Least Angle und L_1-penalisierte Regression

Kaspar Rufibach

Uni Zürich, ISPMZ, Abt. Biostatistik

6. März 2008

1 Fragestellung

2 Standardmethoden

3 Penalisierte Ansätze

4 “Least Angle Regression”

5 Spezialfall: “Elastic net”

6 Beispiel: Pathways in Alzheimer

1. Fragestellung
2. Standardmethoden
3. Penalisierte Ansätze
4. "Least Angle Regression"
5. Spezialfall: "Elastic net"
6. Beispiel: Pathways in Alzheimer

1. Fragestellung
2. Standardmethoden
3. Penalisierte Ansätze
4. "Least Angle Regression"
5. Spezialfall: "Elastic net"
6. Beispiel: Pathways in Alzheimer

1. Fragestellung
2. Standardmethoden
3. Penalisierte Ansätze
4. "Least Angle Regression"
5. Spezialfall: "Elastic net"
6. Beispiel: Pathways in Alzheimer

1. Fragestellung
2. Standardmethoden
3. Penalisierte Ansätze
4. "Least Angle Regression"
5. Spezialfall: "Elastic net"
6. Beispiel: Pathways in Alzheimer

1. Fragestellung
2. Standardmethoden
3. Penalisierte Ansätze
4. “Least Angle Regression”
5. Spezialfall: “Elastic net”
6. Beispiel: Pathways in Alzheimer

I've got all these variables, but I d'ont know which ones to use.

Ziele Modellselektion:

- Prädiktion.
- Interpretierbare Modelle, bestimmen von substanzwissenschaftlich bedeutsamen Prädiktoren.
- Stabilität: Kleine Änderungen in den Daten sollten keine grossen Änderungen verursachen für
  - die Menge der gewählten Prädiktoren
  - die Parameterschätzungen
  - die Prädiktionen.
- Vermeiden von Bias in Hypothesentests während oder nach Variablenselektion.

- "Forward stepwise": instabil $\Rightarrow$ kleine Änderungen in den Daten $\Rightarrow$ andere Selektionsreihenfolge $\Rightarrow$ komplett anderes Modell.
- "Backward" analog.
- "Greedy" (gierige) Verfahren:
  - Beste Wahl bei jedem Schritt, ohne Berücksichtigung zukünftiger Effekte.
  - Zwei beste Prädiktoren brauchen den "alleine-besten" nicht zu enthalten.
- "all subset": Noch mehr Modelle $\Rightarrow$ noch mehr Bias.
- t, F-Test für geschätzte Parameter: Setzen fixe Menge von Prädiktoren voraus, ansonsten verfälscht!
- Hauptprobleme: ad-hoc, instabil, Parameterinferenz verfälscht.

