Bayesianische Regularisierungs-Prioris

Thomas Kneib

gemeinsam mit

Ludwig Fahrmeir, Susanne Konrath & Fabian Scheipl

Institut für Statistik
Ludwig-Maximilians-Universität München

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

26.8.2008

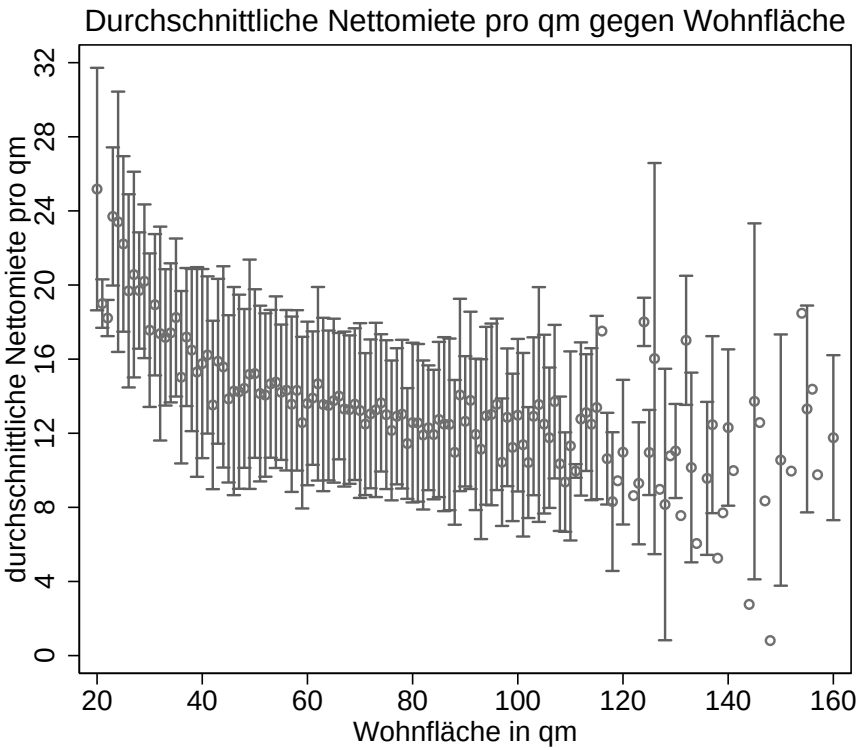
Übersicht

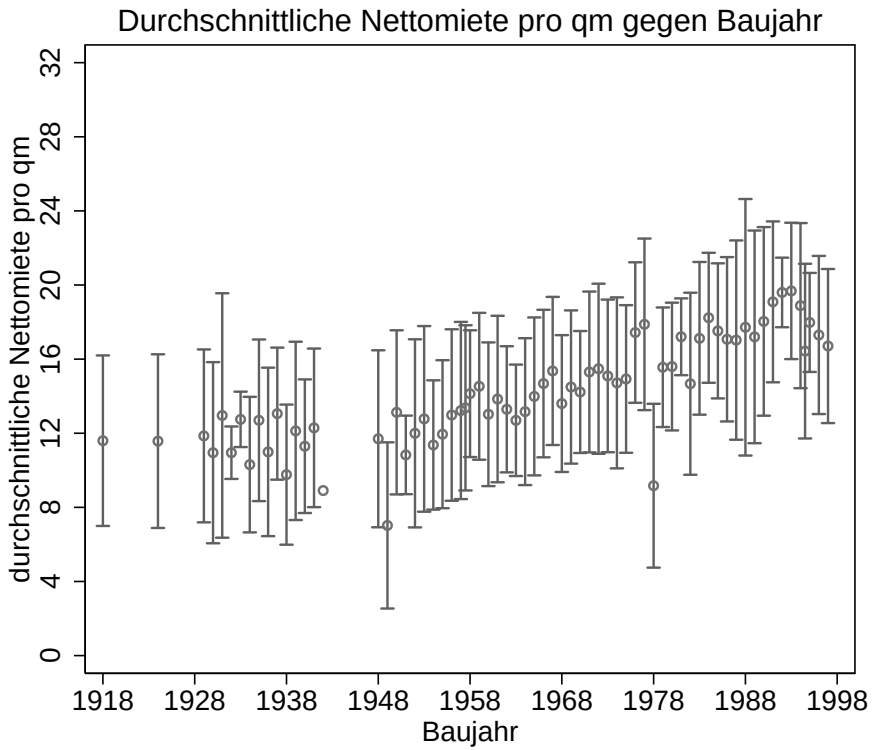
- Bayesianische Regularisierung
- Regularisierung in geoadditiven Regressionsmodellen
- Regularisierung in hochdimensionalen Regressionsmodellen
- Anwendung: Münchner Mietspiegel

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).
- Typisches Vorgehen: Variablenselektion und Modellwahl zur Bestimmung der relevanten Kovariablen und der geeigneten Modellierungsform.

