

Regularisierung (Shrinkage Methoden)

Steve Finger

Inhalt

1. Motivation
2. Regularisierung
3. Vergleich der Shrinkage Methoden
4. Zusammenfassung

1. Motivation

1. Kleinste Quadrate Methode

Lineare Regression:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

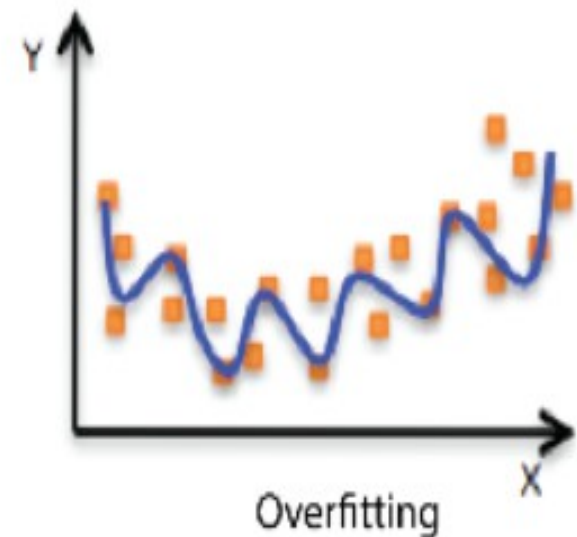
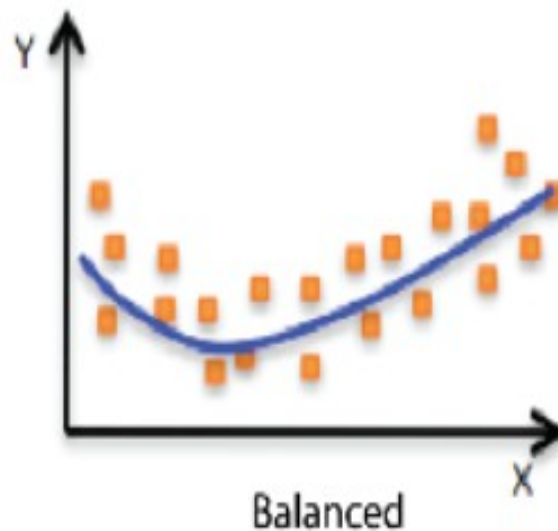
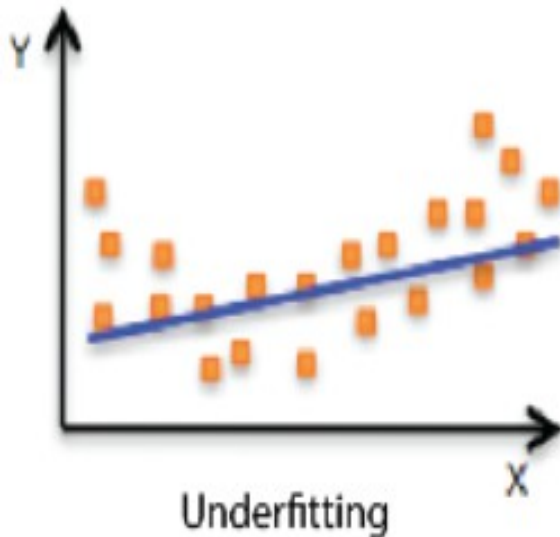
Minimiere:

$$\text{RSS} = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

1. Probleme der KQM

Probleme:

1. Die Interpretierbarkeit des Modells
2. Die Prognosefähigkeit des Modells



1. Alternativen zur KQM

Alternativen:

1. Subset Selection (Best Subset Selection)
2. Shrinkage Methoden (Regularisierung)
3. Weitere Methoden

Wollen die Interpretierbarkeit und
Prognosefähigkeit unseres Modells verbessern.