"Ridge regression": Minimiere

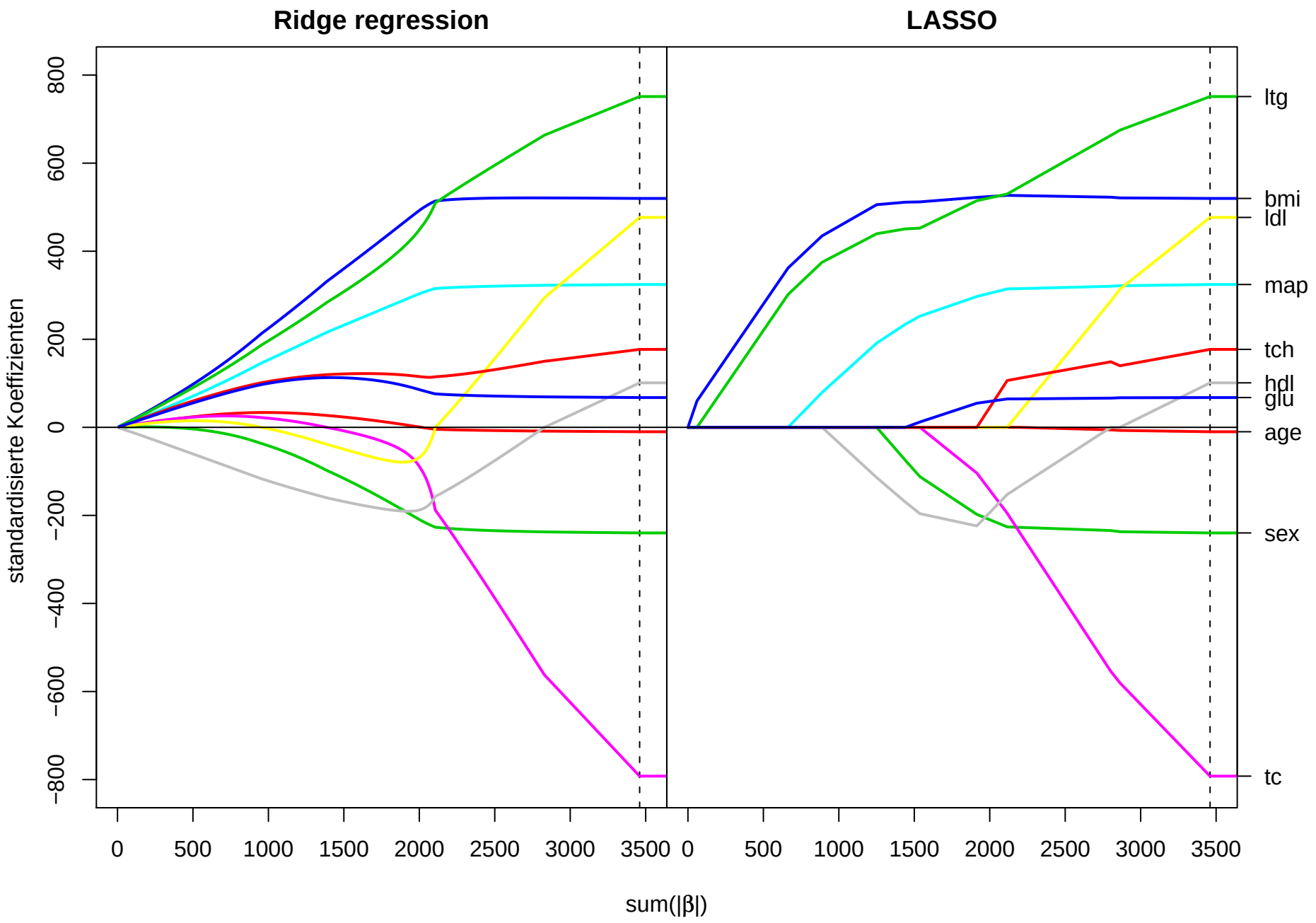
$$L(\boldsymbol{\beta}, \theta) \quad = \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \theta \sum_{j=1}^{p} \beta_j^2.$$

- Modell beinhaltet alle Prädiktoren $\Rightarrow$ keine Variablenselektion.
- Geschätzte Koeffizienten typischerweise "shrinked" (aber i.A. nicht auf 0).

LASSO: Minimiere

$$L(\boldsymbol{\beta}, \theta) \quad = \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \theta \sum_{j=1}^{p} |\beta_j|.$$

$\Rightarrow$ Variablenselektion und Parameterschätzung.


Ridge regression
LASSO
standardisierte Koeffizienten
800
600
400
200
0
−200
−400
−600
−800
0
500
1000
1500
2000
2500
3000
3500
sum(|β|)
ltg
bmi
ldl
map
tch
hdl
glu
age
sex
tc

