Bayesianische Regression

Lineare und logistische Modelle

Lona Koers

LMU

25. Juli 2025

Motivation für bayesianische Regression

Regressionsmodelle sind in der Wissenschaft und bei Anwendern sehr beliebt, aber ...

→ ... wie können wir Vorwissen (korrekt) modellieren?

→ ... wie können wir Unsicherheit im Modell und in der Vorhersage darstellen?

Outline

Frequentistisches $\rightarrow$ bayesianisches lineares Modell

Annahmen:

1. i.i.d. Daten $\boldsymbol{D} = (\boldsymbol{y}, \boldsymbol{X})$
2. Kondition auf $\boldsymbol{X}$ (implizit)

Frequentistisches lineares Modell: $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$

3. Gewichtsparameter $\boldsymbol{\theta}$ als Zufallsvariable interpretieren

Bayesianisches lineares Modell:

$$\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$$

Priori-Verteilungen für das bayesianische LM

A-Priori-Annahme für $\boldsymbol{\theta}$ (und evtl. σ^2) notwendig $\rightarrow$ sehr vielseitige Modell-Anpassung möglich

Normal-Invers-Gamma Priori:

$$\begin{aligned}\boldsymbol{\theta} \mid \sigma^2 &\sim \mathcal{N}(\breve{\boldsymbol{\mu}}, \sigma^2 \breve{\Sigma}) \\ \sigma^2 &\sim \text{IG}(\breve{a}, \breve{b}) \\ \boldsymbol{\theta}, \sigma^2 &\sim \text{NIG}(\breve{\boldsymbol{\mu}}, \sigma^2 \breve{\Sigma}, \breve{a}, \breve{b})\end{aligned}$$

mit Priori Parametern: $\breve{\boldsymbol{\mu}}, \breve{\Sigma}, \breve{a}$ und $\breve{b}$

Vorteil: NIG-Priori ist mit Normalverteilungs-Likelihood konjugiert $\rightarrow$ exakte Inferenz möglich (mehr dazu später)

Uninformative Priori als Spezialfall der NIG-Priori

Uninformative Priori

z.B. mit NIG-Priori mit Priori Parametern

$$\breve{\boldsymbol{\mu}} = \mathbf{0}, \quad \breve{\Sigma}^{-1} = \mathbf{0} \text{ i.e. } \breve{\Sigma} \to \infty, \quad \breve{a} = -\frac{p}{2}, \quad \breve{b} = 0$$

$\implies$ flache (und damit uninformative) Priori und maximaler Einfluss der Daten auf die Posteriori:

$$\boldsymbol{\theta} \mid \sigma^2 \overset{a}{\sim} \mathcal{N}(\breve{\boldsymbol{\mu}}, \sigma^2 \infty) \implies p(\boldsymbol{\theta} \mid \sigma^2) \propto 1$$

Regularisierung: frequentistisch $\rightarrow$ bayesianisch

Erinnerung: Regularisierung im *frequentistischen* LM durch Minimierung von

$$\text{PLS}(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda \text{ pen}(\boldsymbol{\theta})$$

mit Regularisierungs-Parameter $\lambda > 0$.

Bayesianische Regularisierung durch Wahl der Priori-Verteilung für $\boldsymbol{\theta}$

Regularisierung durch Wahl der Priori-Verteilung

Ridge Regularisierung

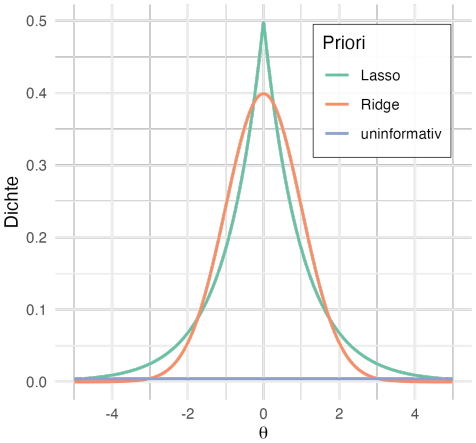
- Frequentistisch [13, 14]: $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$
- Bayesianisch [16]: $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I})$ mit $\tau^2 \propto \frac{1}{\lambda}$

Lasso Regularisierung

- Frequentistisch [24]: $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$
- Bayesianisch [20]:

$$\boldsymbol{\theta} \mid \boldsymbol{\tau}^2 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\tau}^2 \boldsymbol{I})$$

$$\tau_j^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(0.5\lambda^2), \quad j = 1, \ldots, p$$



Problem: keine Variablenselektion (im Gegensatz zu frequentistischem Lasso)

$\rightarrow$ Alternative Priori für Variablenselektion: Spike and Slab [17], Horseshoe [3], u.v.m.

