Ex 3.3 (a)

$$\hat{\theta}_{LS} = a^T \hat{\beta} = a^T (X^T X)^{-1} X^T y \qquad E[a^T \hat{\beta}] = a^T \beta$$

$$\tilde{\theta} = c^T y \quad \underline{\text{unbiased}} \qquad \underline{c^T = a^T (X^T X)^{-1} X^T + \delta^T}$$

$$\begin{aligned}
E[\tilde{\theta}] &= E[c^T y] \\
&= E[a^T (X^T X)^{-1} X^T y + \delta^T y] \\
&= a^T \cancel{(X^T X)}^{-1} \cancel{X^T X} \beta + \delta^T X \beta \\
&= \underline{a^T \beta} + \delta^T X \beta \quad \Rightarrow \underline{\delta^T X = 0}
\end{aligned}$$

$$\begin{aligned}
Var(\tilde{\theta}) &= Var(c^T y) \\
&= c^T \sigma^2 c \\
&= \sigma^2 (a^T (X^T X)^{-1} X^T + \delta^T)(a^T (X^T X)^{-1} X^T + \delta^T)^T \\
&= \sigma^2 (a^T (X^T X)^{-1} X^T + \delta^T)(X (X^T X)^{-1} a + \delta) \\
&= \sigma^2 \Big(a^T (X^T X)^{-1} \cancel{X^T X} \cancel{(X^T X)}^{-1} a + a^T (X^T X)^{-1} \underbrace{X^T \delta}_{=0} + \\
&\qquad + \underbrace{\delta^T X}_{=0} (X^T X)^{-1} a + \delta^T \delta \Big) \\
&= \sigma^2 a^T (X^T X)^{-1} a + \sigma^2 \delta^T \delta
\end{aligned}$$

$$\underbrace{\sigma^2 a^T (X^T X)^{-1} a}_{= Var(\hat{\theta}_{LS})} \qquad \delta^T \delta \geq 0$$

Variable selection

IDEA: remove irrelevant variables from the model
↑
not useful to predict/explain the response

- less variance (despite small increase of bias)
- better interpretability
- better portability

When a variable is considered irrelevant
- p-value of a test $> \alpha$, usually $\alpha = 0.05$
- its inclusion increases a information criterion

INFORMATION CRITERIA

IDEA: instead of $\hat{\theta} = \arg\min_{\theta} L(\theta)$,

we find $\hat{\theta}_{IC} = \arg\min_{\theta} \{ L(\theta) + \lambda J(\theta) \}$

$-2\ell(\theta)$ $\sum_{j=1}^{p} \mathbb{1}[\theta_j \neq 0]$

$\hat{\theta}_{IC} = X\beta$

GOAL: penalize larger models

$L(\theta) = -2\ell(\theta)$

$J(\theta) = \sum_{j=1}^{p} \mathbb{1}[\theta_j \neq 0]$ # variables in the model

$$\lambda = \begin{cases} 2 & \text{AIC} \quad \text{Akaike information criterion} \\ \log(n) & \text{BIC} \quad \text{Bayesian information criterion} \end{cases}$$

n is the sample size

NB: using AIC for model selection is like using $\alpha = 0.157$

if two explanatory variables are strongly correlated → collinearity

extreme case: variables linearly dependent
→ super-collinearity

in the case of super-collinearity $(X^TX)^{-1}$ is not invertible (not full rank)