Wichtige Variablen

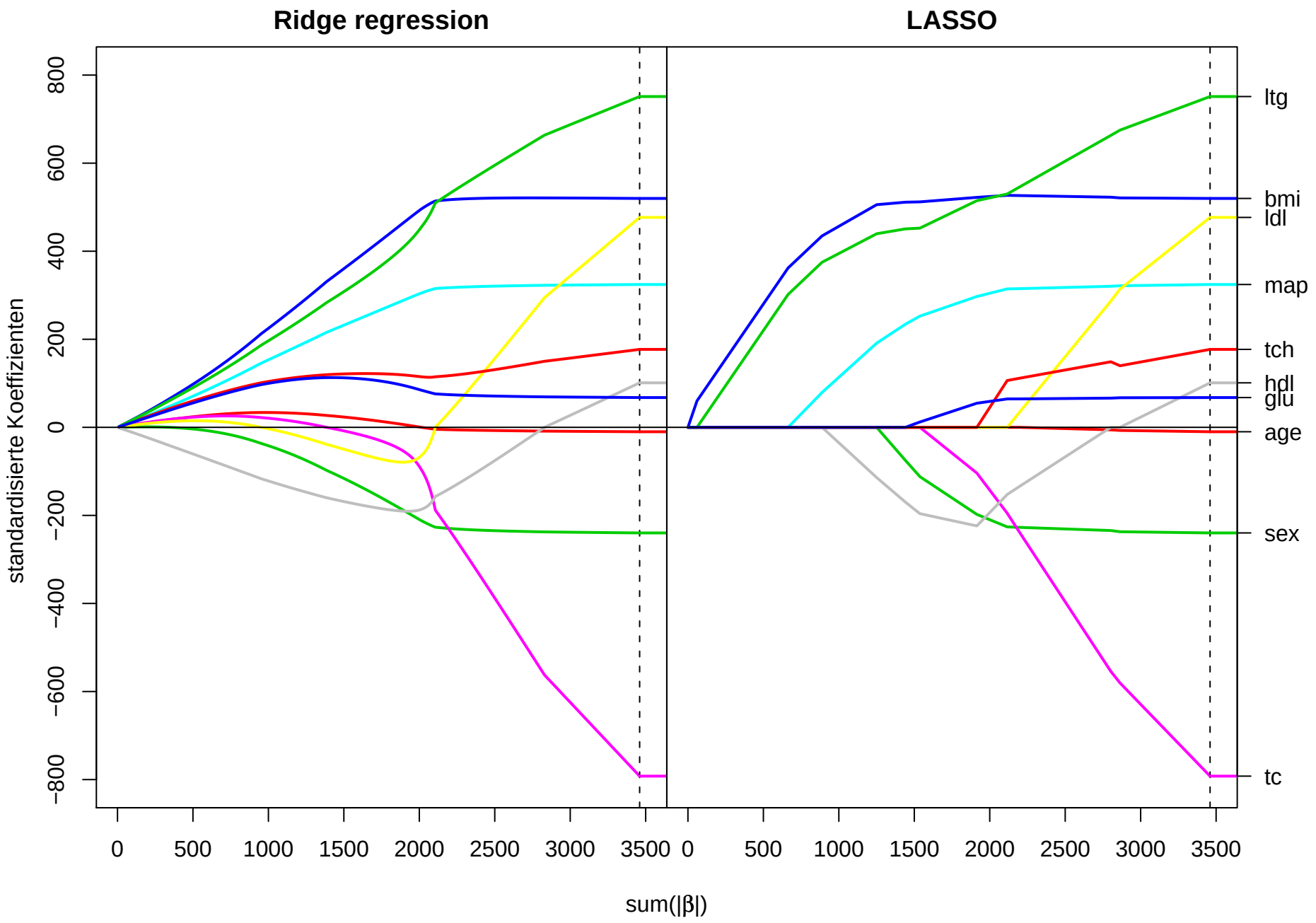
- Ridge: "drückt" β_j nach 0 mit Stärke proportional zu β_j.
- LASSO: gleiche Stärke auf alle nicht-verschwindenden Koeffizienten.