Outline

Bayesianisches LM $\rightarrow$ **GLM**

$$\text{LM: } \boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X\theta}, \sigma^2 \boldsymbol{I}) \quad \rightarrow \quad \text{GLM: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim F(g^{-1}(\boldsymbol{X\theta}))$$

- Verteilungsannahme von $\boldsymbol{y}$ wird (äquivalent zum frequentistischen GLM) auf alle Verteilungen F der Exponentatialfamilie ausgeweitet
- Skala des linearen Prädiktors $\boldsymbol{X\theta}$ wird mit der Link-Funktion g^{-1} angepasst

Bayesianisches GLM → **logistisches** Modell

→ Bayesianisches logistisches Modell

$$\begin{aligned} \boldsymbol{y}_i \mid \boldsymbol{\theta} &\sim \text{Bin}(1, g^{-1}(\boldsymbol{x}_i\boldsymbol{\theta})), \quad i = 1, \ldots, n \\ g^{-1}(\boldsymbol{x}_i\boldsymbol{\theta}) &= \sigma(\boldsymbol{x}_i\boldsymbol{\theta}) \end{aligned}$$

Für Beobachtungen $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ und Sigmoid-Link $\sigma(y) = \frac{\exp(y)}{1+\exp(y)}$

Priori Wahl

- I. Allg. äquivalent zum LM möglich, z.B. Normalverteilungs-Priori
- Für Regularisierung können dieselben Priori wie im LM verwendet werden [19, 6, 5]
- Verteilungen mit schweren Rändern (z.B. t-Verteilung, Cauchy Verteilung) adressieren Separation und fördern Shrinkage [9, 11]

Outline

Inferenz der Parameter Posteriori

Erinnerung

- Inferenz im frequentistischen LM: z.B. *kleinste Quadrate Schätzung* mit

$$\hat{\boldsymbol{\theta}}_{KQ} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad \text{mit} \quad \hat{\boldsymbol{\theta}}_{KQ} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

- Inferenz im bayesianischen LM: *Bayes-Regel* zur Ermittlung der Parameter Posteriori

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\mathcal{L}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int \mathcal{L}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

Inferenz mit konjugierten Priori

Bayesianisches LM: aus der Konjugiertheit von $\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}$ und $\boldsymbol{\theta}, \sigma^2 \sim \text{NIG}$ ergibt sich

$$\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y} \sim \text{NIG}(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \hat{a}, \hat{b})$$

mit

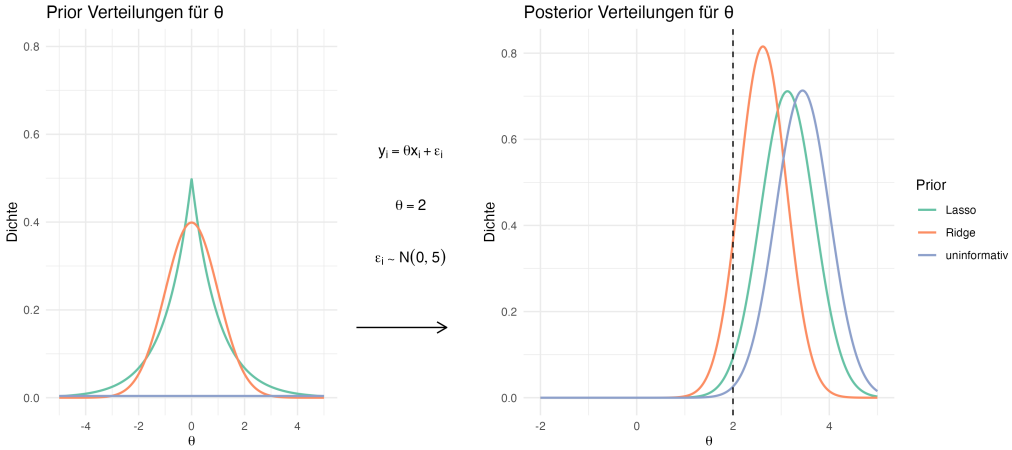
$$\hat{\boldsymbol{\mu}} = \hat{\Sigma}(\breve{\Sigma}^{-1}\breve{\boldsymbol{\mu}} + \boldsymbol{X}^\top \boldsymbol{y}), \quad \hat{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{X} + \breve{\Sigma}^{-1})^{-1}$$