Haerl & Kennard (1970) $X^TX \to X^TX + \lambda I_p$

$$I_p = \underbrace{\begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots\ 1 \end{pmatrix}}_{p}$$

$$\boxed{\hat{\beta}_{ridge}^{(\lambda)} = (\underline{X^TX + \lambda I_p})^{-1} X^T y} \quad \lambda \in [0; \infty)$$

when $\lambda \in (0; \infty)$ $(X^TX + \lambda I_p)^{-1}$ exists

RIDGE REGRESSION AS A SHRINKAGE METHOD

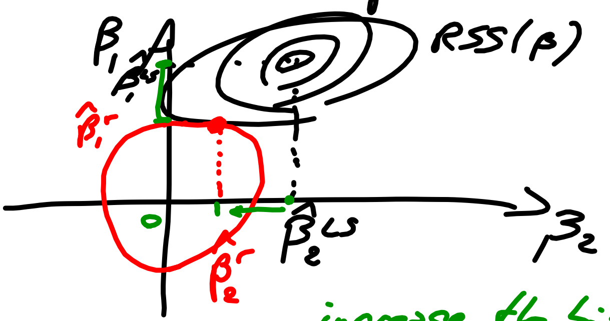
- $\hat{\beta}_{ridge}^{(\lambda)} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$

alternative formulation

- $\hat{\beta}_{ridge}^{(\lambda)} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}$

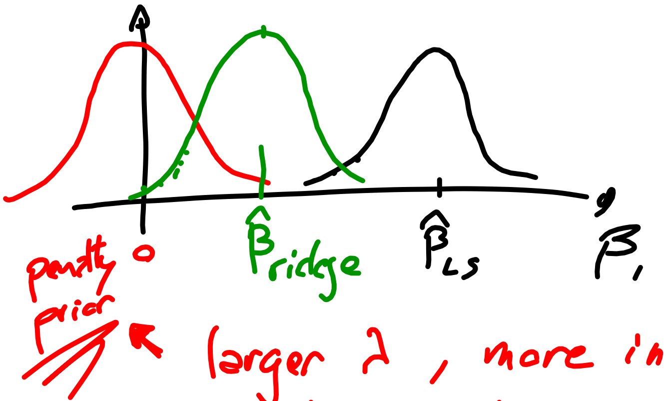
  subject to $\sum_{j=1}^{p} \beta_j^2 \le t$

– one to one correspondence between λ and t



← increase the bias, reduce the variance

- from a Bayesian point of view, ridge estimator is the posterior mean/mode

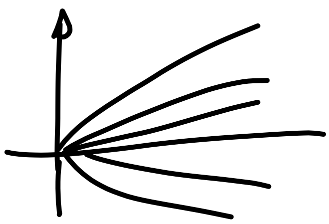


larger λ, more influence for the prior

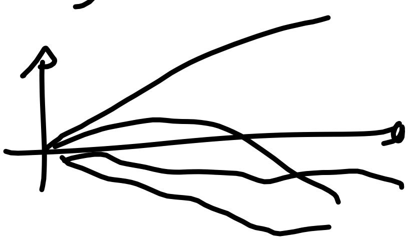
smaller λ, less " " " "

---

if Xs are uncorrelated



if Xs are correlated



IMPORTANT! We need to standardize our explanatory variables before applying the ridge regression

$$E[X_j] = 0$$

$$Var[X_j] = 1$$

Expected value of ridge estimator

$$E[\hat{\beta}^{(\lambda)}_{ridge}] = E[(X^TX + \lambda I_p)^{-1}X^Ty]$$

$$= E[(I_p + \lambda(X^TX)^{-1})^{-1}\underbrace{(X^TX)^{-1}X^Ty}_{\hat{\beta}_{LS}}]$$

$$= \underbrace{(I_p + \lambda(X^TX)^{-1})^{-1}}_{W_\lambda} E[\hat{\beta}_{LS}]$$

$$= W_\lambda \beta \qquad E[\hat{\beta}^{(\lambda)}_{ridge}] \neq \beta$$

$\lambda \to 0 \qquad E[\hat{\beta}^{(\lambda)}_{ridge}] = \beta$

$\lambda \to \infty \qquad E[\hat{\beta}^{(\lambda)}_{ridge}] = 0_{p\times 1}$ ← without intercept $\beta_0 = 0$