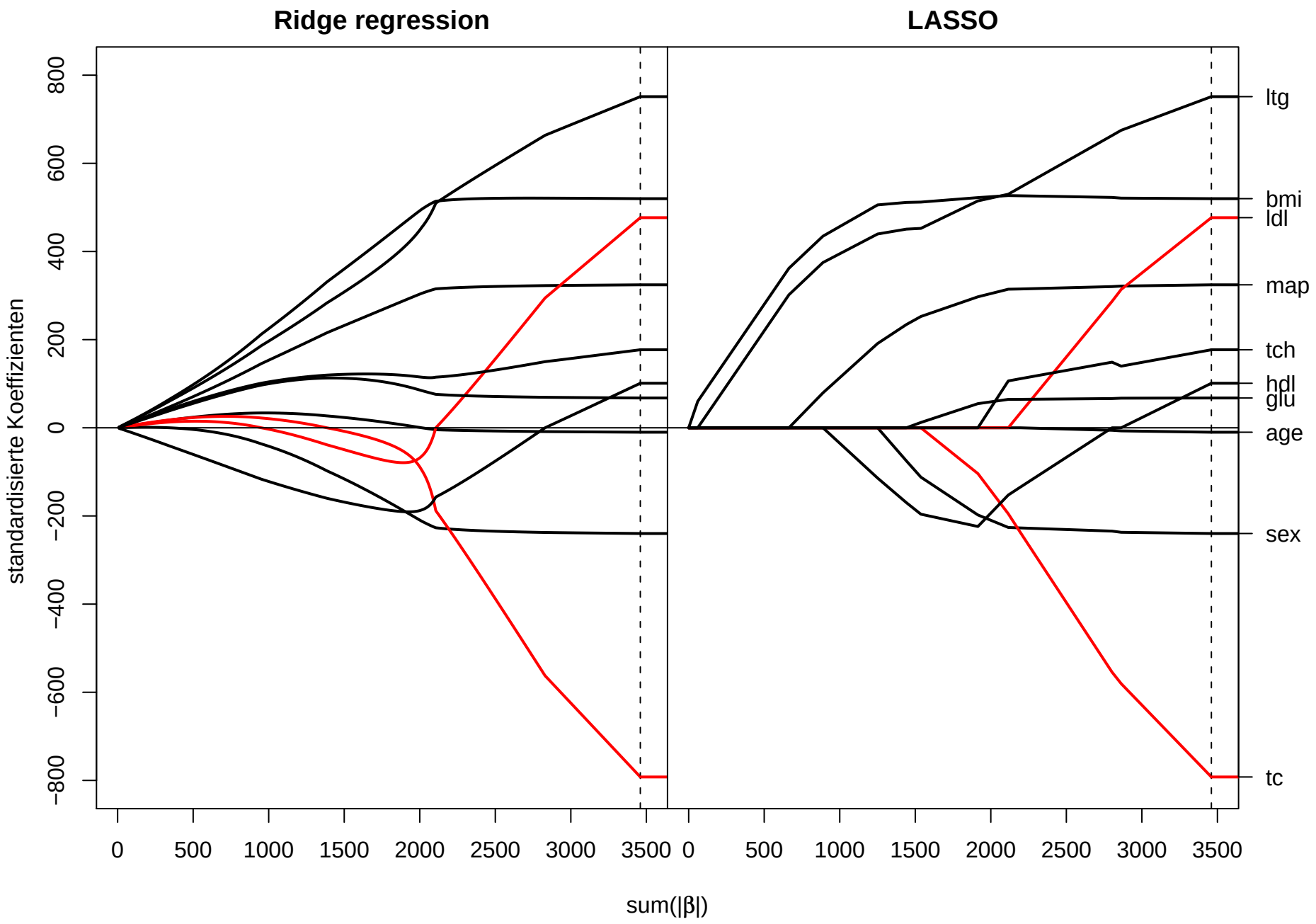
$\Rightarrow$ LASSO "shrinkt" nicht-verschwindende Koeffizienten weniger $\Rightarrow$ erklärt bessere Eigenschaften in Prädiktion.