- **uninformative NIG-Priori**: a-Posteriori Mean und KQ-Schätzer sind äquivalent:

$$\breve{\boldsymbol{\mu}} = \boldsymbol{0}, \quad \breve{\Sigma}^{-1} = \boldsymbol{0} \implies \hat{\boldsymbol{\mu}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \quad \hat{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

- **Ridge**: Spezialfall der NIG-Priori → konjugierte Berechnung der Parameter Posteriori
- **Lasso**: Posteriori hat keine geschlossene Form, man kann aber mit z.B. Gibbs-Sampling daraus simulieren [20]

Posteriori Update für Priori-Verteilungen



Approximative Inferenz

Problem: Inferenz mit konjugierten Priori ist nur sehr selten möglich [21]

→ **Approximative bayesianische Inferenz**, z.B. mit

- Sampling Methoden (Markov chain Monte Carlo Methoden)
  - **Metropolis-Hastings Algorithmus** [12]
  - Gibbs Sampling (bei bedingte Konjugiertheit) [4]
  - Hamiltonian Monte Carlo (v.a. hochdimensionale Posteriori) [18]
- Deterministischer Approximation, z.B. **Laplace Approximation** [25]
- U.v.m.

Metropolis Hastings Algorithmus [12]

- **Idee**: Aus der Posteriori $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ ziehen, ohne Annahmen über ihre exakte Form machen zu müssen
- **Problem**: Ergebnisse sind am besten, wenn die Posteriori bis auf eine Konstante (meist die Normalisierungskonstante $\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}$) bekannt sind
- **Inputs**:
  - Anzahl der Ziehungen $K \rightarrow$ frei wählbar
  - Likelihood $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ und Priori $p(\boldsymbol{\theta}) \rightarrow$ bekannt
  - Proposal Verteilung $q \rightarrow$ muss sorgfältig gewählt werden

Metropolis Hastings: Wahl der Proposal Verteilung q

Effizienz des Algorithmus ist stark abhängig von der Proposal Verteilung.

$\rightarrow$ **Optionen**:

- **Normalverteilung**

$$q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)}) \sim \mathcal{N}(\boldsymbol{\theta}^{(k)} \mid -H^{-1}(\boldsymbol{\theta}^{(k)}))$$

  mit Mittelwert beim letzten Sample. Die Hesse-Matrix H von $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ wird meist mit IWLS geschätzt [7, 15, 23]

- Scott [23] schlägt **Verteilungen mit schweren Rändern** vor $\rightarrow$ mehr Mixing, kürzerer Burn-in und schnellere Konvergenz

Laplace Approximation (LA) [25]

- **Idee**: Approximation der Posteriori mit einer Normalverteilung

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}_{MAP}, H^{-1}(\hat{\boldsymbol{\theta}}_{MAP}))$$

  mit Maximum-a-Posteriori Schätzer $\hat{\boldsymbol{\theta}}_{MAP}$ und Hesse-Matrix H von $p(\boldsymbol{\theta} \mid \boldsymbol{y})$
- Erweiterung für hierarchische Modelle: Integrated Nested Laplace Approximation (INLA) [22]

Beispiel: Vergleich von approximativen Inferenz Methoden

Setup: 1.000 synthetische Datensätze mit $n = 100$ und

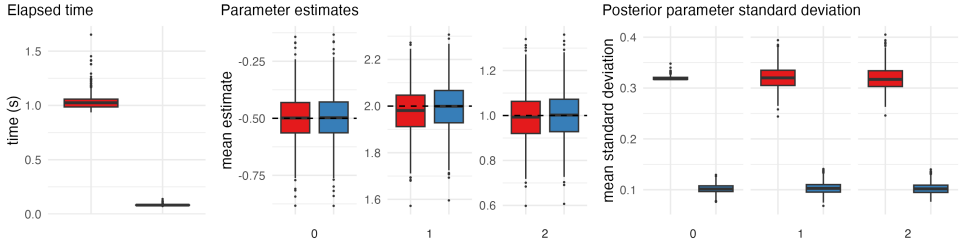
$$\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad \boldsymbol{\theta} = (-0.5, 2, 1)$$
$$\text{linear: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{I})$$
$$\text{logistic: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \text{Ber}(\sigma(\boldsymbol{X}\boldsymbol{\theta}))$$