important $\quad \lambda_a > \lambda_b \not\Rightarrow |\hat{\beta}_j(\lambda_a)| < |\hat{\beta}_j(\lambda_b)|$

due to correlation

Variance of the ridge estimator

$$Var\left[\hat{\beta}^{(\lambda)}_{ridge}\right] = Var\left[W_\lambda \hat{\beta}_{LS}\right] = W_\lambda Var(\hat{\beta}_{LS}) W_\lambda^T = \sigma^2 W_\lambda (X^TX)^{-1} W_\lambda^T$$

$$Var(\hat{\beta}_{LS}) - Var(\hat{\beta}^{(\lambda)}_{ridge}) = \sigma^2\left[(X^TX)^{-1} - W_\lambda (X^TX)^{-1} W_\lambda^T\right]$$

$$= \sigma^2 W_\lambda \left[(\underline{I} + \lambda(X^TX)^{-1})\underline{(X^TX)^{-1}}(\underline{I} + \lambda(X^TX)^{-1}) - (X^TX)^{-1}\right] W_\lambda^T$$

$$= \sigma^2 W_\lambda \left[\cancel{(X^TX)^{-1}} + \lambda(X^TX)^{-2} + \lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3} - \cancel{(X^TX)^{-1}}\right] W_\lambda^T$$

$$= \sigma^2 W_\lambda \left[2\lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3}\right] W_\lambda^T > 0$$

$$\Rightarrow Var(\hat{\beta}^{(\lambda)}_{ridge}) \preccurlyeq Var(\hat{\beta}_{LS})$$

Degrees of freedom

$$\hat{Y}_{LS} = \underbrace{X(X^TX)^{-1}X^T}_{H} y \qquad df = tr(H) = p$$

$$\hat{Y}_{RIDGE} = \underbrace{X(X^TX + \lambda I)^{-1}X^T}_{H_\lambda} y \qquad \text{effective } df = tr(H_\lambda)$$

$$tr(H_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \qquad \begin{array}{l} \lambda = 0 \rightarrow df = p \\ \lambda \rightarrow \infty \rightarrow df \rightarrow 0 \end{array}$$

AIC : $L(\beta) + 2J(\theta)$

least square : $-2\ell(\beta) + 2p$

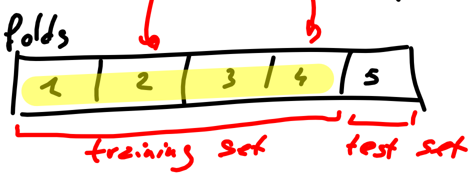
ridge : $-2\ell(\beta) + 2\sum \frac{d_j^2}{d_j^2+\lambda}$

AIC(λ) we can find $\hat{\lambda}$ as as $\arg\min_\lambda AIC(\lambda)$

usually λ is computed by cross-validation

sample: x_1 x_2 ..... x_n

split in K folds ($K=5$ or $K=10$)



$$err : \sum_{i=1}^{n_5} \left(y_{i[5]} - X_{[5]} \hat{\beta}_{[-5]}(\lambda) \right)^2$$

train the model on K-1 folds $\rightarrow \hat{\beta}_{[-5]}$

evaluate its performance on the remaining fold err_5

$\hat{\beta}$ computed by using all observation not belonging to fold 5

.repeat the procedure for K times, alway using a different fold as test set

err_1 : train set $\{2,3,4,5\}$
test set $\{1\}$

err_2 : train set $\{1,3,4,5\}$
test set $\{2\}$

⋮

err_5

$$\sum_{K=1}^{5} err_K(\lambda)$$

λ that minimizes

More about shrinkage

$$\underset{n\times p}{X} = UDV^T$$

U = $n\times n$ orthogonal matrix, whose columns span the column space of X

V = $p\times n$ orthogonal matrix, whose columns span the row space of X

D = $n\times n$ diagonal matrix, d_{ii} are single value of X_i

$$U^TU = I_n = V^TV$$

LS estimator $\hat\beta_{LS}$

$$\begin{aligned}\hat\beta_{LS} &= (X^TX)^{-1}X^Ty \\ &= (VD\underbrace{U^TU}DV^T)^{-1}VDU^Ty \\ &= (VD^2V^T)^{-1}VDU^Ty \\ &= VD^{-2}\underbrace{V^TV}DU^Ty \\ &= V\underbrace{D^{-2}D}U^Ty\end{aligned}$$