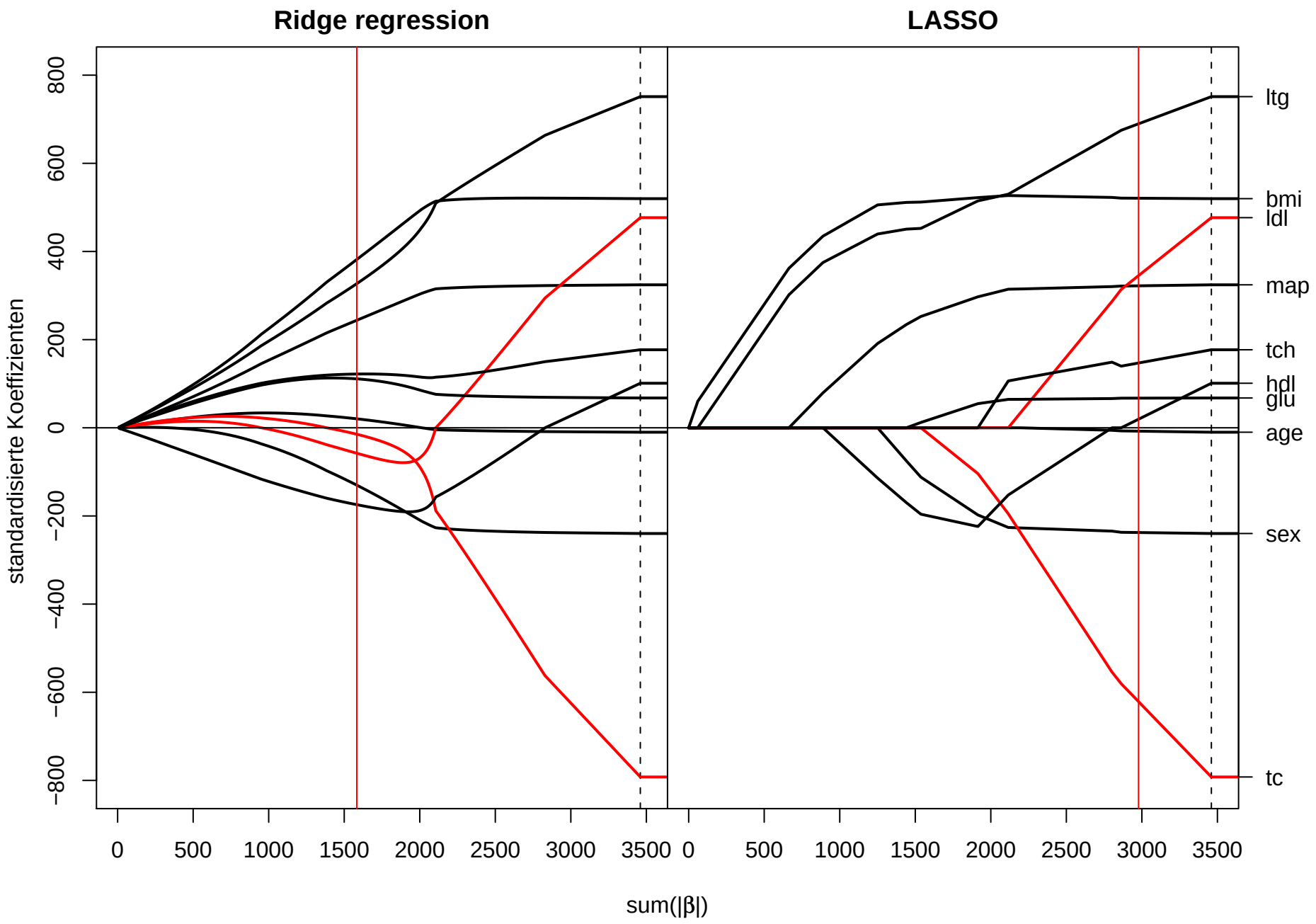
Variablenselektion

Cross-Validierung $\Rightarrow$ LASSO-Koeffizienten exakt 0!

Korrelierte Variablen: `tc`, `ldl` (r= 0.90) $\Rightarrow$ Instabilität.


Ridge regression
LASSO
standardisierte Koeffizienten
800
600
400
200
0
−200
−400
−600
−800
0
500
1000
1500
2000
2500
3000
3500
sum(|β|)
ltg
bmi
ldl
map
tch
hdl
glu
age
sex
tc


Ridge regression
LASSO
standardisierte Koeffizienten
800
600
400
200
0
−200
−400
−600
−800
0
500
1000
1500
2000
2500
3000
3500
sum(|β|)
ltg
bmi
ldl
map
tch
hdl
glu
age
sex
tc


Ridge regression
LASSO
standardisierte Koeffizienten
800
600
400
200
0
−200
−400
−600
−800
0
500
1000
1500
2000
2500
3000
3500
sum(|β|)
ltg
bmi
ldl
map
tch
hdl
glu
age
sex
tc

"Forward stepwise": Bester Prädiktor ins Modell, Koeffizient springt von 0 auf OLS - Wert. $\Rightarrow$ "greedy".

"Forward stagewise":

1. Erste Variable dieselbe, ändere Koeffizient um ε.
2. Berechne $\mathrm{cor}(\mathrm{residuen}, \mathbf{x}_j)$, wähle j mit der höchsten Korrelation, erhöhe β_j um ε.
3. ...

Koeffizienten stabiler als bei "forward stepwise".

Praktisch äquivalent zu "boosting", aber viel weniger rechenintensiv.

"Least angle regression": Wie "forward stagewise" $\Rightarrow$ Punkte an denen neue Variablen ins Spiel kommen analytisch ausgerechnet.

Eigenschaften:

- Schnell: Falls $p < n$ brauchts $O(p^3 + np^2)$ Rechenschritte – wie OLS für p Variablen.
- LAR(S) kann benutzt werden um LASSO und "stagewise models" zu fitten. "Unifying framework".
- Wahl der Anzahl Schritte: Minimiere

$$C_p(k) \quad = \quad (1/\widehat{\sigma}^2) \sum_{i=1}^{n} (y_i - \widehat{y} - i)^2 - n + 2k,$$

wobei k die Anzahl Schritte von LAR ist.

LAR, LASSO, stagewise ergeben i.A. ähnliche aber nicht gleiche Resultate.