- Probleme:
  - Einbezug nichtlinearer Effekte beispielsweise der Wohnfläche oder des Baujahrs.





– Einbezug räumlicher Information über die Experteneinschätzung zur Lage hinaus.

  - Ist es wirklich sinnvoll nur einen Teil der Kovariablen in der Prognose neuer Mieten zu verwenden?

- 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.

Bayesianische 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.
  - Hochdimensionale Kovariablen: Regularisierung durch geeignete Normen der Regressionskoeffizienten β um sparsame Modelle zu erzielen.

- Verschiedene Ziele: Glattheitseigenschaften bzw. sparsame Modellstruktur.

- Frequentistischer Regularisierungsansatz: Ergänze einen Strafterm zur Likelihood bzw. allgemeiner dem zur Schätzung verwendeten Kriterium.

- Beispiel: Ridge-Regression im linearen Modell

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

- Für hochdimensionale Kovariablenvektoren wird die Kleinste Quadrate-Schätzung von β instabil.

  $\Rightarrow$ Ergänze das KQ-Kriterium um einen quadratischen Strafterm

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

- Lösung in geschlossener Form: Penalisierter Kleinste Quadrate (PKQ)-Schätzer

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

- Bayesianische Formulierung des linearen Modells:

$$y = X\beta + \varepsilon, \qquad \beta \sim \mathsf{N}(0, \tau^2 I).$$

- Man erhält die Posteriori-Verteilung

$$p(\beta|y) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \exp\left(-\frac{1}{2\tau^2}\beta'\beta\right)$$

- Maximierung der Posteriori-Verteilung ist äquivalent zur Minimierung des penalisierten KQ-Kriteriums

$$(y - X\beta)'(y - X\beta) + \lambda\beta'\beta$$

wobei der Glättungsparameter gegeben ist durch das Signal-Rauschen-Verhältnis

$$\lambda = \frac{\sigma^2}{\tau^2}.$$

- Der Posteriori Modus ist äquivalent zum penalisierten KQ-Schätzer.

- Analog für allgemeinere Formen von Priori-Verteilungen:
  - Penalisierte Log-Likelihood:

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

  - Posteriori:

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

- Insgesamt erhält man die Beziehung

Penalisierung $\equiv$ logarithmierte Priori-Verteilung.