$$\hat y_{LS} = X\hat\beta_{LS} = UD\underbrace{V^TV}_{I}D^{-2}DU^Ty = UD^2D^{-2}U^Ty = UU^Ty$$

$$\sum_{j=1}^{p}\frac{d_j}{d_j^2}$$

$$\begin{aligned}\hat\beta^{(\lambda)}_{ridge} &= (X^TX+\lambda I)^{-1}X^Ty \\ &= (VD\underbrace{U^TU}DV^T+\lambda I_p)^{-1}VDU^Ty \\ &= (VD^2V^T+\lambda VV^T)VDU^Ty \\ &= V(D^2+\lambda I)^{-1}\underbrace{V^TV}_{I}DU^Ty \\ &= V(D^2+\lambda I_p)^{-1}DU^Ty\end{aligned}$$

$$\sum_{j=1}^{p}\frac{d_j}{d_j^2+\lambda}$$

$$\hat y_{rid} = X\hat\beta_{ridge} = UD\underbrace{V^TV}_{I}(D^2+\lambda I_p)^{-1}DU^Ty = U\underbrace{D^2(D^2+\lambda I_p)^{-1}}U^Ty$$

$$\rightarrow \sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda}$$

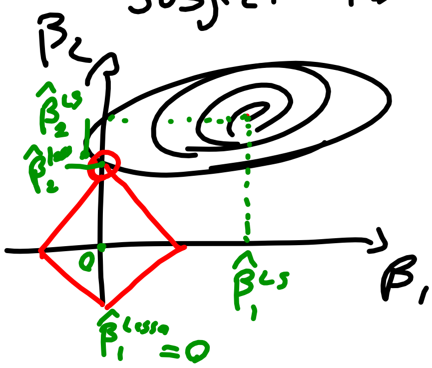
LASSO

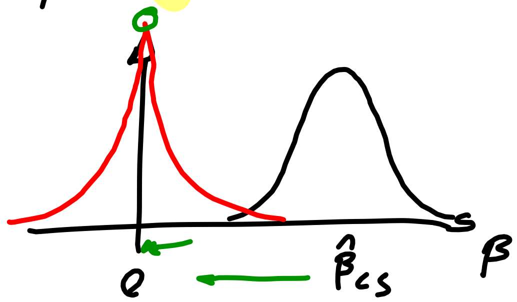
$$\hat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

or, equivalently

$$\hat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \right.$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$





$$\hat{\beta}_{pen} = \operatorname{argmin}_{\beta} \left\{ \sum_i \left(y_i - \beta_0 - \sum_j x_{ij}\beta_j\right)^2 + \underline{\lambda \sum_j |\beta|^q} \right\}$$

$q = 1 \rightarrow$ LASSO

$q = 2 \rightarrow$ RIDGE REGRESSION

$1 < q < 2 \approx$ average between lasso and ridge penalties

ELASTIC NET

penalty : $\lambda\left(\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha)\sum |\beta_j|\right)$