Falls verschieden: LAR am schnellsten, da nur Variablen dazukommen.

Verallgemeinerung LAR auf GLM's: Interpretation als Newton-Methode.

Modifikationen/Alternativen Penalty: SCAD, adaptives LASSO, "relaxed LASSO", "Dantzig selector" (Candes & Terry Tao).

"Fused LASSO": nahe Prädiktoren sollen ähnlich sein:

$$L(\boldsymbol{\beta}, \theta_1, \theta_2) \quad = \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \theta_1 \sum_{j=1}^{p} |\beta_j| + \theta_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

"Elastic net": Hochkorrelierte Prädiktoren als Gruppe im oder aus dem Modell:

$$L(\boldsymbol{\beta}, \theta_1, \theta_2) \quad = \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \theta_1 \sum_{j=1}^{p} |\beta_j| + \theta_2 \sum_{j=1}^{p} \beta_j^2.$$

Witz: Strikt konvexe Penaltys haben Gruppierungseffekt (auch Ridge). In Verbindung mit L_1-Penalty $\Rightarrow$ "sparse solution".

Microarrays: Verwandte Gene in oder aus dem Modell als Gruppe.

Bessere Prädiktionen als LASSO bei korrelierten Prädiktoren (Bias-Varianz trade-off).

Falls $p \gg n$: Kann mehr als n Prädiktoren wählen, LASSO nicht!

Alzheimer's disease

- definite diagnosis only after death (brain autopsy)
- diagnosis in patient: test, interviews with relatives, ...
- Reliability of experienced assessor: 85% - 90 %

Goal: Find biochemical markers to unequivocally diagnose AD in living patient

Response: AD yes/no, in living patients

Explanatory: Company in Texas determined 89 markers (+ sex, age) in 273 patients' blood

Aim

- Ideally: From the set markers deduce entire up-/downregulated pathways.
- Predict AD from markers, e.g. via some regression technique.
- "Verify" results in knock–out mice.

Statistical problems:

- $n = 273$, $p = 91$.
- Some markers measure similar "things": highly correlated predictors.
- Estimate parameters and select variables.

$\Rightarrow$ Regularization necessary!

Just what the doctor ordered: Elastic net, Zhou & Hastie (2005).

What do we get as a result?

- A vector $\widehat{\boldsymbol{\beta}}_{\mathrm{EN}}(\widehat{\lambda}_1, \widehat{\lambda}_2)$ specifying the coefficients in the model.
- Some of the coefficients will be $=0$!
- An order in which the elements $\widehat{\beta}_{\mathrm{EN},j}(\widehat{\lambda}_1, \widehat{\lambda}_2)$ appear in the model.

What do we NOT get as a result?

- p-values, tests, and/or confidence intervals for the estimated coefficients.
- Variables in the model are not "significant", but simply "chosen by the elastic net".

Note: No assumptions on residuals made!

Variables 1-11	$\widehat{\beta}_{\mathrm{EN},j}(\widehat{\lambda}_1, \widehat{\lambda}_2)$	Variables 12-22	$\widehat{\beta}_{\mathrm{EN},j}(\widehat{\lambda}_1, \widehat{\lambda}_2)$
tnf_a	-0.1520	gm_csf	-0.0150
calcitonin	-0.1470	il_3	-0.0090
endothelin_1	-0.2211	il_8	0.0175
glutathione_s_trans	0.0840	papp_a	0.0108
eotaxin	0.0486	mmp_3	-0.0192
igf_1	-0.0795	il_2	-0.0031
tissue_fac	0.0473	ige	0.0052
carcinoembryonic	0.0430	tnf_b	-0.0032
g_csf	-0.0077	vegf	0.0053
il_6	-0.0990	vcam_1	-0.0012
mip_1a	0.0065	il_5	-0.0006

- So far, researchers could not connect (this list of chosen variables to up- or downregulated pathways.
- So no replication in mice on the way at the moment.
- Given the total cost of the data

$$\$400 \ / \ \text{patient} \ \Rightarrow \ \$80'000 \ \text{dataset}$$

they are working hard to provide explanations...

Points not discussed

- Use of least squares for binary response.
- Standardization of predictors.
- Imputation of values below "measurement barrier".
- Our $L(\boldsymbol{\beta}, \theta_1, \theta_2)$ is "naïve" elastic net, can be improved.

Danke für die Aufmerksamkeit.