- Wesentliche Klasse im Folgenden: Quadratische Strafterme

$$\text{pen}(\beta) = \lambda \beta' K \beta$$

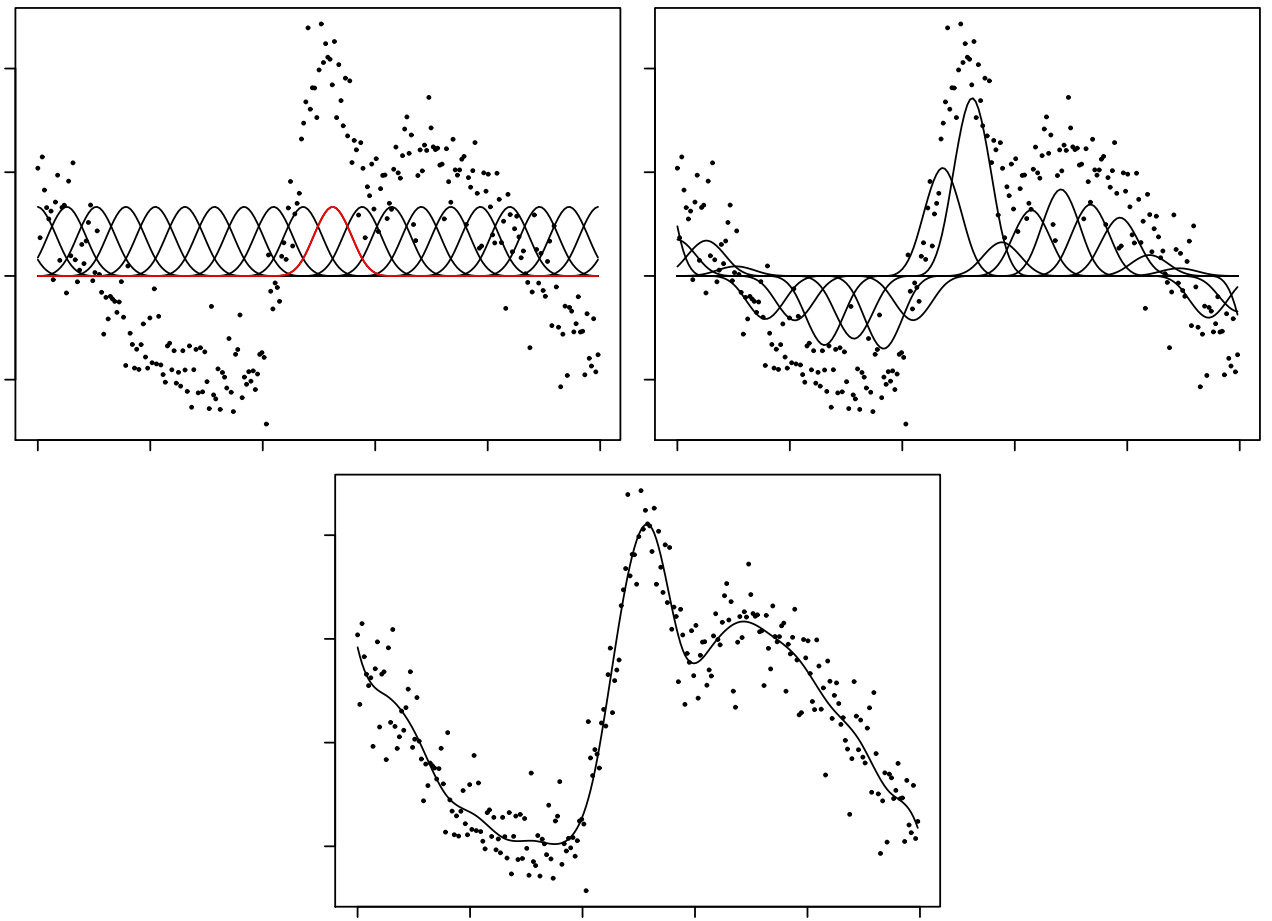
bzw. multivariate Gauß-Prioris

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

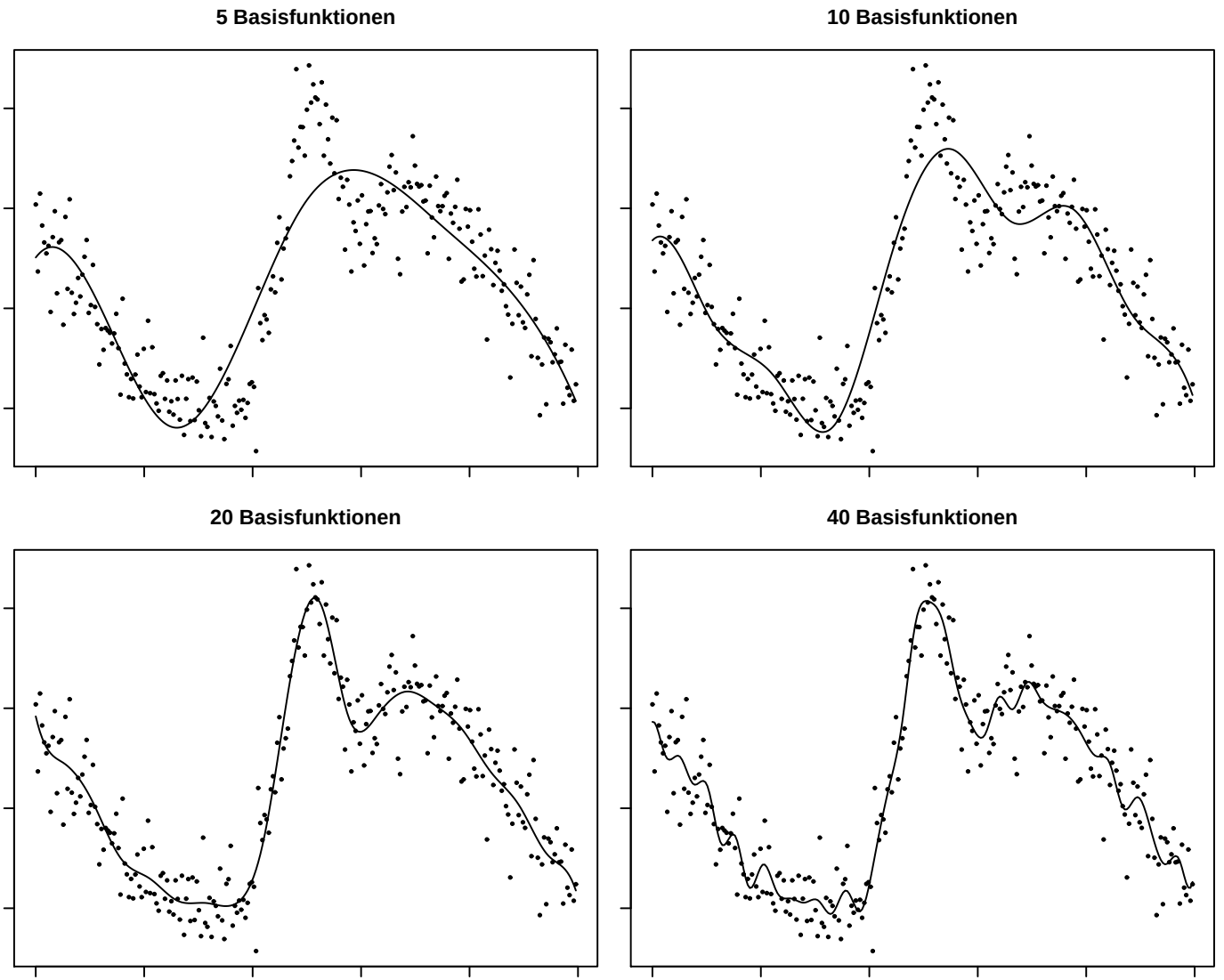
Penalisierte Spline-Glättung

- Approximiere eine 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:



- Unregularisierte 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 (sollte mitgeschätzt werden).

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

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

- Die gemeinsame Verteilung des Random Walks ergibt sich zu

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

