ETH zürich

Least Absolute Shrinkage And Seletion Operator (LASSO)

Peter von Rohr

Qualitas AG

Lineare Modell und Least Squares

- Als Ausgangspunkt haben wir das Lineare Modell und Least Squares

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

$$\hat{\beta} = \mathit{argmin}_{\beta}||\mathbf{y} - \mathbf{X}\beta||^2 \tag{2}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3}$$

- Bedinung, dass $\hat{\beta}$ in Gleichung (3) berechenbar: $\mathbf{X}$ muss vollen Kolonnenrang haben, d.h. $p < n$ (mehr Beobachtungen als Parameter)

Genomische Selektion

- Paper: Meuwissen, Hayes und Goddard (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 1819–1829 (April 2001). (abgekürzt als MHG2001)
- Simulationsstudie: effektive Populationsgrösse: $N_e = 100$, Genom: 10 Chromosomen, pro Chromosom: 100 QTL und 101 SNP $\rightarrow n = 100$, $p = 1010$ und somit $n << p$
- Aktuelle Populationen: Anzahl typisierte Tiere: $\approx 10^4$, Genom: $50k$ oder $150k$
- **Achtung**: kein Overfitting! Auch wenn Anzahl typisierter Tiere $>$ Anzahl SNP, soll Least Squares nicht verwendet werden. Beispiel: Pläne von SEMEX 800000 Kühe zu typisieren. (Siehe auch: `https://de.wikipedia.org/wiki/Überanpassung`)

Lösungsansätze in MHG2001 (I): Stepwise

Schrittweises Hinzufügen von SNP-Effekten als fixe erklärende Variablen zu einem Regressionsmodell. Im Paper wird das als **stepwise-approach** bezeichnet, sollte aber korrekterweise als stepwise-forward-approach bezeichnet werden.

- Problem: wie finde ich die "richtigen" SNPs?
- Die mögliche Anzahl an Kombinationen ist sehr gross
- Sei p die Anzahl SNPs $\rightarrow$ Anzahl an Modellkombinationen 2^p (Kardinalität des Power-Sets)

Lösungsansätze in MHG2001 (II): BLUP

- **BLUP** für SNP-allel Effekte (a-Werte) können berechnet werden.
- Wir weisen aber a-priori allen SNPs den gleichen Varianzanteil zu
- Bei totaler genetischer Varianz V_g ohne polygene Effekte erklärt jeder SNP fix die Varianz V_g/p
- Vorteil: nur ein Varianzparameter muss geschätzt werden
- Kaum realistische Annahme, dass alle SNP-Loci gleich wichtig sind
- Bei dichteren Markerkarten erhöht sich die Anzahl der Loci mit erklärendem Einfluss
- Kaum sinnvoll, da von fixer Anzahl von Gene im Genom ausgegangen wird
- Mit BLUP ist keine Auswahl der erklärenden Loci möglich

Lösungsansätze in MHG2001 (III): Bayes

- Bayes'sche Statistik basiert Parameterschätzung auf a posteriori Verteilungen
- A posteriori Verteilungen sind proportional zu der a priori Verteilung und der Likelihood
- A priori, d.h. vor der Beobachtung der Daten sind die Parameter, so zum Beispiel die durch einen SNP-Locus erklärten Varianzanteil, durch die a-priori Verteilung bestimmt.
- Im Gegensatz zu BLUP, erlaubt Dies eine Variabilität zwischen den Varianzanteilen der einzelnen SNP-Loci