Experiment: für jeden Datensatz wurde angepasst:

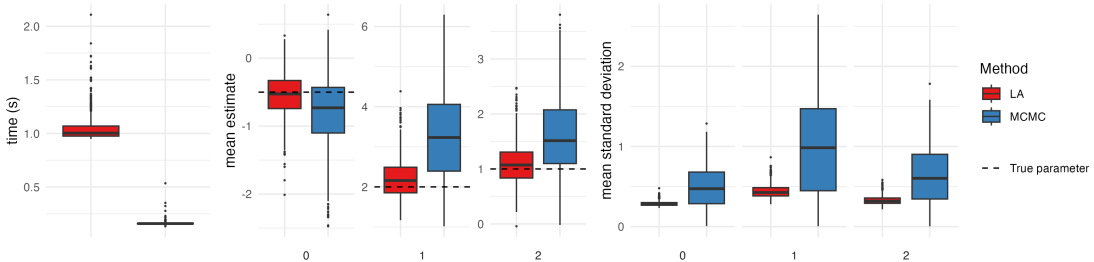
- Ein lineares und ein logistisches Regressionsmodell mit $\boldsymbol{\theta} \sim \mathcal{N}(0, 10 \cdot \boldsymbol{I})$ und $\sigma^2 = 10$
- Mit Laplace Approximation und Metropolis-Hastings ($K = 5.000$, Burn-in $= 500$, Thinning Intervall $= 10$)

Beispiel: LA vs. Metropolis-Hastings

Lineares Modell



Logistisches Modell



Vorhersagen mit bayesianischen Modellen

Aus der Bayes Regel:

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Posterior Predictive Distribution

$\rightarrow$ Vorhersagen für neue Daten $\tilde{\boldsymbol{X}}$ im bayesianischen GLM: [2, 1]

$$p(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \int p(\tilde{\boldsymbol{y}}, \boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta} = \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \overset{\tilde{\boldsymbol{y}} \perp \boldsymbol{y} \mid \boldsymbol{\theta}}{=} \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Für das bayesianische **logistische Modell** berechnet man

$$p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{y})$$

mit $y_i \in \{0 \text{ (negativ)}, 1 \text{ (positiv)}\}$

Berechnung der Posterior Predictive Distribution

Analytische Berechnung: z.B. für die NIG-Priori möglich

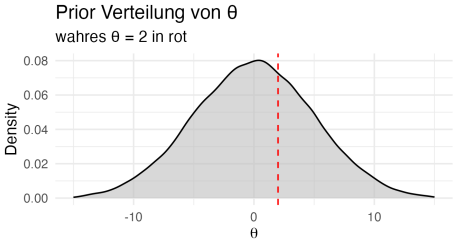
Sonst: **Approximation**

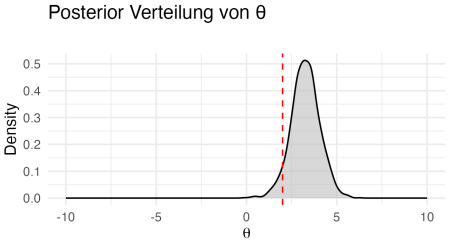
- für Metropolis-Hastings: $p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{y}) \approx \frac{1}{K} \sum_{k=1}^{K} \sigma(\tilde{\boldsymbol{X}} \boldsymbol{\theta}^{(k)})$
- für Laplace Approximation:
  - Samples $\boldsymbol{\theta}^{(s)} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_{MAP}, H^{-1}(\hat{\boldsymbol{\theta}}_{MAP}))$ mit $s = 1, \ldots, S$ aus der LA-approximierten Parameter Posteriori ziehen und wie bei Metropolis-Hastings vorgehen oder
  - LA-approximierte PPD durch Integration analytisch berechnen

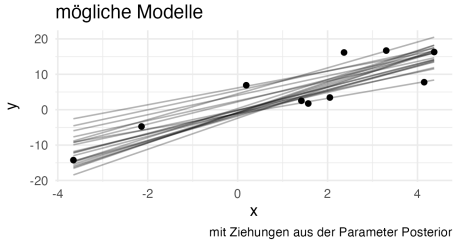
Outline

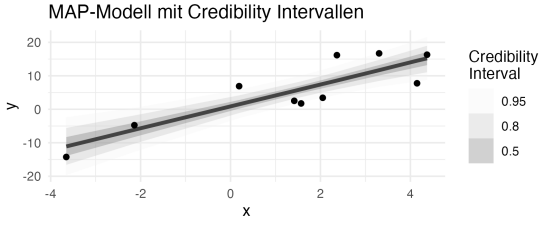
Zusammenfassung anhand eines simplen Beispiels:

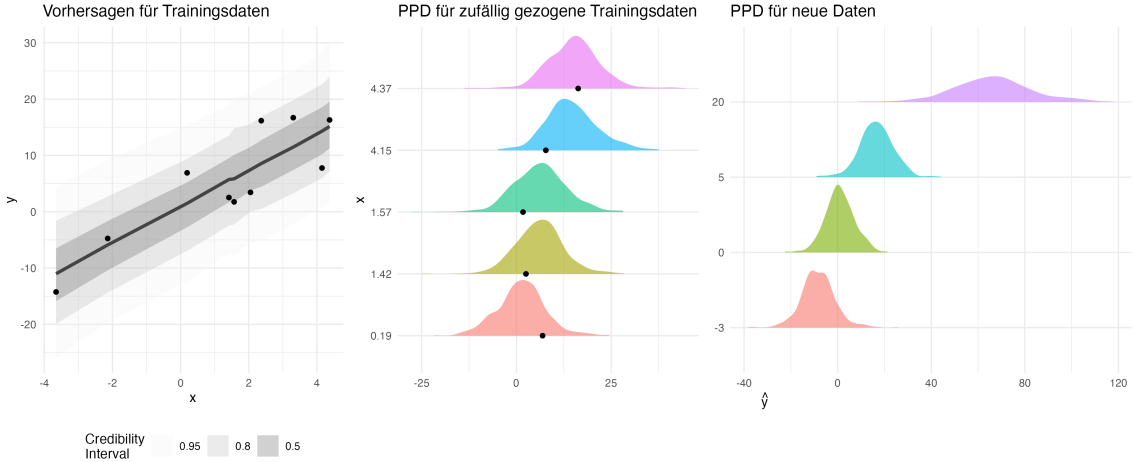
$$y_i = \theta_{true} x_i + \epsilon_i \quad \text{mit} \quad \epsilon_i \sim \mathcal{N}(0, 5) \quad \text{für} \quad i = 1, \ldots, 20$$
$$\theta_{true} = 2, \quad \theta \sim \mathcal{N}(0, 5)$$


Prior Verteilung von θ
wahres θ = 2 in rot
Density
θ


Posterior Verteilung von θ
Density
θ


mögliche Modelle
y
x
mit Ziehungen aus der Parameter Posterior


MAP-Modell mit Credibility Intervallen
y
x
Credibility Interval
0.95
0.8
0.5


Vorhersagen für Trainingsdaten
PPD für zufällig gezogene Trainingsdaten
PPD für neue Daten
y
x
x
4.37
4.15
1.57
1.42
0.19
ŷ
Credibility Interval
0.95
0.8
0.5

Fazit

- Bayesianische Regression ist nicht immer sinnvoller als frequentistische Regression
- Anwendungsmöglichkeiten u.a.:
  - kleine Stichproben
  - explizite Nutzung von Vorwissen
  - einfache Regularisierung + Erhalt der Interpretierbarkeit
  - Verteilungen (statt Konfidenzintervallen) zur Quantifizierung von Unsicherheit
- Inferenz: modernere Methoden notwendig

Literatur Empfehlungen

- Bayesianische Regression (v.a. für praktische Anwendung): Gelman et al. [10]
- Priori Verteilungen (v.a. Shrinkage): Erp, Oberski, and Mulder [5]
- Software:
  - `brms` und `tidybayes` in `R`
  - `PyMC` in `python`

Outline

Outline

Praktische Anwendbarkeit von (bayesianischer) Regularisierung

Vorteile von bayesianischer Regularisierung sind u.a.:

- Probabilistisches Modell trotz Regularisierung
- Regularisierung-Parameter muss nicht als Hyperparameter optimiert werden (z.B. durch Priori auf τ^2)
- Mehr Anpassungsmöglichkeiten durch Priori-Spezifikation