STK4030 - Statistical Learning: Advanced Regression and Classification

Riccardo De Bin

debin@math.uio.no

Procedure	Logic
BE-only	Estimate full model on $x_1, \ldots, x_k$. Repeat: while the least significant term has $P \geq \alpha_2$, remove it and re-estimate the model
BE	Estimate full model on $x_1, \ldots, x_k$. If the least significant term has $P \geq \alpha_2$, remove it and re-estimate; otherwise stop. Again: if the least significant term has $P \geq \alpha_2$, remove it and re-estimate; otherwise stop Repeat • if most significant excluded term has $P < \alpha_1$, add it and re-estimate; • if least significant included term has $P \geq \alpha_2$, remove it and re-estimate; until neither action is possible.
FS-only	Estimate null model. Repeat: while the most significant excluded term has $P < \alpha_1$, add it and re-estimate.
FS	Estimate null model. If the most significant excluded term has $P < \alpha_1$, add it and re-estimate; otherwise stop. Again: if the most significant excluded term has $P < \alpha_1$, add it and re-estimate; otherwise stop. Repeat • if least significant included term has $P \geq \alpha_2$, remove it and re-estimate; • if most significant included term has $P < \alpha_1$, add it and re-estimate; until neither action is possible.

from: Royston & Sauerbrei (2008)

Table 2.2 Myeloma study (65 patients, 48 events). Results of applying variable selection strategies.[a]

Variable	Nominal significance level, α						All-subsets
	0.05			0.157			
	Full	BE	FS	Full	BE	FS	AIC
x_1	∗	✓	✓	∗	✓	✓	✓
x_2			✓			✓	
x_3	∗	✓		∗	✓		✓
x_4		✓		∗	✓		✓
x_5							
x_6		✓		∗	✓		✓
x_7	∗	✓		∗	✓		✓
x_8				∗	✓		✓
x_9							
x_{10}							
x_{11}							
x_{12}	∗	✓		∗	✓		✓
x_{13}	∗	✓		∗	✓		✓
x_{14}				∗			
x_{15}							
x_{16}							

[a] BE(0.01) and FS(0.01) selected only one variable, x_1; ∗ denotes that a variable is significant at the relevant α-level in the full model.

from:
Royston &
Sauerbrei
(2008)

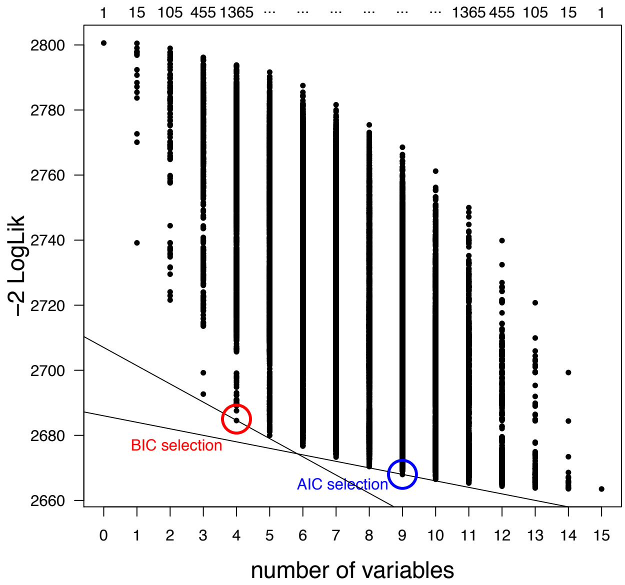

1 15 105 455 1365 ··· ··· ··· ··· ··· ··· 1365 455 105 15 1
−2 LogLik
2800
2780
2760
2740
2720
2700
2680
2660
BIC selection
AIC selection
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
number of variables

Table 2.3 Educational body-fat data. Full model and that selected by BE(0.05). Final three columns give details of the full model excluding x_6.