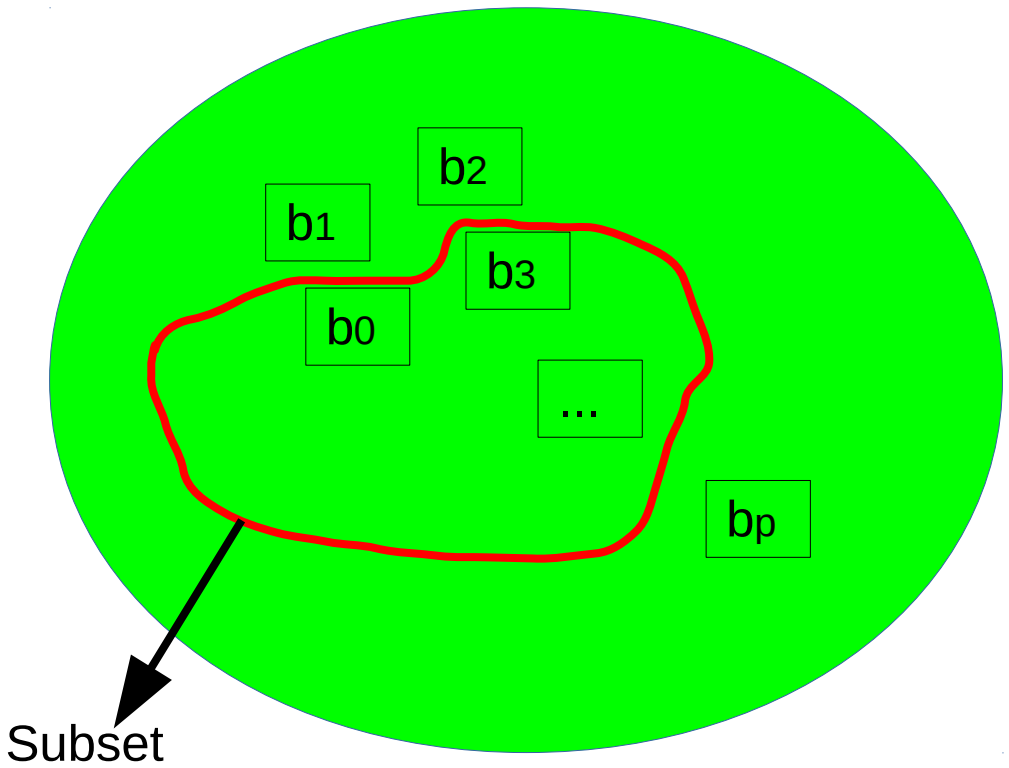
Lösungsansatz nicht in MHG2001: LASSO

- LASSO steht für Least Absolute Shrinkage and Selection Operator
- LASSO ist Teil einer grösseren Klasse von Methoden zur Selektion und Regularisation von Parameterschätzungen in linearen Modellen
- Weshalb braucht es solche Methoden?
  - Falls $n >> p$ funktioniert Least Squares gut, d.h. Parameterschätzungen sind unbiased und haben tiefe Varianz
  - Ist $n \approx p$ dann zeigen Least Square Schätzer erhöhte Variabilität und das Problem des Overfitting tritt auf
  - Ist $n < p$, dann können Least Squares Schätzer nicht berechnet werden.
- Positive Eigenschaften des Linearen Modells möchten wir trotzdem erhalten, somit suchen wir nach Alternativen zu Least Squares

Selektion und Regularisation (I)

Drei mögliche Alternativen zu Least Squares sind

- **Subset Selection**: Identifikation einer Teilmenge der p Parameter. Lineares Modell wird mit dieser Teilmenge an Parametern angepasst