2. Regularisierung

2.1 Ridge Regression

Minimiere:

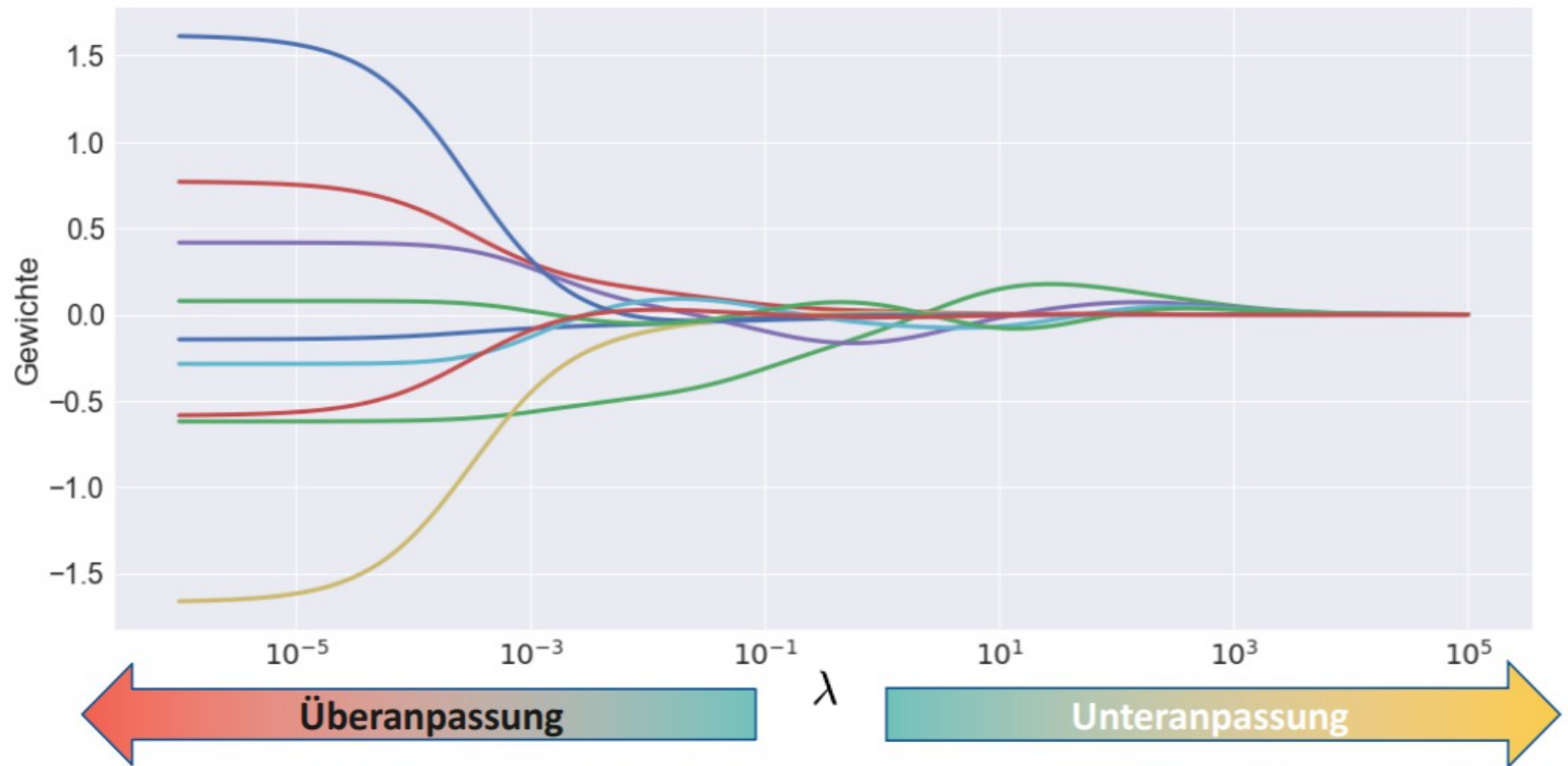
$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$

- λ = Tuning-Parameter
- Zweiter Term wird Regularisierungs- oder Strafterm genannt.
- Wahl von λ ist entscheidend!!!

2.1 Tuning Parameter λ

Wahl von λ ist entscheidend!!!

