

Gliederungsbeispiel

**Robustheit in der Regressionsanalyse: Ein
systematischer Vergleich von OLS, Ridge, Lasso und
bayesianischen Modellen unter Verletzung der
Modellannahmen**

Masterarbeit

Fachbereich: Data Science / Statistik

Vorgelegt von:

[Vorname, Nachname]

26.06.2026

Problemstellung & Relevanz

Problemstellung: Die klassische Kleinste-Quadrate-Schätzung (OLS) gilt unter den Gauss-Markov-Annahmen als bester linearer erwartungstreuer Schätzer (BLUE). In der statistischen Praxis führen jedoch Ausreißer, Multikollinearität und Heteroskedastizität häufig zu einer massiven Verzerrung der Ergebnisse und instabilen Koeffizienten, da OLS einen Breakdown-Point von Null besitzt (Rousseeuw, 1997). Während frequentistische Regularisierungsverfahren wie Ridge und Lasso Lösungen für Multikollinearität bieten (Hoerl & Kennard, 1970; Zou, 2006), versprechen bayesianische Ansätze durch die Wahl robuster Priors eine stabilere Modellierung von Unsicherheit (Park & Casella, 2008).

Wissenschaftliche Relevanz: Trotz der theoretischen Vorteile robuster und bayesianischer Verfahren fehlt oft ein direkter, simulationsbasierter Vergleich der Performance dieser Modellklassen unter simultaner Verletzung mehrerer Annahmen (Maronna et al., 2006).

Praktische Relevanz: In der Data Science ist die Wahl des richtigen Modells bei unsauberen Datensätzen kritisch für die Vorhersagegüte und die Validität statistischer Inferenz, insbesondere wenn keine realen Parameter für die Validierung vorliegen.

Forschungsfrage und Forschungsziel

Zentrale Forschungsfrage: Inwieweit weisen regularisierte frequentistische und bayesianische Regressionsmodelle eine überlegene Robustheit gegenüber der klassischen OLS-Regression auf, wenn Ausreißer, Heteroskedastizität und Multikollinearität in verschiedenen Intensitäten auftreten?

Unterfragen:

- Welchen Einfluss hat der Anteil von Ausreißern auf den Mean Squared Error (MSE) der Koeffizientenschätzungen bei OLS im Vergleich zu Ridge und Lasso?
- Inwiefern stabilisieren bayesianische Priors (z.B. Laplace-Prior) die Schätzung bei hoher Multikollinearität im Vergleich zum frequentistischen Lasso?
- Wie verhält sich die Vorhersagegenauigkeit der Modelle bei Vorliegen von Heteroskedastizität?

Hypothesen:

- **H1:** Regularisierte Verfahren (Ridge, Lasso) zeigen bei hoher Multikollinearität eine geringere Varianz der Schätzer als OLS, erkaufte durch einen geringen Bias (Hoerl & Kennard, 1970).
- **H2:** Bayesianische Regressionsmodelle mit schwerfälligen Priors erzielen bei hohen Ausreißeranteilen stabilere Ergebnisse als klassische OLS-Modelle (West, 1984).

Forschungsziel und Erkenntnisinteresse

Forschungslücke: Bestehende Studien vergleichen oft nur zwei Modellklassen (z.B. OLS vs. Lasso) oder konzentrieren sich auf eine einzelne Annahmeverletzung (z.B. nur Multikollinearität). Ein systematischer 'Stress-Test', der OLS, Ridge, Lasso und bayesianische Modelle unter kombiniertem Einfluss von Ausreißern, Heteroskedastizität und Multikollinearität vergleicht, ist in der aktuellen Literatur unterrepräsentiert.

Beitrag dieser Arbeit: Diese Arbeit schließt diese Lücke durch ein einheitliches Simulationsdesign, das die Grenzen der Robustheit für alle vier Modellklassen simultan aufzeigt.

Implikationen & Ausblick: Die Ergebnisse liefern eine Entscheidungsgrundlage für die Modellwahl in Data-Science-Pipelines und dienen als Ausgangspunkt für die Entwicklung hybrider Schätzer, die sowohl bayesianische als auch regularisierende Vorteile vereinen.

Forschungsstand und theoretische Grundlagen

Forschungsstand - Klassische Robustheit: Die Literatur etabliert M-Schätzer und LTS als Standard für den Umgang mit Ausreißern (Hampel, 2001; Rousseeuw, 1997). Maronna (2011) zeigt, dass MM-Schätzer hohe Effizienz mit Robustheit verbinden können.

Forschungsstand - Regularisierung: Es besteht Konsens, dass Lasso die Variablenselektion optimiert, während Ridge bei Multikollinearität die Varianz reduziert (Tibshirani, 1996; Hoerl & Kennard, 1970). Zou (2006) erweiterte dies um das Adaptive Lasso zur Korrektur von Selektionsbias.

Forschungsstand - Bayesianische Ansätze: Das Bayesian Lasso (Park & Casella, 2008) integriert die L1-Strafe in einen Wahrscheinlichkeitsrahmen und bietet natürliche Unsicherheitsintervalle. West (1984) und Mitchell & Beauchamp (1988) wiesen die Vorteile von Spike-and-Slab-Priors für die Variablenselektion nach.