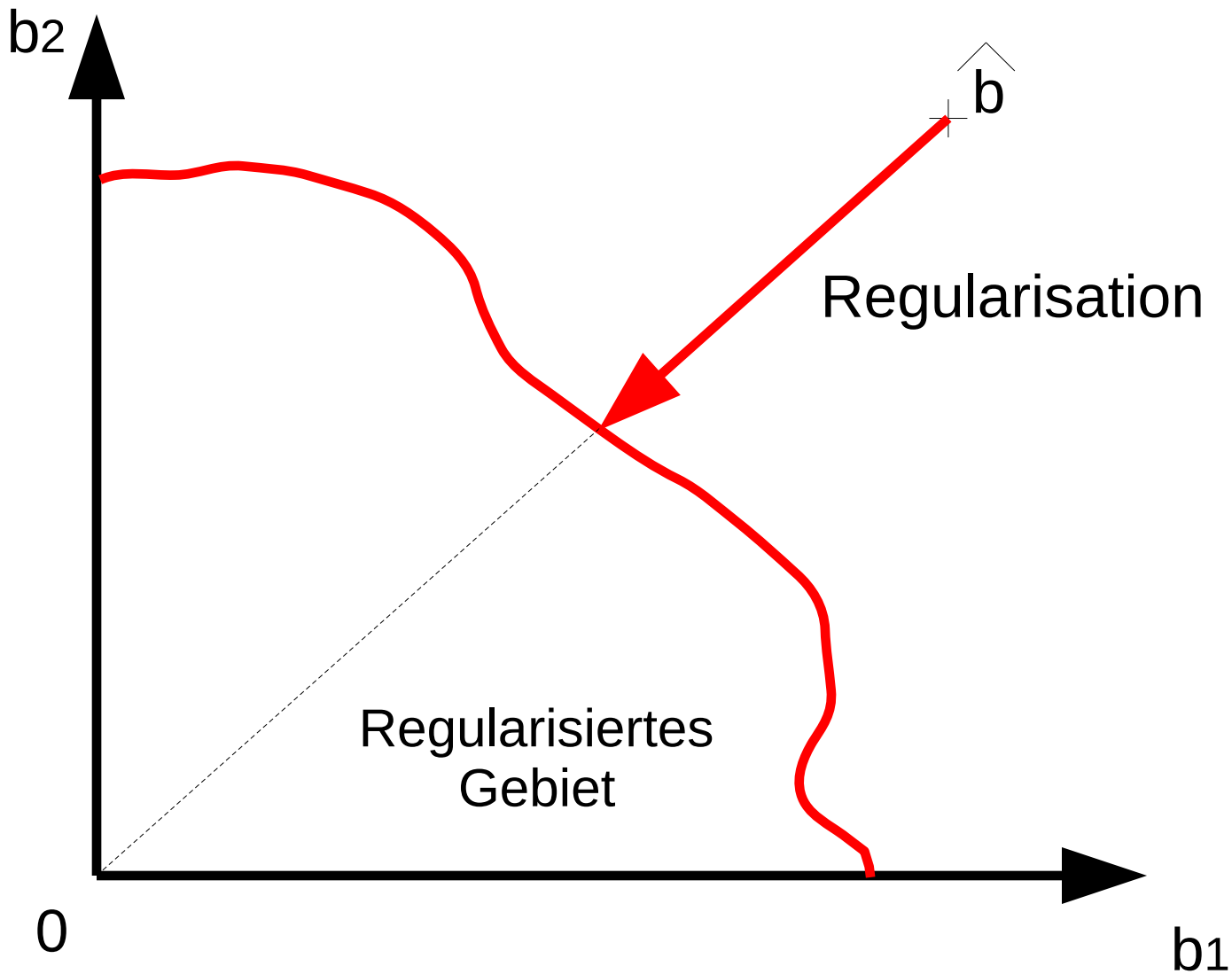


Selektion und Regularisation (II)

Shrinkage:

- Alle p Parameter werden verwendet um Modell anzupassen
- Variabilität der geschätzten Parameter wird kontrolliert durch Schrumpfung (Shrinkage, wird auch als Regularisation bezeichnet) der Parameterschätzungen zum Nullpunkt.
- Je nach verwendeter Art der Regularisation, werden dadurch gewisse Parameterschätzwerten auf Null gesetzt.
- Dadurch wird Regularisation und Parameterselektion kombiniert