Variable	Full model			BE(0.05)			Full model excl. x_6		
	$\widehat{\beta}$	SE	$\widehat{\beta}$/SE	$\widehat{\beta}$	SE	$\widehat{\beta}$/SE	$\widehat{\beta}$	SE	$\widehat{\beta}$/SE
x_1	0.074	0.032	2.31	0.056	0.024	2.35	0.211	0.034	6.20
x_2	−0.019	0.067	−0.28				0.227	0.074	3.08
x_3	−0.249	0.191	−1.30	−0.322	0.121	−2.65	−0.915	0.212	−4.32
x_4	−0.394	0.234	−1.68				−0.378	0.278	−1.36
x_5	−0.119	0.108	−1.10				0.150	0.124	1.21
x_6	0.901	0.091	9.90	0.774	0.033	23.26	–	–	–
x_7	−0.146	0.144	−1.02				0.163	0.166	0.98
x_8	0.178	0.146	1.22				0.231	0.173	1.33
x_9	−0.041	0.245	−0.17				−0.095	0.291	−0.33
x_{10}	0.185	0.220	0.85				−0.053	0.259	−0.21
x_{11}	0.178	0.170	1.04				−0.066	0.200	−0.33
x_{12}	0.277	0.207	1.34				0.058	0.244	0.24
x_{13}	−1.830	0.529	−3.46	−1.943	0.406	−4.78	−2.692	0.620	−4.34

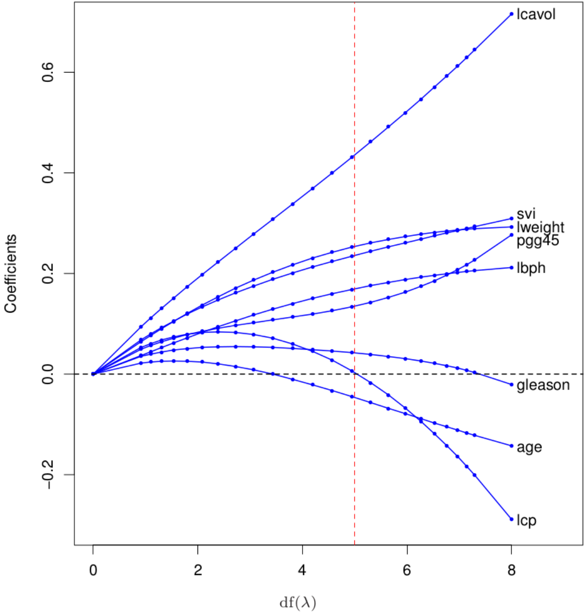
from: Royston & Sauerbrei (2008)

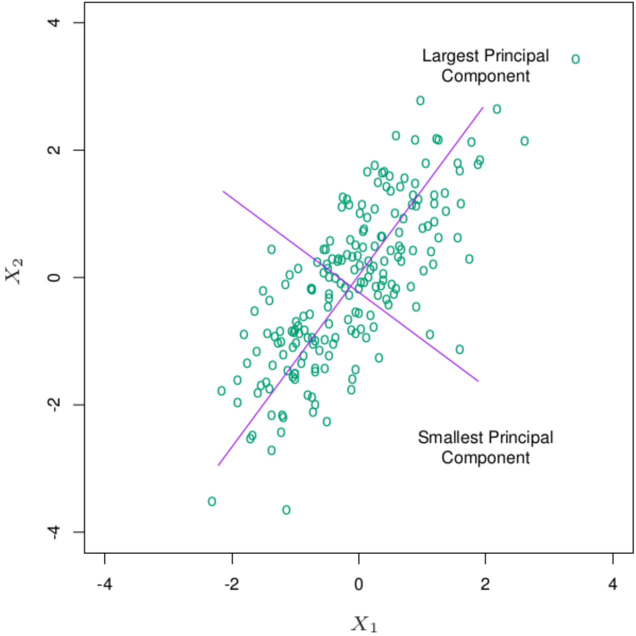
```r
library(Biobase)
library(breastCancerVDX)

# ids of genes FLOT1
idFLOT1 <- which(fData(vdx)[,5] == 10211)

# ids of ERBB2
idERBB2 <- which(fData(vdx)[,5] == 2064)

# get expression levels of probes mapping to FLOT genes
X <- t(exprs(vdx)[idFLOT1,])
X <- sweep(X, 2, colMeans(X))

# get expression levels of probes mapping to FLOT genes
Y <- t(exprs(vdx)[idERBB2,])
Y <- sweep(Y, 2, colMeans(Y))

# regression analysis
summary(lm(formula = Y[,1] ~ X[,1] + X[,2] + X[,3] + X[,4]))
```

