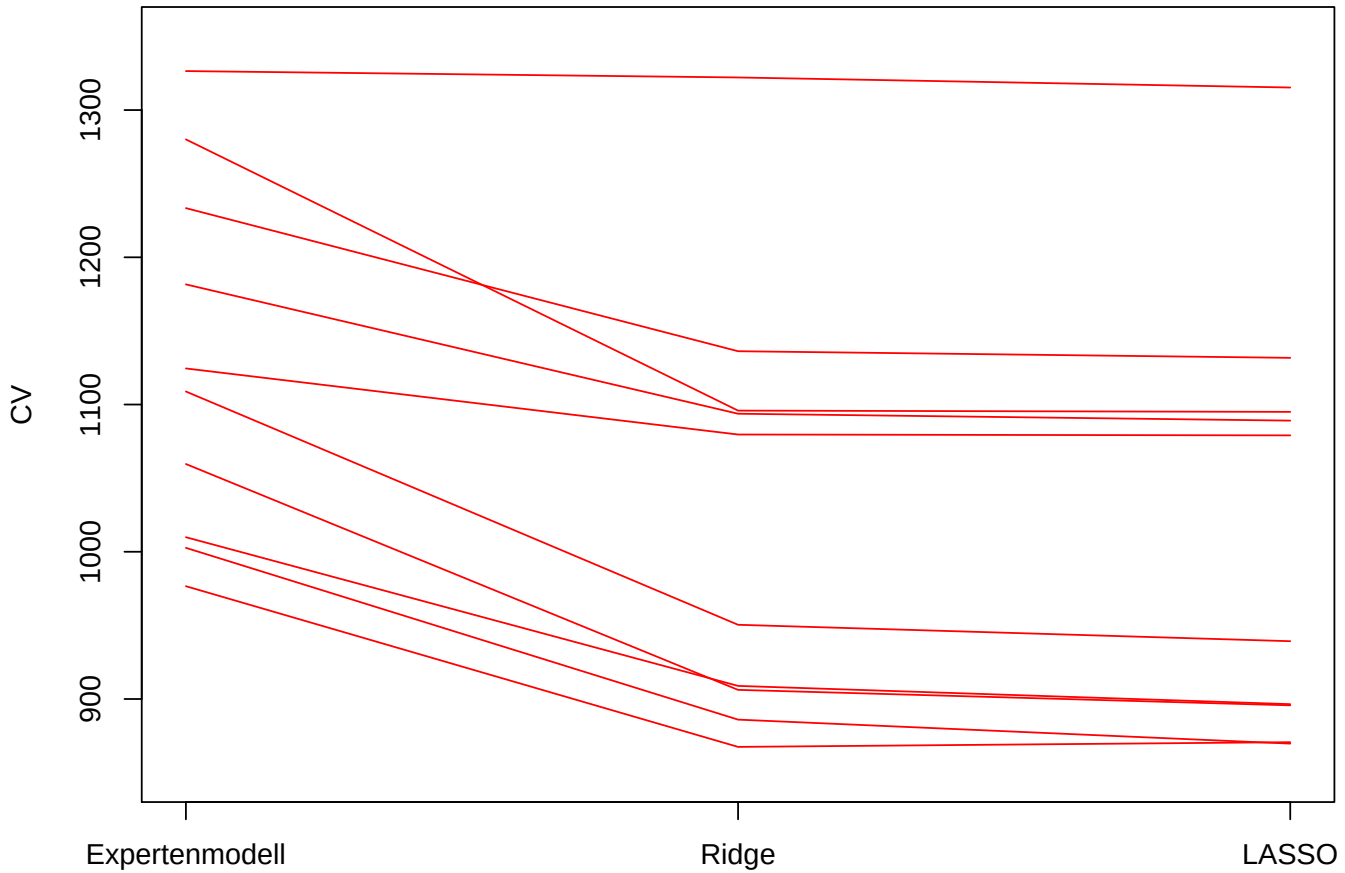
- Im Folgenden Ergebnisse bei LASSO-Regularisierung des Vektors β.





- Interpretierbare Ergebnisse, aber was gewinnt man für die Prognose?

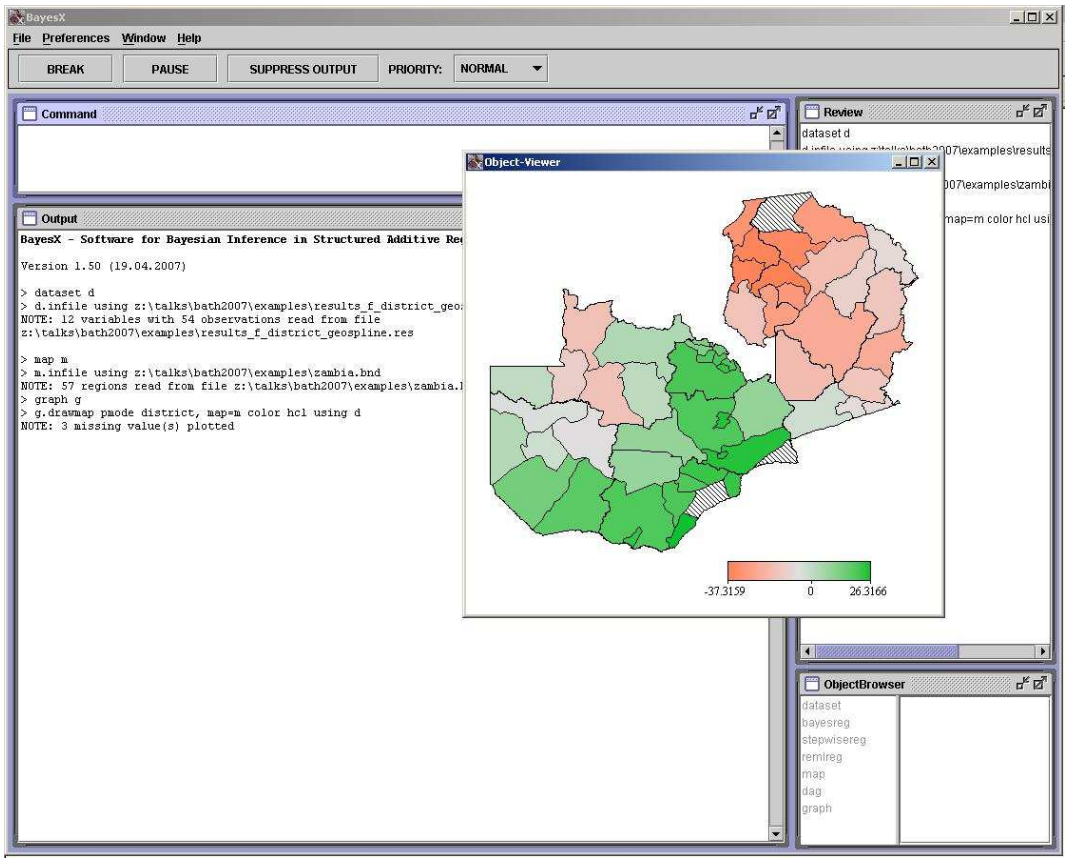
- Vergleich eines Expertenmodells (Subvektor der Kovariablen), der Ridge-Regression und der LASSO-Regression über 10-fache Kreuzvalidierung.