Regularisation



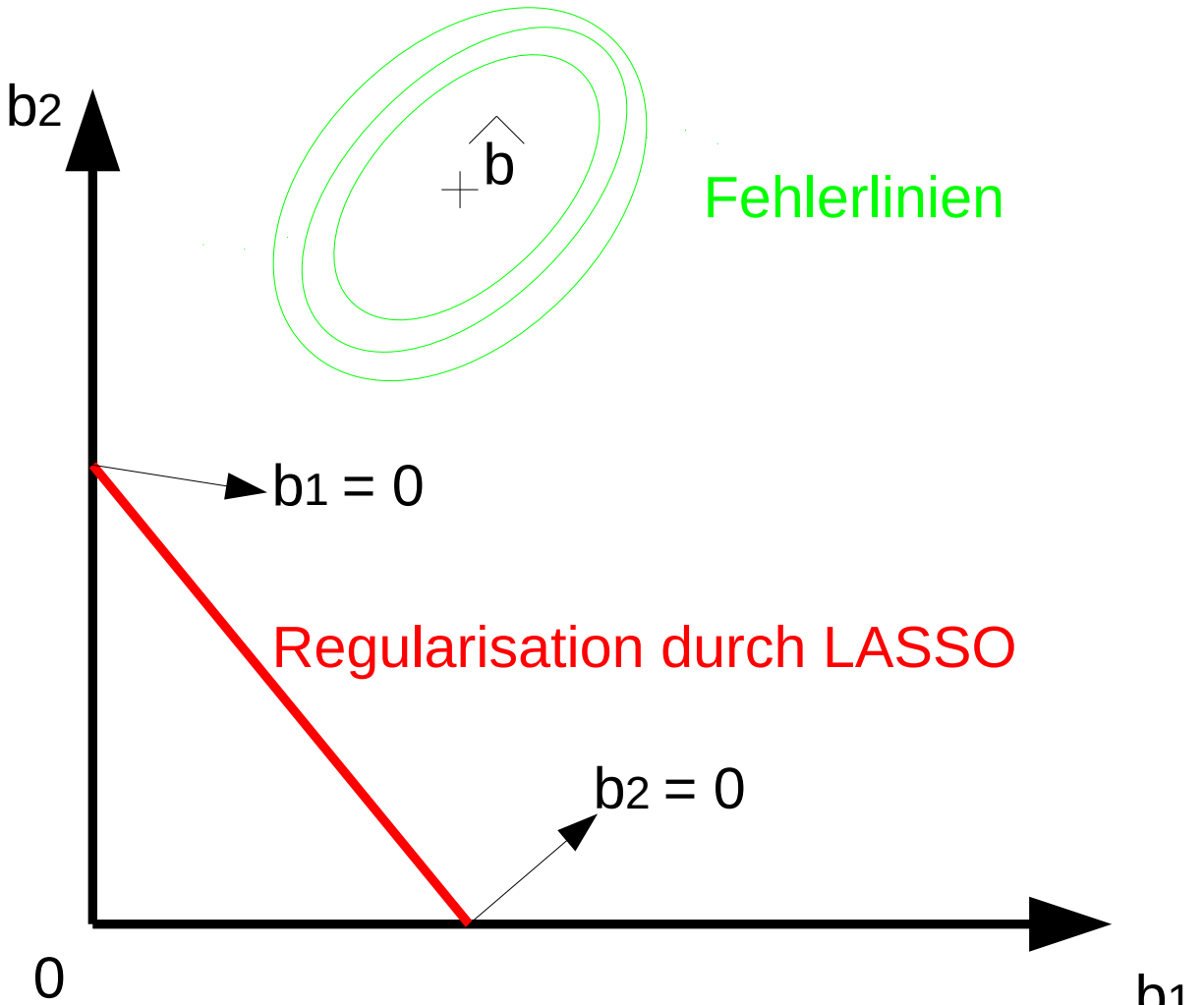
Selektion und Regularisation (III)

Reduktion der Dimensionen:

- Projektion der p erklärenden Variablen in einen M-dimensionalen Unterraum, wobei $M < p$
- Projektion wird erreicht durch Berechnung von M linearen Kombinationen aus den p erklärenden Variablen
- Least Squares wird mit den M Projektionen gemacht
- Techniken sind: Principal Components Analysis (PCA), Faktoranalyse, ...
- Problem: Interpretation der Ergebnisse, da Linearkombinationen oft keine Bedeutung (Bsp genomische Selektion, was bedeuten lineare Kombinationen aus SNP-Effekten?)