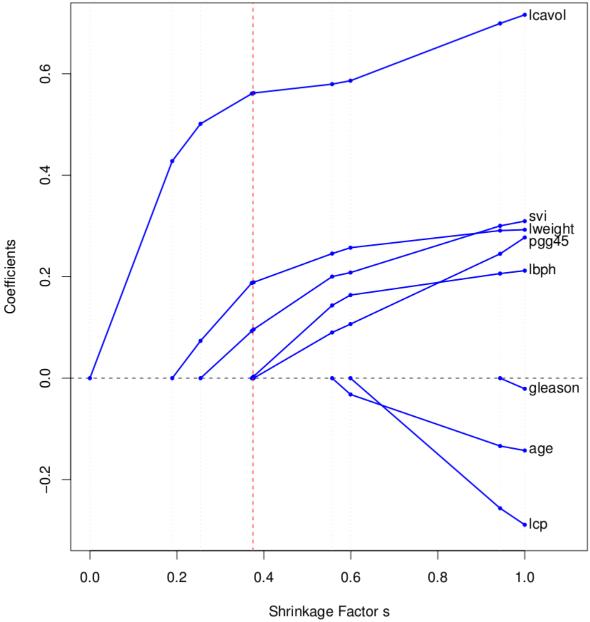
from: Van der Wieringen (2015)

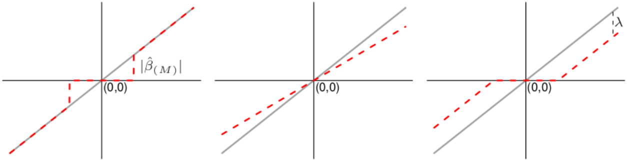
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0000     0.0633  0.0000   1.0000
X[, 1]        0.1641     0.0616  2.6637   0.0081 **
X[, 2]        0.3203     0.3773  0.8490   0.3965
X[, 3]        0.0393     0.2974  0.1321   0.8949
X[, 4]        0.1117     0.0773  1.4444   0.1496
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.175 on 339 degrees of freedom
Multiple R-squared:  0.04834,Adjusted R-squared:  0.03711
F-statistic: 4.305 on 4 and 339 DF,  p-value: 0.002072
```

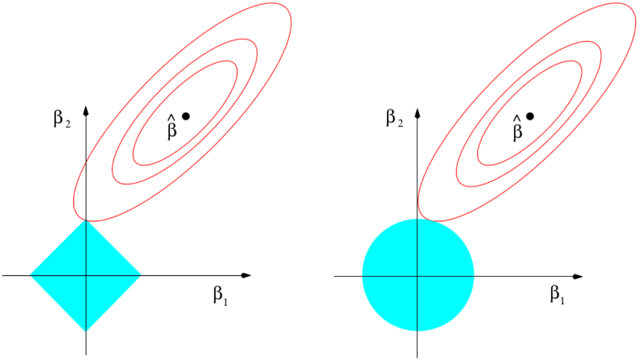
from: Van der Wieringen (2015)


lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp
Coefficients
0.6
0.4
0.2
0.0
−0.2
0
2
4
6
8
df(λ)


Largest Principal
Component
Smallest Principal
Component
4
2
0
-2
-4
-4
-2
0
2
4
X_1
X_2


lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp
Coefficients
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Shrinkage Factor s


$|\hat{\beta}_{(M)}|$
(0,0)
(0,0)
(0,0)
λ


β_2
$\hat{\beta}$
β_1
β_2
$\hat{\beta}$
β_1