- Man erhält die angekündigte quadratische Penalisierung bzw. multivariate Gauß-Priori.

Räumliche Effekte für Regionendaten

- Ziel: Schätzung eines separatem 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

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

- 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

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

Regularisierung in hochdimensionalen Modellen

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

- Beispiel Ridge-Regression:

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

- Penalisierter KQ-Schätzer

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

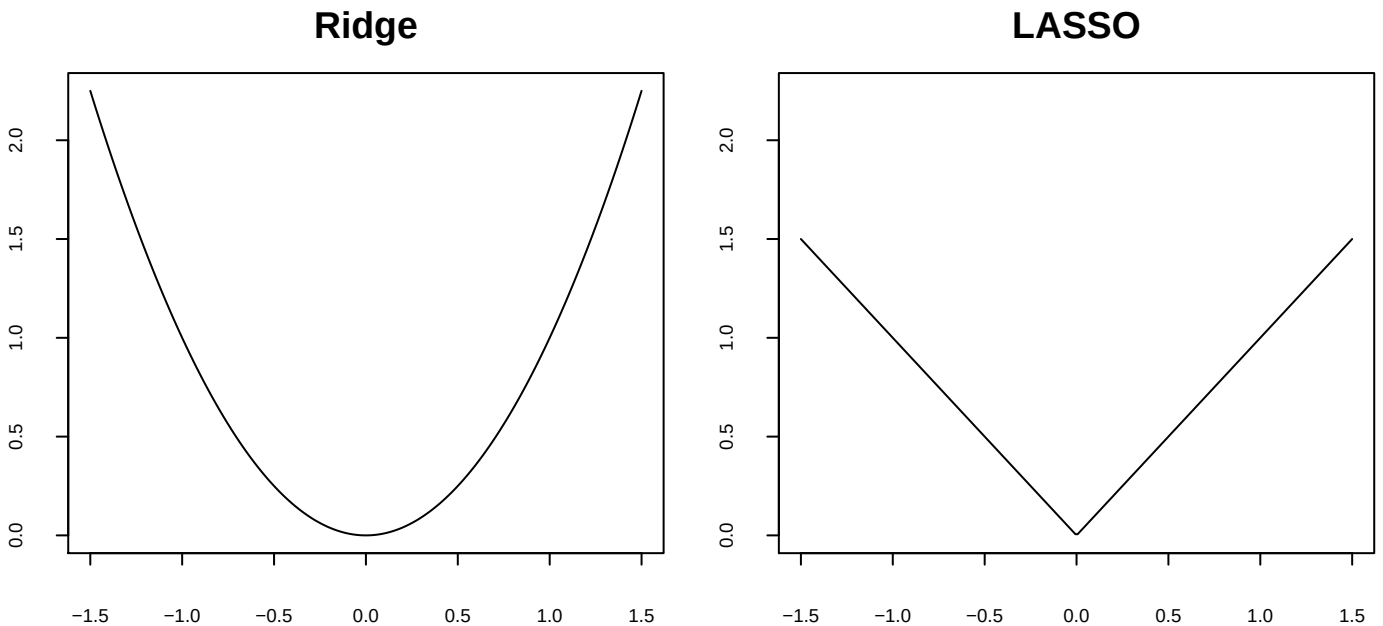
- Der PKQ-Schätzer ist verzerrt, besitzt aber im Vergleich zum KQ-Schätzer eine geringere Varianz.