Effekt der L2-Regularisierung auf die Gewichte/Parameter



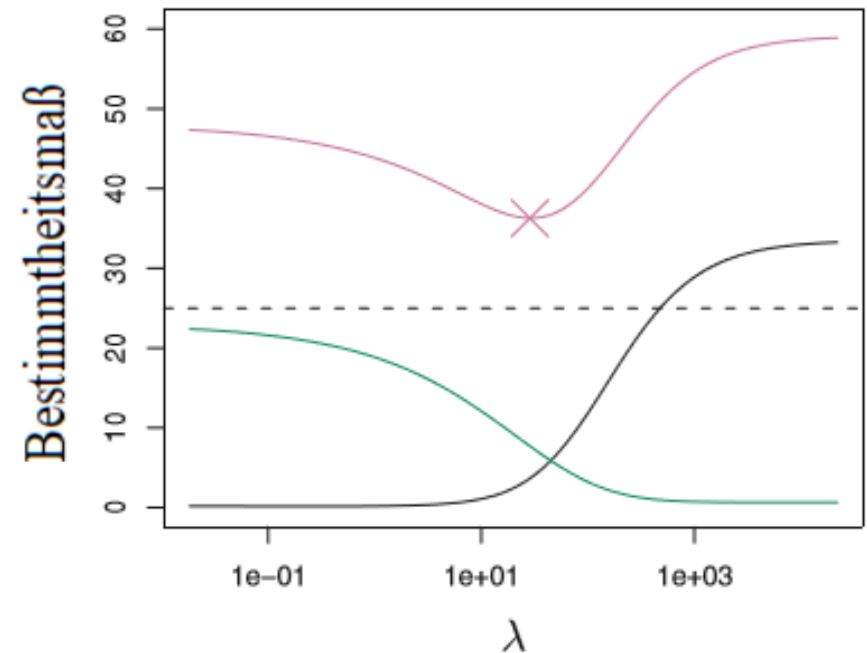
2.1 Ridge Regression

- Ist nicht skalierungsinvariant im Vergleich zur KQM!!!
- Daher müssen vor einer Skalierung die Einflussgrößen standardisiert werden, durch:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

2.1 Vergleich zur KQM

- Ridge Regression profitiert vom Verzerrungs-Varianz-Dilemma
- Funktioniert auch im Fall $p < n$.



2.2 Lasso Regression

Minimiere:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|$$

- Kann Koeffizienten sogar direkt auf 0 schrumpfen (ausselektieren)!
- Wieso? Siehe Kapitel 3.
- Wahl von λ entscheidend!!! (Kreuzvalidierung)

2.3 Kreuzvalidierung

Motivation:

- Vergleich der Prognosegüte von Regressionsmodellen

Idee:

- Aufspalten eines Datensatzes in einen Trainings- und Testdatensatz
- Anpassung des Modells mit dem Trainingsdatensatz
- Vorhersage mit Testdatensatz

2.3 k-fache Kreuzvalidierung

Wozu?

Um die Prognosegüte noch besser als mit der einfachen Kreuzvalidierung zu beurteilen.

Idee:

- Im Prinzip wie einfache Kreuzvalidierung nur mit k Teildatensätze.

2.3 k-fache Kreuzvalidierung

Kreuzvalidierungsfehler:

$$\eta_k = \frac{1}{N} \sum_{i=1}^N \left(\tilde{y}_{k,i} - y_{k,i} \right)^2$$

$$\bar{\eta} = \frac{1}{K} \sum_{k=1}^K \eta_k = \frac{1}{K} \sum_{k=1}^K \frac{1}{N_k} \sum_{i=1}^N \left(\tilde{y}_{k,i} - y_{k,i} \right)^2$$

Wähle λ so, wo der Kreuzvalidierungsfehler am kleinsten ist!!!

3. Vergleich Lasso- und Ridge Regression

3.1 Ein Spezialfall

- Sei $n = p$.
- Bestimme Parameter der KQM, Ridge- und Lasso Regression.
- Zur Vereinfachung ignoriere β_0 .

3.1 Ein Spezialfall

KQM: $\hat{\beta}_j = y_j$

Ridge Regression: $\hat{\beta}_j^R = y_j / (1 + \lambda)$

Lasso Regression:

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases}$$

3.1 Ein Spezialfall

Fazit:

- Ridge Regression schrumpft Koeffizienten um relativen Wert.
- Lasso Regression schrumpft Koeffizienten um absoluten Wert.
- Im Allgemeinen wesentlich komplizierter.

