# Kurze Einführung in Regularisierung und penalisierte Schätzansätze in der Statistik

Sarah Brockhaus
nach einer Vorlage von David Rügamer

13. November 2015

# Regularisierung in der Statistik

## Problemstellung

KQ-Kriterium (KQ, Kleinste Quadrate):

$$\mathrm{KQ}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \to \min_{\boldsymbol{\beta}}$$

KQ-Schätzung basiert auf Lösung des Gleichungssystems

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$$

Voraussetzung: $\boldsymbol{X}$ besitzt vollen Rang $p \Leftrightarrow \boldsymbol{X}'\boldsymbol{X}$ ist invertierbar

Problematisch:

- Korrelierte Einflussgrößen $\Rightarrow$ nahezu kollineare Spalten in $\boldsymbol{X}$
- $p \to \infty$ bzw. $p > n$

$\Rightarrow$ Theoretische Lösung existiert, Optimierung jedoch numerisch instabil

## Statistische Regularisierung

Im Allgemeinen meist Erweiterung des KQ-Kriteriums durch einen Strafterm

$$\mathrm{PLS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \cdot \mathrm{pen}(\boldsymbol{\beta}) \to \min_{\boldsymbol{\beta}}$$

zum sogenannten *penalisierten KQ-Kriterium* (*Penalized Least Squares*).

Der Bestrafungsterm $\mathrm{pen}(\boldsymbol{\beta})$ misst die Komplexität der Regressionskoeffizienten während der *Glättungsparameter* $\lambda \geq 0$ dessen Einfluss steuert:

- $\lambda \to 0$: Keine / kaum Bestrafung; es ergibt sich wieder das KQ-Kriterium
- $\lambda \to \infty$: Die Penalisierung besitzt einen großen Einfluss

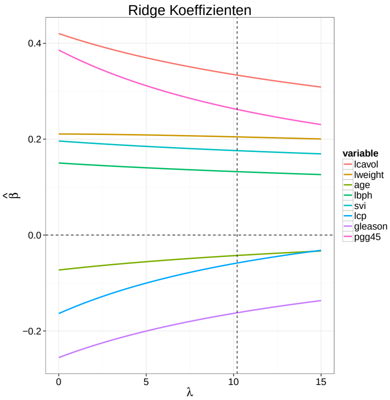
## Bekannte Regularisierungsansätze

- Ridge Regression (engl. ridge, dt. First, Grat, Bergrücken):
  - $\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{k} \beta_j^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$
  - Schrumpft Koeffizienten gegen Null
  - Vorteil: falls $\lambda$ groß $\Rightarrow \boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p$ invertierbar, auch wenn $\boldsymbol{X}'\boldsymbol{X}$ keinen vollen Rang besitzt.
  - Ridge-Schätzer besitzt kleinere (Ko-)Varianz als KQ-Schätzer für $\lambda > 0$, ist aber verzerrt
    $\rightarrow$ insbesondere für hochdimensionales oder annährend kollineares $\boldsymbol{X}$ wird MSE von Varianz stärker beeinflusst als von Bias, was zu einem geringerem MSE des Ridge-Schätzer führen kann

- LASSO (Least Absolute Shrinkage and Selection Operator):
  - $\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{k} |\beta_j|$
  - Schrumpft Koeffizienten nicht nur gegen sondern auch auf Null und kann damit zur Konstruktion sparsamer Modelle verwendet werden

Wichtig: Regularisierungsverfahren basieren auf der Annahme, dass alle Koeffizienten hinsichtlich ihrer absoluten Werte vergleichbar sind.
$\rightarrow$ Kovariablen standardisieren!

# Anwendungsbeispiel




### Vergleich KQ- und Ridge-Schätzung

Simulationsdaten auf Basis des Prostata Datensatzes von Hastie, Tibshirani und Friedman (2009)

- Modellierung der logarithm. Menge an Prostata-spezifischem Abwehrstoff
- in Abhängigkeit von Einflußgrößen wie dem Volumen des Krebses (`lcavol`) oder dem Alter des Patienten (`age`).

Beispielhafter Vergleich der MSEs des Koeffizienten von `lcavol`, mit $\beta$=0.5643:

| | KQ | Ridge |
|---|---|---|
| $\hat{\beta}$ | 0.4204 | 0.3336 |
| MSE | 0.0135 | 0.0070 |

# Weiterführende Verwendung penalisierter Schätzverfahren

Penalisierte Schätzansätze finden in der Statistik insbesondere Verwendung in

- gemischten Modellen:

  Beispiel: Schätzung der Regressionskoeffizienten im linearen gemischten Modell

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon$$
$$\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G}), \varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R})$$

  über Minimierung des PLS-Kriteriums

$$\text{PLS}(\beta, \boldsymbol{b}) = (\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) + \boldsymbol{b}'\boldsymbol{G}^{-1}\boldsymbol{b}$$

Penalisierte Schätzansätze finden in der Statistik insbesondere Verwendung in

- nicht-parametrischer Regression:

  Beispiel: Schätzung eines glatten Effekts mit TP-Basis vom Grad $\ell$

$$y_i = f(z_i) + \varepsilon_i = \gamma_1 + \gamma_2 z_i + \ldots + \gamma_{\ell+1} z_i^\ell + \\ \gamma_{\ell+2}(z_i - \kappa_2)_+^\ell + \ldots + \gamma_d (z_i - \kappa_{m-1})_+^\ell + \varepsilon_i,$$

  mit trunkierten Polynomen $(\cdot)_+^\ell$
  über Minimierung des PLS-Kriteriums

$$\text{PLS}(\lambda) = \sum_{i=1}^{n} (y_i - f(z_i))^2 + \lambda \sum_{j=\ell+2}^{d} \gamma_j^2 .$$

  (Bestrafe Abweichung vom globalen Polynom)

## Penalisierte Schätzansätze aus bayesianischer Sicht

- Penalisierungsterm kann als Glattheits- oder informative Priori angesehen werden
- Beispiel: Bayesianisches lineares gemischtes Modell
  - $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$
  - Nicht-informative Priori-Info für $\boldsymbol{\beta}$: $p(\boldsymbol{\beta}) \propto const$
  - Informative Gauß-Priori für $\boldsymbol{b}$: $\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G})$
  - Gemeinsame Posteriori-Verteilung der Regressionskoeffizienten über

$$p(\boldsymbol{\beta}, \boldsymbol{b} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{b}) p(\boldsymbol{\beta}) p(\boldsymbol{b})$$

$\Rightarrow$ Posteriori äquivalent zu penalisiertem KQ-Kriterium (9),
Posteriori-Modus äquivalent zu Koeffizientenschätzer des frequentistischen Ansatzes

## Literatur


- Fahrmeir, L., T. Kneib, S. Lang und B. Marx.
  *Regression - Models, Methods and Applications*. Springer, 2013.
- Hastie, T., R. Tibshirani und J. Friedman.
  *The Elements of Statistical Learning*. Springer, 2009.