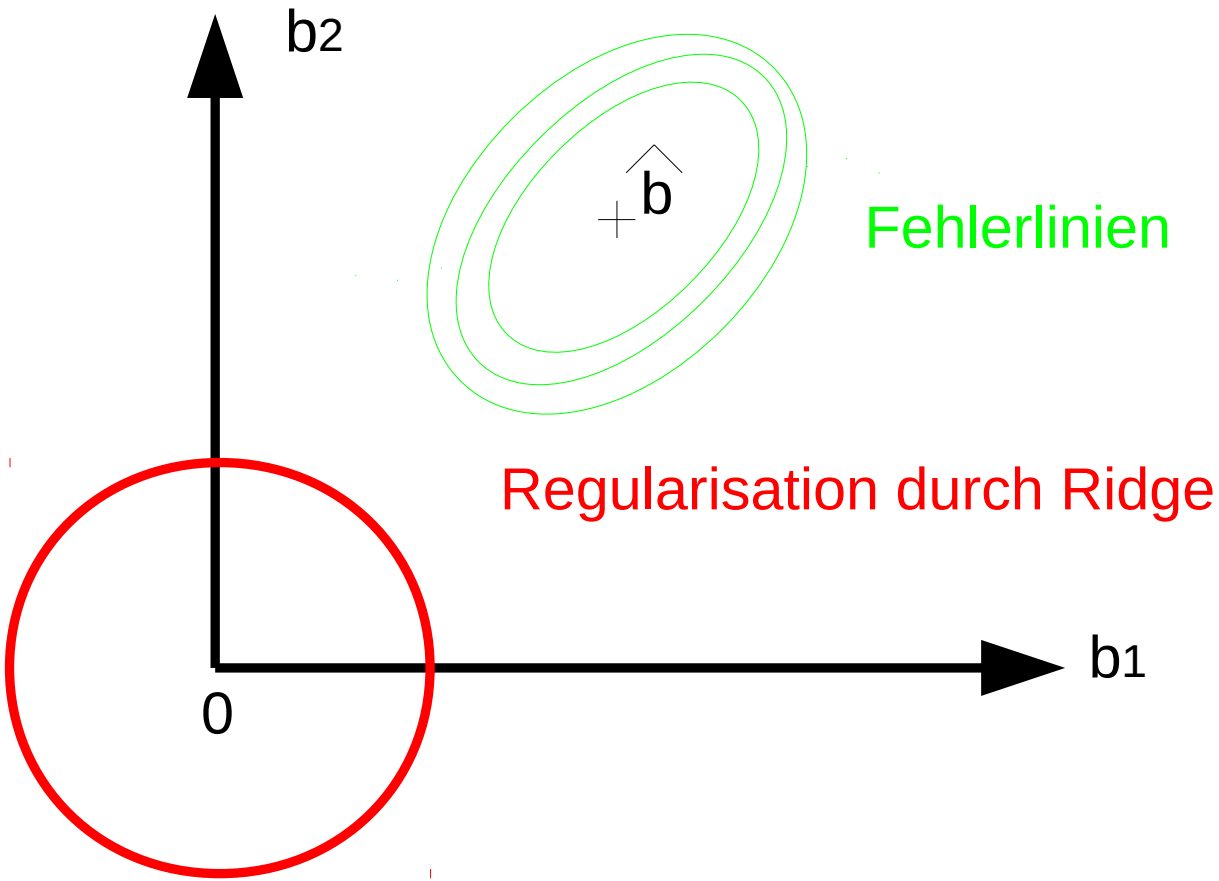
LASSO

- LASSO kombiniert Regularisation und Parameterselektion durch geschickte Veränderung der Zielgrösse aus Least Squares



Vergleich Ridge Regression

- Ridge Regression führt zu Regularisation aber nicht zu Parameterselektion



Parameterschätzung

- Residual Sums of Squares: $RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$
- Least Squares

$$\hat{\beta}_{LS} = argmin_{\beta}\{RSS\}$$

- LASSO

$$\hat{\beta}_{LASSO} = argmin_{\beta}\{RSS + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

- Ridge

$$\hat{\beta}_{Ridge} = argmin_{\beta}\{RSS + \lambda \sum_{j=1}^{p} \beta_j^2\}$$