Frequentistische vs. bayesianische Lasso Regularisierung

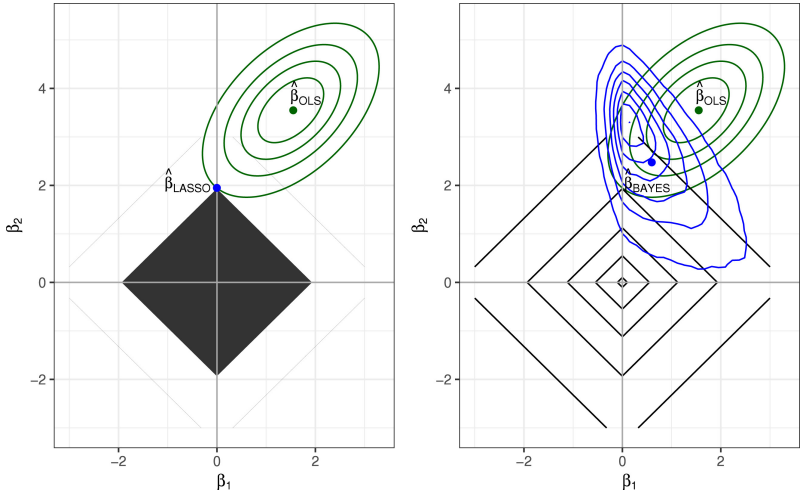


Figure 1: Contour plot representing the sum of squared residuals, classical lasso constraint region (left), bivariate lasso prior and posterior distribution (right), and the classical and Bayesian penalized point estimates. Aus Erp, Oberski, and Mulder [5]

Beispiel: Vergleich von Regularisierungs-Priori in linearer und logistischer Regression

Setup: zwei synthetische Datensätze mit $n = 150$

- **Szenario A**: Sparses Setting ohne Kollinearität oder Überparametrisierung
  - $n = 150,\ n_{train} = 100,\ n_{test} = 50$
  - $\boldsymbol{\theta} = (2, 1.5, 0, 0, 0)$
  - $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  - linear: $\boldsymbol{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{I})$, logistisch: $\boldsymbol{y} \mid \boldsymbol{\theta} \sim \text{Ber}(\sigma(\boldsymbol{X}\boldsymbol{\theta}))$
- **Szenario B**: uninformatives Setting mit $n \approx p$ und Kollinearität zwischen informativen und uniformativen Koeffizienten
  - $n = 150,\ n_{train} = 30,\ n_{test} = 120$
  - $\boldsymbol{\theta} = (2, 1.5, 0, \overset{26 \text{ times}}{\ldots}, 0)$
  - $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & & & \\ & S_3 & & 0 \\ & & I_{26} & \\ & 0 & & \end{pmatrix}, \qquad S_3 = \begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}$
  - linear: $\boldsymbol{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{I})$, logistisch: $\boldsymbol{y} \mid \boldsymbol{\theta} \sim \text{Ber}(\sigma(\boldsymbol{X}\boldsymbol{\theta}))$

Experiment: Für jeden Datensatz wurde angepasst:

- bayesianisches lineares und logistisches Modell (MCMC mit $K = 20000$, Burn-in $= 1000$, Thinning $= 10$) mit
- uninformative, Ridge und Lasso Priori

Ergebnis: Regularisierung in bayesianischen Modellen (I)

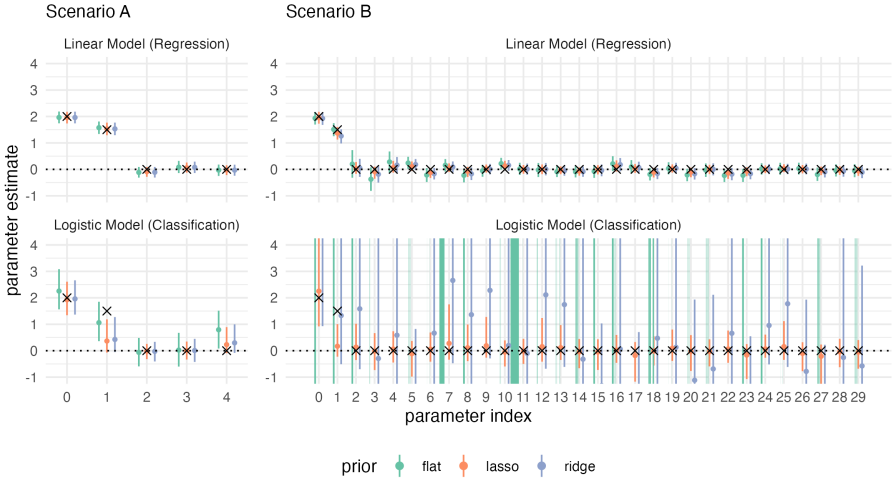
Evaluation

- Anzahl korrekt als einflussreich identifizierter Kovariablen (Hits)
- Anzahl fälschlich als einflussreich identifizierter Kovariablen (FP)
- Mean log posterior predictive density (MLPPD) [8]