- Geeignete Glättungsparameter sollten zu einem reduzierten MSE führen.
- 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:

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

- (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

- Der LASSO-Strafterm ist nicht differenzierbar (in Null)

  $\Rightarrow$ Numerische Schwierigkeiten bei der Berechnung.

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

- Allgemeine Darstellung über Skalenmischungen von Normalverteilungen:

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

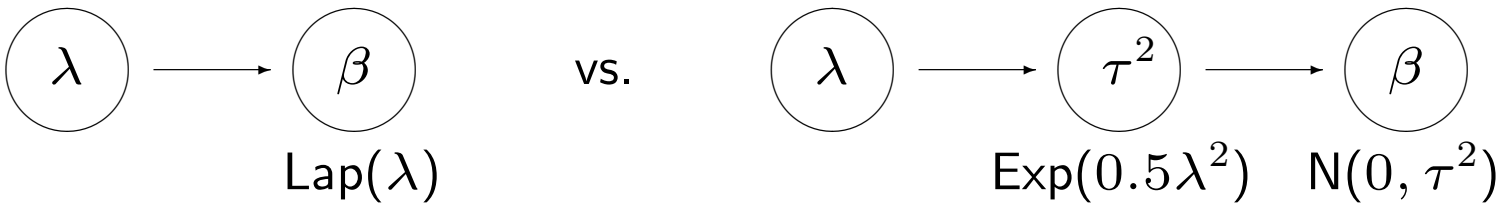
  wobei

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

- Speziell für die Laplace-Priori / LASSO:

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

- Interpretation: Hierarchische Priori-Formulierung.



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

Bayesianische Inferenz

- Inferenz basierend auf Markov chain Monte Carlo (MCMC) Simulationen erlaubt die Kombination der skizzierten Regularisierungsansätze.

- Allgemeine Modellstruktur:

  – 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.

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

für penalisierte Splines und räumliche Glättung bzw.

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

für LASSO-Regularisierung.

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

- Für allgemeinere Zielvariablen benötigt man geeignete Vorschlagsdichten (für die Regressionskoeffizienten).

- Idee: Approximiere die vollständig bedingte Dichte durch eine Normalverteilung mit geeignetem Modus und geeigneter Streuung.

- Verwende dazu den iterativ gewichteten KQ-Schätzer der penalisierten Likelihood-Schätzung.

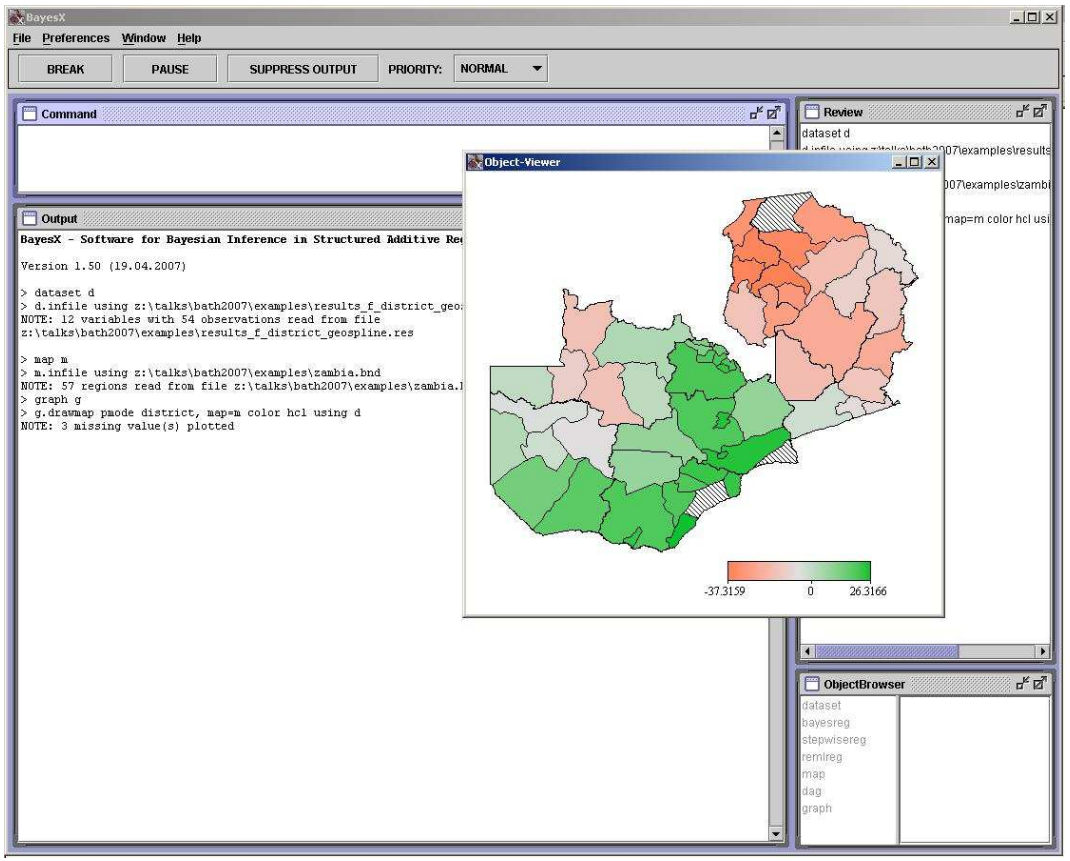
- Es ergibt sich eine multivariate Normalverteilung als Vorschlagsdichte mit Präzisionsmatrix und Erwartungswert