$\Rightarrow$ Deutlich verbesserte Vorhersageeigenschaften durch Regularisierung!

BayesX

- Die beschriebene Methodik ist implementiert im freien Software-Paket BayesX.



- Erhältlich unter

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

Erweiterungen

- Die betrachtete Modellklasse und entwickelte Methodik trägt deutlich weiter.
- Zusätzliche Modellterme (Fahrmeir, Kneib & Lang, 2004):
  - Variierende Koeffizienten,
  - Oberflächenschätzung,
  - Kriging, . . .
- Allgemeinere Zielvariablen:
  - Sample Selection Modelle (Kneib & Wiesenfarth, 2008),
  - Univariate Exponentialfamilien (Fahrmeir, Kneib & Lang, 2004),
  - Kategoriale Zielvariablen (Kneib & Fahrmeir, 2006, Kneib, Baumgartner & Steiner, 2007),

  - Verweildauern und Mehrstadien-Modelle (Kneib, 2006, Kneib & Fahrmeir, 2007, Kneib & Hennerfeind, 2008, Konrath, Kneib & Fahrmeir, 2008).

- Modellwahl und Variablenselektion über Boosting (Kneib, Hothorn & Tutz, 2008).

- Die Darstellung über gemischte Modelle erlaubt auch theoretische Untersuchungen z.B. zur Normierbarkeit der Posteriori-Verteilung (Fahrmeir & Kneib, 2008).

Zusammenfassung

- Semiparametrische Regressionsmodelle bilden eine flexible, reichhaltige Modellklasse.
- Einheitliche Behandlung unterschiedlicher Regularisierungsansätze und -ziele in einem bayesianischen Ansatz.
- Voll-automatische Schätzung aller relevanten Modellparameter (insbesondere der Glättungsparameter).

- Anknüpfungspunkte zu Sozial- und Bildungswissenschaften:
  - Mehrebenenanalyse beispielsweise in der Schulentwicklungsforschung.
  - Räumliche Effekte etwa zur räumlich differenzierten Analyse des sozialen Wandels oder in bildgebenden Verfahren.
  - Verweildauer- und Mehrstadienmodelle in der Analyse von Migration oder sozialer Mobilität.
  - Modellierung von Longitudinaldaten (Kombination aus flexibler Trendschätzung und zufälligen Effekten).
  - Geoadditive Regression für Modelle mit latenten Variablen.

- Dank an
  - das StaBLab und Christoph Wunder (Dynamik der Lebenszufriedenheit),
  - Felix Heinzl, Susanne Konrath & Ludwig Fahrmeir (DFG Projekt Bayesianische Regularisierung).

- Mehr Informationen unter

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

Referenzen

- Fahrmeir, L. & Kneib, T. (2008): Propriety of Posteriors in Structured Additive Regression Models: Theory and Empirical Evidence. *Journal of Statistical Planning and Inference*, to appear.
- Fahrmeir, L., Kneib, T. & Lang, S. (2004): Penalised structured additive regression for space-time data: A Bayesian perspective, *Statistica Sinica*, 14, 731–761.
- Kneib, T. (2006): Geoadditive Hazard Regression for Interval Censored Survival Times. *Computational Statistics & Data Analysis*, 51, 777–792.
- Kneib, T., Baumgartner, B. & Steiner, W. J. (2007): Semiparametric Multinomial Logit Models for Analysing Consumer Choice Behaviour. *Advances in Statistical Analysis*, 91, 225–244.
- Kneib, T. & Fahrmeir, L. (2006): Structured Additive Regression for Categorical Space-Time Data: A Mixed Model Approach. *Biometrics*, 62, 109–118.
- Kneib, T. & Hennerfeind, A. (2008): Bayesian Semiparametric Multi-State Models. *Statistical Modelling*, 8, 169–198.
- Kneib, T., Hothorn, T. & Tutz, G. (2008): Model Choice and Variable Selection in Geoadditive Regression Models. *Biometrics*, to appear.
- Kneib & Wiesenfarth (2008): Geoadditive Sample Selection Models, in preparation.
- Konrath, Kneib & Fahrmeir (2008): Bayesian Regularisation in Structured Additive Regression Models for Survival Data, Technical Report.