Ergebnis

Model	Prior	Scenario A			Scenario B		
		Hits (of 2)	FP (of 3)	MLPPD	Hits (of 2)	FP (of 28)	MLPPD
Linear	flat	2	0	-1.425	2	1	-1.605
Linear	LASSO	2	0	-1.424	2	0	-1.464
Linear	ridge	2	0	-1.427	2	0	-1.575
Logit	flat	2	1	-0.390	0	21	$-\infty$
Logit	LASSO	1	0	-0.463	1	0	-0.485
Logit	ridge	1	0	-0.455	1	0	-0.493

Ergebnis: Regularisierung in bayesianischen Modellen (II)



Posterior Inference: Verteilung von $\boldsymbol{\theta}$

- Im frequentistischen Modell:

$$\hat{\boldsymbol{\theta}}_{KQ} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

- Im bayesianischen Modell (NIG-Prior):

$$\boldsymbol{\theta} \sim \mathcal{T}(2\hat{a}, \hat{\boldsymbol{\mu}}, \hat{b}/\hat{a}\hat{\Sigma}) \approx \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{b}/\hat{a}\hat{\Sigma})$$

für großes p

Posterior Inference: Der Metropolis (-Hastings) Algorithmus

1. $\boldsymbol{\theta}^{(1)}$ initialisieren
2. Für $k = 1, \ldots, K$
   1. $\boldsymbol{\theta}^{(*)}$ aus der *Proposal Verteilung* $q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)})$ ziehen
   2. *Akzeptanzwahrscheinlichkeit* berechnen

$$\alpha = \min\Big(1, \frac{p(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{y})\; p(\boldsymbol{\theta}^{(*)})\; q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(*)})}{p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})\; p(\boldsymbol{\theta}^{(k)})\; q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)})}\Big)$$

   3. Vorschlag $\boldsymbol{\theta}^{(*)}$ akzeptieren oder verwerfen (für $u \sim \text{Uni}[0, 1]$)

$$\begin{cases} u \leq \alpha & \boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(*)} \\ u > \alpha & \boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} \end{cases}$$

Und z.B.

$$q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)}) \sim \mathcal{N}(\boldsymbol{\theta}^{(k)} \mid -H^{-1}(\boldsymbol{\theta}^{(k)}))$$

mit $H(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \log\Big(p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})\; p(\boldsymbol{\theta}^{(k)})\Big)$

Posterior Inference: Mittelwert und Varianz für Laplace Approxima

Exemplarische Berechnung für ein bayesianisches logistisches Modell mit der Prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$:

Es gilt dass mit Laplace Approximation $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}_{MAP}, H^{-1}(\hat{\boldsymbol{\theta}}_{MAP}))$ mit

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{MAP} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \overset{\text{Bayes' rule}}{=} \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\Big(\sigma(y_i \, \boldsymbol{x}_i \boldsymbol{\theta})\Big) - \frac{1}{2\sigma^2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \\
H(\boldsymbol{\theta}) &= -\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{1}{\sigma^2} \boldsymbol{I} + \sum_{i=1}^{n} \sigma(y_i \, \boldsymbol{x}_i \boldsymbol{\theta}) \Big(1 - \sigma(y_i \, \boldsymbol{x}_i \boldsymbol{\theta})\Big) \boldsymbol{x}_i \boldsymbol{x}_i^\top .
\end{aligned}$$

[1] Maria Maddalena Barbieri. “Posterior Predictive Distribution”. In: *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 2015, pp. 1–6. ISBN: 978-1-118-44511-2. DOI: 10.1002/9781118445112.stat07839.

[2] George E. P. Box. “Sampling and Bayes’ Inference in Scientific Modelling and Robustness”. In: *Journal of the Royal Statistical Society. Series A (General)* 143.4 (1980), pp. 383–430. ISSN: 0035-9238. DOI: 10.2307/2982063.

[3] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. “The horseshoe estimator for sparse signals”. English. In: *BIOMETRIKA* 97.2 (June 2010), pp. 465–480. ISSN: 0006-3444. DOI: 10.1093/biomet/asq017.

[4] P. Dellaportas and A. F. M. Smith. “Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 42.3 (1993), pp. 443–459. ISSN: 0035-9254. DOI: 10.2307/2986324.