$$P_j = X_j' W X_j + \frac{1}{\tau_j^2} K_j \qquad \text{und} \qquad m_j = P_j^{-1} X_j' W (\tilde{y} - \eta_{-j}).$$

- Vorteile der Schätzung basierend auf MCMC:
  - Die modulare Struktur erlaubt die einfache Modifikation einzelner Modellteile und die Zerlegung in kleine Teilprobleme.
  - Exakte Unsicherheitsmaße sind erhältlich.
  - Einfach auf verschiedene Typen von Zielvariablen zu verallgemeinern (implementiert für Exponentialfamilien und Verweildauern).
- Nachteile:
  - Konvergenz und Mixing der simulierten Markov-Kette müssen überwacht werden.
  - Sensitivität bezüglich der Priori-Spezifikationen.

BayesX

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



- Erhältlich unter

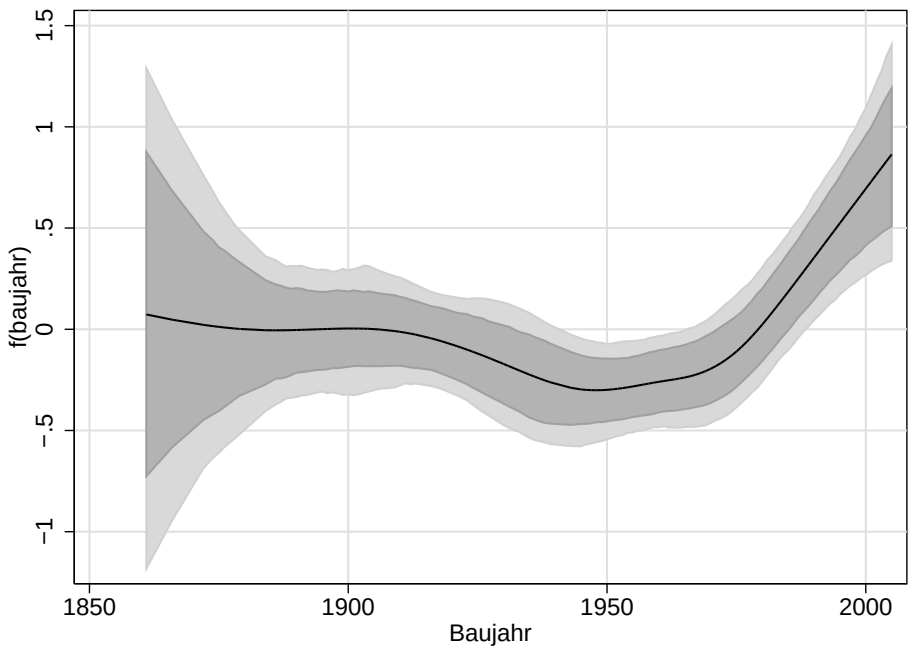
`http://www.stat.uni-muenchen.de/~bayesx`