Kontroversen: Es wird debattiert, ob bayesianische Modelle bei sehr großen Datensätzen (Large p, Small n) den rechentechnisch effizienteren frequentistischen Modellen überlegen sind (Tew et al., 2023).

Theoretischer Rahmen und Begriffsklärungen

Theoretischer Rahmen: Die Arbeit basiert auf dem Gauss-Markov-Theorem und dessen modernen Erweiterungen (Hansen, 2022). Ergänzend wird die statistische Entscheidungstheorie (Fourdrinier et al., 2018) herangezogen, um den Bias-Varianz-Tradeoff bei Shrinkage-Schätzern zu bewerten.

Zentrale Konzepte:

- **Breakdown Point:** Das Maß für die Robustheit eines Schätzers gegenüber kontaminierten Daten (Hampel, 2001).
- **Shrinkage:** Die bewusste Einführung von Bias zur Reduktion der Varianz (Hoerl & Kennard, 1970).
- **Posterior-Verteilung:** Die Verteilung der Parameter nach Einbezug von Daten und Prior-Information (Park & Casella, 2008).
- **Heteroskedastizität:** Nicht-konstante Varianz der Fehlerterme, die OLS ineffizient macht (Lakshmi et al., 2021).

Forschungskonzept/Methodisches Vorgehen

Art der Ergebnisse: Die Arbeit liefert einen quantitativen Vergleich der Schätzgüte (MSE, Bias, Varianz) und der Vorhersagepräzision (RMSE) über verschiedene Simulationsszenarien hinweg.

Argumentationslinie:

- **Schritt 1:** Theoretische Herleitung der Schwächen von OLS bei Annahmeverletzungen basierend auf der Literatur (Rousseeuw, 1997).
- **Schritt 2:** Durchführung von Simulationen, die zeigen, dass Ridge und Lasso bei Multikollinearität stabilere Koeffizienten liefern als OLS.
- **Schritt 3:** Nachweis, dass das Bayesian Lasso durch die Posterior-Verteilung eine robustere Unsicherheitsquantifizierung bei Ausreißern ermöglicht.

Erwarteter Erkenntnisgewinn: Es wird erwartet, dass bayesianische Modelle insbesondere in Szenarien mit geringen Stichproben und hohen Ausreißerraten den frequentistischen Ansätzen überlegen sind.

Eingrenzung und Umfang

Untersuchungsrahmen: Die Arbeit konzentriert sich auf lineare Regressionsmodelle für stetige Zielvariablen. Untersucht werden OLS, Ridge-Regression (L2), Lasso-Regression (L1) und das Bayesian Lasso. Die Evaluation erfolgt ausschließlich über Monte-Carlo-Simulationen mit kontrollierten Parametern (n , p , SNR, Korrelationsstruktur).

Ausschlüsse: Nicht-lineare Modelle (z.B. Random Forests, neuronale Netze) werden nicht berücksichtigt. Ebenso werden keine realen Datensätze verwendet, um die Vergleichbarkeit gegenüber den 'wahren' Parametern der Simulation zu gewährleisten. Zeitreihenanalytische Besonderheiten (Autokorrelation) werden zugunsten der klassischen Querschnittsannahmen vernachlässigt.

Vorläufige Gliederung

1. Einleitung (ca. 5 Seiten)
 - Problemstellung und Zielsetzung
 - Aufbau der Arbeit
2. Theoretische Grundlagen der Regressionsanalyse (ca. 12 Seiten)
 - 2.1 Das klassische lineare Modell und OLS
 - 2.2 Gauss-Markov-Theorem und BLUE-Eigenschaft
 - 2.3 Verletzungen der Modellannahmen: Konsequenzen für OLS
3. Regularisierungsverfahren in der frequentistischen Statistik (ca. 10 Seiten)
 - 3.1 Ridge-Regression: L2-Penalty und Multikollinearität
 - 3.2 Lasso-Regression: L1-Penalty und Variablenselektion
 - 3.3 Der Bias-Varianz-Tradeoff bei Shrinkage-Methoden
4. Bayesianische Regressionsanalyse (ca. 12 Seiten)
 - 4.1 Bayesianische Inferenz und Prior-Spezifikationen
 - 4.2 Das Bayesian Lasso: Hierarchische Modellierung
 - 4.3 MCMC-Verfahren zur Posterior-Schätzung

- 5. Konzepte der robusten Statistik (ca. 8 Seiten)
 - 5.1 Breakdown Point und Einflussfunktionen
 - 5.2 Robuste Verlustfunktionen und M-Schätzer
- 6. Methodik der Simulationsstudie (ca. 10 Seiten)
 - 6.1 Design der Monte-Carlo-Simulation
 - 6.2 Definition der Störszenarien und Metriken
 - 6.3 Implementierung in R (brms, glmnet)
- 7. Ergebnisse und Diskussion (ca. 18 Seiten)
 - 7.1 Performanzvergleich unter Multikollinearität
 - 7.2 Robustheit gegenüber Ausreißern und Heteroskedastizität
 - 7.3 Diskussion der praktischen Implikationen
- 8. Fazit und Ausblick (ca. 5 Seiten)