[5] Sara van Erp, Daniel L. Oberski, and Joris Mulder. “Shrinkage priors for Bayesian penalized regression”. In: *Journal of Mathematical Psychology* 89 (Apr. 2019), pp. 31–50. ISSN: 0022-2496. DOI: 10.1016/j.jmp.2018.12.004.

[6] Ludwig Fahrmeir, Thomas Kneib, and Susanne Konrath. "Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection". In: *Statistics and Computing* 20.2 (2010), pp. 203–219. ISSN: 1573-1375. DOI: 10.1007/s11222-009-9158-3.

[7] Dani Gamerman. "Markov chain Monte Carlo for dynamic generalised linear models". In: *Biometrika* 85.1 (Mar. 1998), pp. 215–227. ISSN: 0006-3444. DOI: 10.1093/biomet/85.1.215.

[8] Andrew Gelman, Jessica Hwang, and Aki Vehtari. *Understanding predictive information criteria for Bayesian models*. July 2013. DOI: 10.48550/arXiv.1307.5928.

[9] Andrew Gelman et al. "A weakly informative default prior distribution for logistic and other regression models". In: *The Annals of Applied Statistics* 2.4 (Dec. 2008), pp. 1360–1383. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/08-AOAS191.

[10] Andrew Gelman et al. *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC, 2013. ISBN: 978-0-429-11307-9. DOI: 10.1201/b16018.

[11] Joyee Ghosh, Yingbo Li, and Robin Mitra. *On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression*. Feb. 2017. DOI: 10.48550/arXiv.1507.07170.

[12] W. K. Hastings. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1 (1970), pp. 97–109. ISSN: 0006-3444. DOI: 10.2307/2334940.

[13] Arthur E. Hoerl and Robert W. Kennard. “Ridge Regression: Applications to Nonorthogonal Problems”. In: *Technometrics* 12.1 (1970), pp. 69–82. ISSN: 0040-1706. DOI: 10.2307/1267352.

[14] Arthur E. Hoerl and Robert W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* 12.1 (1970), pp. 55–67. ISSN: 0040-1706. DOI: 10.2307/1267351.

[15] Peter J. Lenk and Wayne S. DeSarbo. “Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects”. en. In: *Psychometrika* 65.1 (Mar. 2000), pp. 93–119. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02294188.

[16] David J. C. MacKay. “Bayesian Interpolation”. In: *Neural Computation* 4.3 (May 1992), pp. 415–447. ISSN: 0899-7667. DOI: 10.1162/neco.1992.4.3.415.

[17] Tj Mitchell and Jj Beauchamp. “Bayesian Variable Selection in Linear-Regression”. English. In: *Journal of the American Statistical Association* 83.404 (Dec. 1988), pp. 1023–1032. ISSN: 0162-1459. DOI: 10.2307/2290129.

[18] Radford M Neal. “Probabilistic inference using Markov chain Monte Carlo methods”. In: *Department of Computer Science, University of Toronto Toronto, Ontario, Canada* (1993).

[19] R. B. O’Hara and M. J. Sillanpää. “A review of Bayesian variable selection methods: what, how and which”. In: *Bayesian Analysis* 4.1 (Mar. 2009), pp. 85–117. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/09-BA403.

[20] Trevor Park and George Casella. “The Bayesian Lasso”. English. In: *Journal of the American Statistical Association* 103.482 (June 2008), pp. 681–686. ISSN: 0162-1459. DOI: 10.1198/016214508000000337.

[21] Nicholas G. Polson, Scott James G., and Jesse Windle. “Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables”. In: *Journal of the American Statistical Association* 108.504 (2013), pp. 1339–1349. ISSN: 0162-1459. DOI: 10.1080/01621459.2013.829001.

[22] Håvard Rue, Sara Martino, and Nicolas Chopin. “Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.2 (Apr. 2009), pp. 319–392. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2008.00700.x.

[23] Steven L. Scott. “Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models”. en. In: *Statistical Papers* 52.1 (Feb. 2011), pp. 87–109. ISSN: 1613-9798. DOI: 10.1007/s00362-009-0205-0.

[24] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 0035-9246.

[25] Luke Tierney and Joseph B. Kadane. “Accurate Approximations for Posterior Moments and Marginal Densities”. In: *Journal of the American Statistical Association* 81.393 (1986), pp. 82–86. ISSN: 0162-1459. DOI: 10.1080/01621459.1986.10478240.