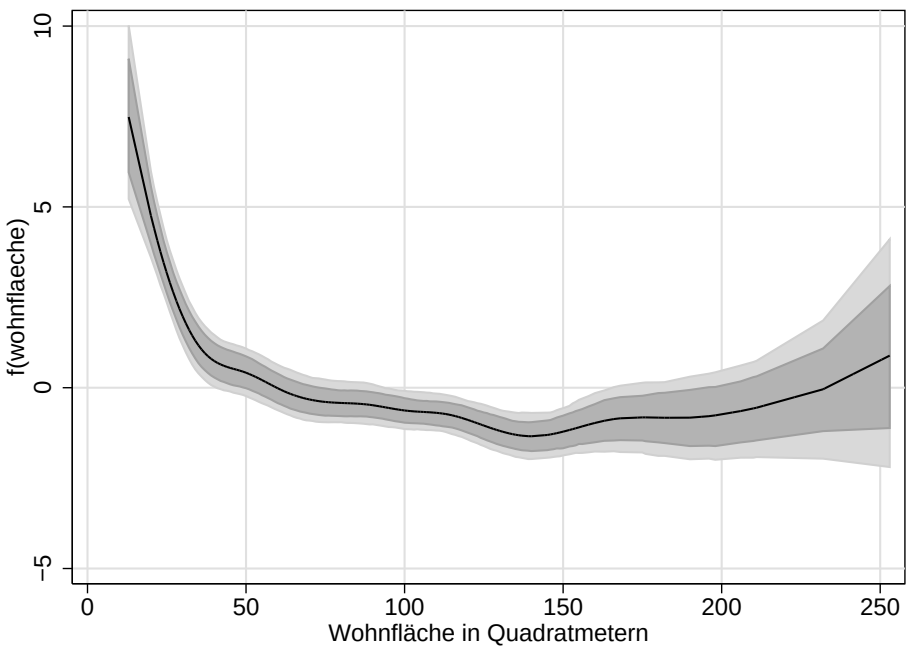
Münchner Mietspiegel: Ergebnisse

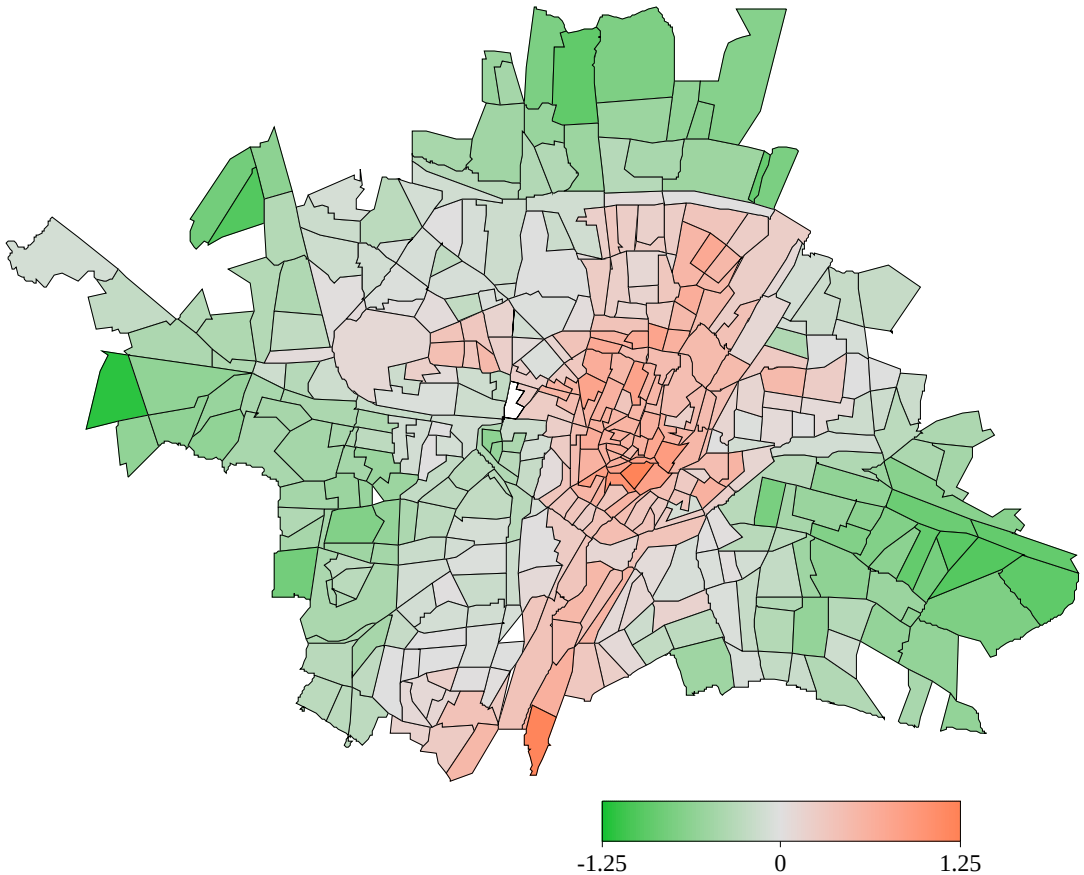
- Modellgleichung

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

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

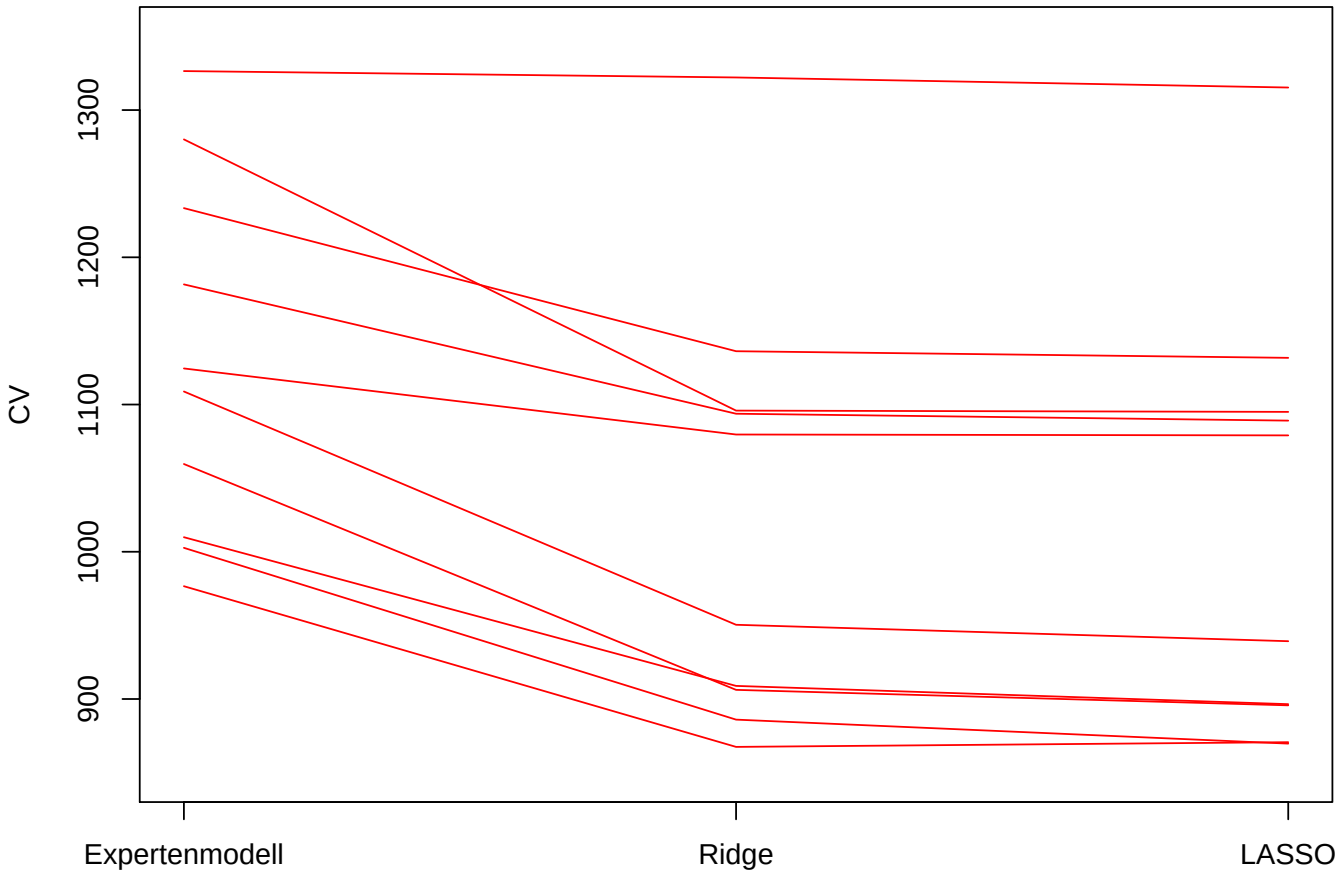






- Interpretierbare Ergebnisse, aber was gewinnt man für die Prognose?

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



$\Rightarrow$ Deutlich verbesserte Vorhersageeigenschaften durch Regularisierung!

Zusammenfassung

- Bayesianische Regularisierung komplexer Regressionsmodelle.
- Einheitliche Regularisierungs-Form zur glatten Funktionsschätzung und zur Erzielung sparsamer Modelle.
- Insbesondere in vorhersage-orientierten Anwendungen von Interesse.
- Erweiterungen:
  - Das Bayesianische LASSO besitzt keine Variablenselektions-Eigenschaft

    $\Rightarrow$ Betrachte Mischverteilungs-Prioris mit Punktmasse in Null.
  - Anwendung der Regularisierungs-Prioris mit Peaks in der adaptiven Funktionsschätzung.
  - Schätzung zahlreicher Submodelle auch über Laplace-Approximation / empirischen Bayes-Ansatz.

- Mehr Informationen unter

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