3.2 Allgemeine Betrachtung

Ridge Regression:

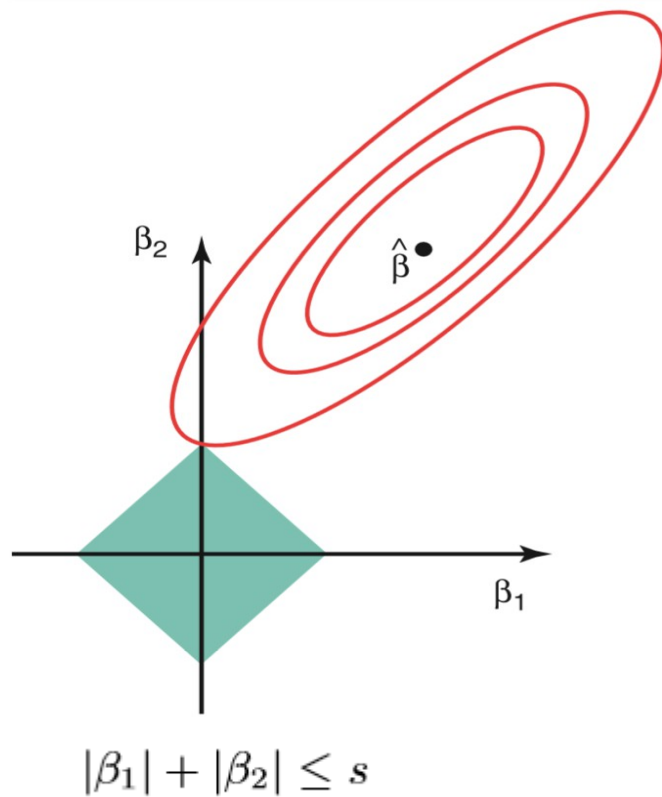
$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^p \beta_j^2 \leq s$$

Lasso Regression:

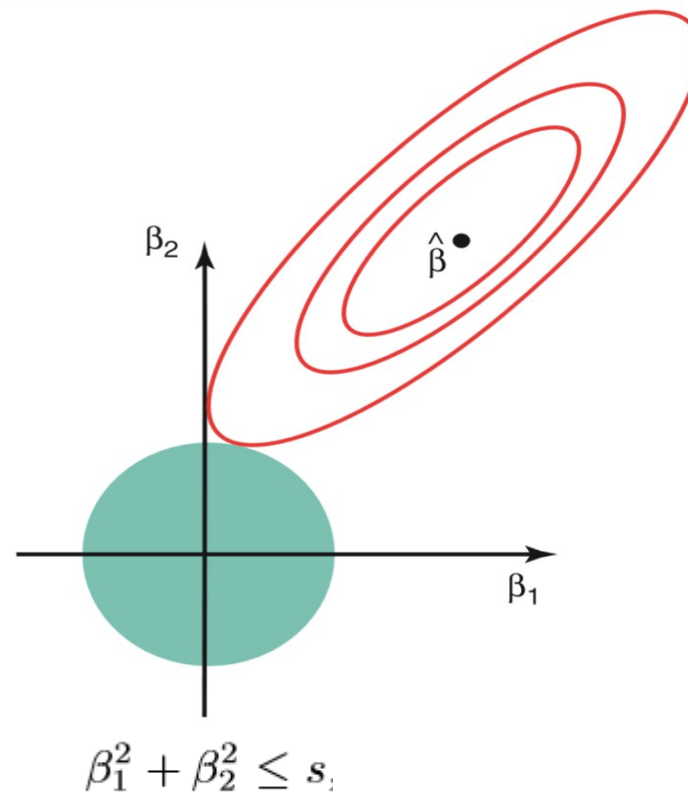
$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s$$

3.2 Allgemeine Betrachtung

Lasso Regression:

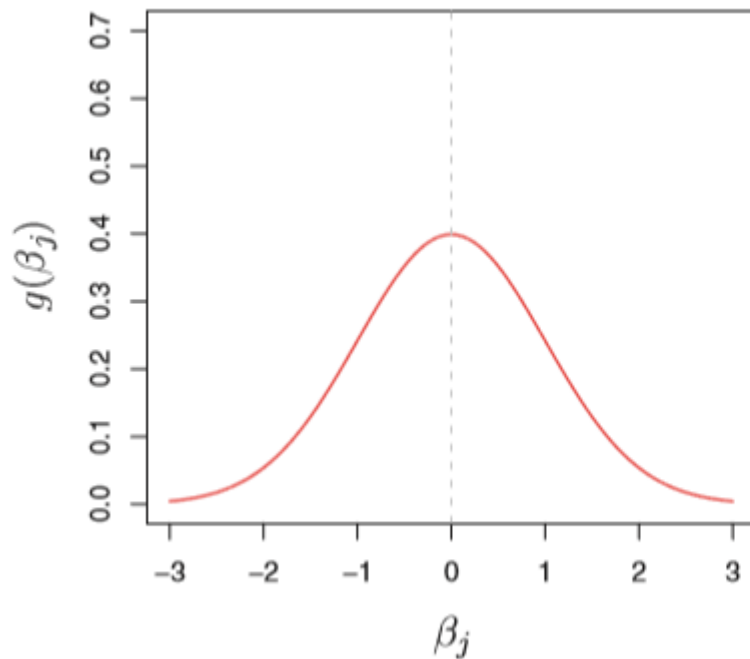


Ridge Regression:

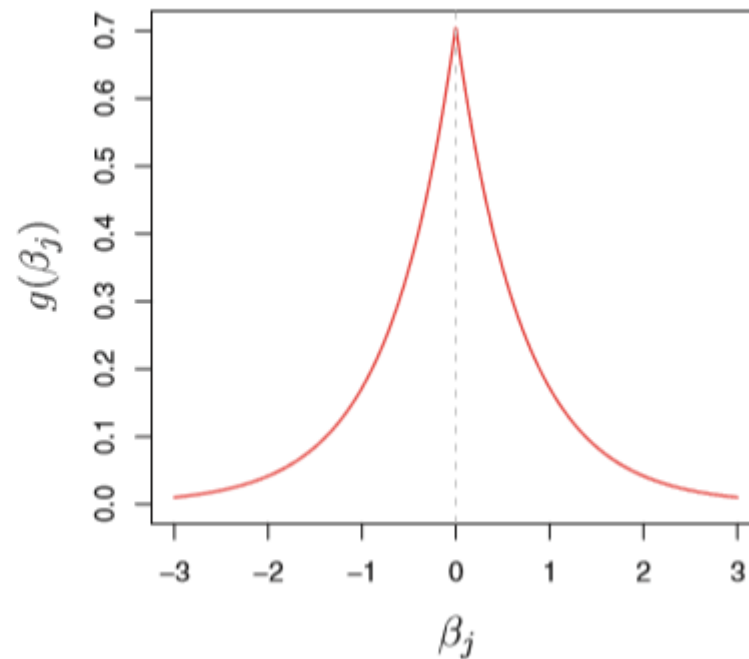


3.3 Wahrscheinlichkeitstheoretische Betrachtung

Ridge Regression:



Lasso Regression:



- Normal verteilt

- Laplace verteilt

3.4 Vergleich zur Best Subset Selection

- Was ist die Best Subset Selection?

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^p I(\beta_j \neq 0) \leq s$$

- Ridge- und Lasso Regression sind weniger rechenaufwendige Alternativen!

3.5 Vergleich

Ridge Regression:

- Schrumpft Koeffizienten nur gegen Null!
- Normal verteilt
- „Koeffizienten werden um absoluten Wert verringert“

Lasso Regression:

- Schrumpft Koeffizienten auch auf Null!
- Liefert besser interpretierbares Modell
- Laplace verteilt
- „Koeffizienten werden um relativen Wert verringert“

4. Zusammenfassung

4. Zusammenfassung

Fazit:

- Funktionieren auch für $n < p$.
- Verringern die Varianz der Parameter.
- Beide Modelle verbessern Prognosegüte und sind weniger rechenaufwendigere Alternativen zu anderen Verfahren!
- Erhöhen Prognosegüte und Interpretierbarkeit
- Jedoch keine Methode ist besser, hängt vom Setting ab und lässt sich mit Kreuzvalidierung herausstellen!