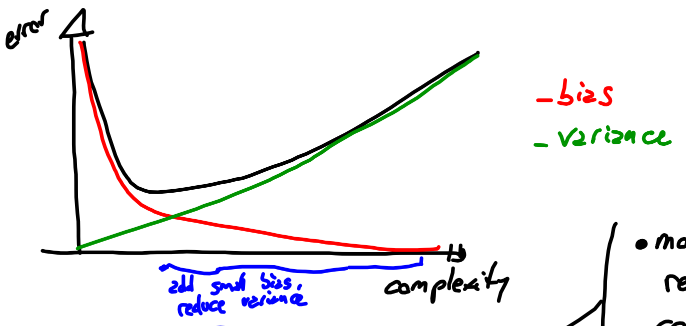
Methods of regularization



- ordinary least squares estimator
  BLUE (Best Linear Unbiased Estimator) → minimize the variance

accepting a small increase in the bias in order to reduce the variance to get the smallest possible prediction error

- model selection: reduce the variance by removing those variables that are not contributing much to the bias-reduction
- same idea for PCR

Penalized (regularized) regression → we add to our usual loss (deviance) a term (penalty) which force the estimates to be a little bit biased, but obtain a smaller variance

Notes: - we now assume to have centred response $y_i - \bar{y}$, because we do not want to penalize the intercept
- for reasons which will be clear soon, we will work with standardized X

$$x_i^{st} = \frac{x_i - \bar{x}}{se(x)}$$

Ridge regression

Consider a linear model $y = X\beta + \varepsilon$

Ridge regression finds the regression coefficient estimates by minimizing

$$D_{ridge}(\beta, \lambda) = \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{L_2 \text{ penalty}} = \|y - X\beta\|_2^2 + \lambda \beta^T\beta$$

λ: tuning parameter (penalty parameter)

The minimizer of this loss function is

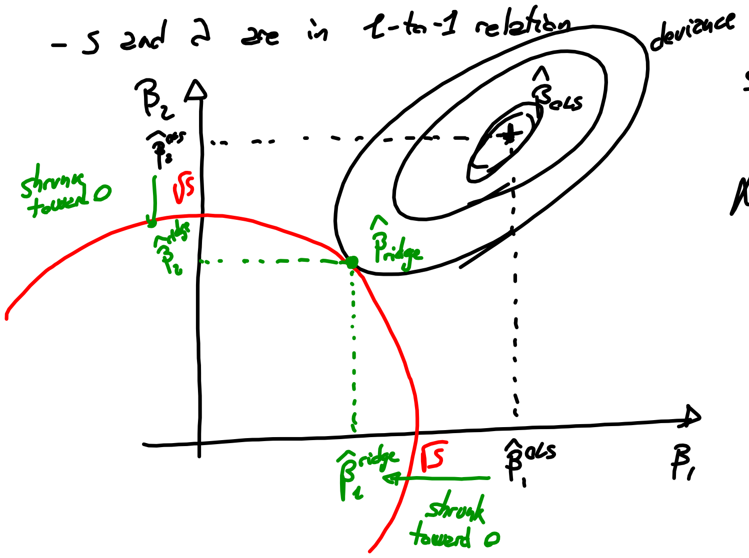
$$\hat{\beta}_{ridge}(\lambda) = (X^TX + \lambda I)^{-1} X^T y \qquad \boxed{\lambda > 0}$$

Note that λ is a very important parameter that controls the amount of penalization
- $\lambda = 0$ → no penalty, so $\hat{\beta}_{ridge} = \hat{\beta}_{OLS}$
- $\lambda = +\infty$ → each small deviation of $\hat{\beta}$ from 0 are strongly penalized, so $y - \bar{y} = 0$
  $\hat{\beta}_{ridge}^{(j)} = 0 \quad \forall j$
- normally, it is computed by cross-validation
- even for small λ, we reduce problems of collinearity
- allow regression of the case $p > n$

Alternative formulation.

Minimize $\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$ subject to $\sum_{j=1}^{p}\beta_j^2 \le s$

- s and λ are in 1-to-1 relation



simple case, two variables → two regression coefficients

Note: due to correlation, it is not necessarily true that
$\lambda_a > \lambda_b \not\Rightarrow |\hat\beta_j(\lambda_a)| < |\hat\beta_j(\lambda_b)|$
$s_a < s_b$

Bias

$E[\hat\beta_{OLS}] = \beta$ ordinary least squares estimator is unbiased BLUE

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

$\leftarrow (X^TX+\lambda I)^{-1}(X^TX)$
$([X^TX+\lambda I](X^TX)^{-1})^{-1}$

$$= E[\underbrace{(I_p+\lambda(X^TX)^{-1})^{-1}}_{W(\lambda)}\underbrace{(X^TX)^{-1}X^Ty}_{\hat\beta_{OLS}}]$$
$$= \underline{W(\lambda)\,E[\hat\beta_{OLS}]} = W(\lambda)\,\beta$$

the ridge estimator is unbiased only if $W(\lambda)=I$, i.e., $\lambda=0$

$$Var(\hat\beta_{ridge}) = Var(W(\lambda)\hat\beta_{OLS})$$
$$= W(\lambda)\,Var(\hat\beta_{OLS})\,W(\lambda)^T$$
$$= \sigma^2 W(\lambda)(X^TX)^{-1}W(\lambda)^T$$

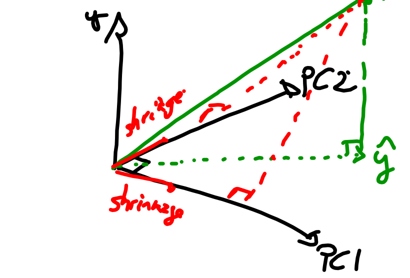
$Var(\hat\beta_{OLS}) = \sigma^2(X^TX)^{-1}$

$$Var(\hat\beta_{OLS}) - Var(\hat\beta_{ridge}) = \sigma^2\left[(X^TX)^{-1} - W(\lambda)(X^TX)^{-1}W(\lambda)^T\right]$$
$$= \sigma^2 W(\lambda)\left[W(\lambda)^{-1}(X^TX)^{-1}(W(\lambda)^T)^{-1} - (X^TX)^{-1}\right]W(\lambda)^T$$
$$= \sigma^2 W(\lambda)\left[(I_p+\lambda(X^TX)^{-1})(X^TX)^{-1}(I_p+\lambda(X^TX)^{-1}) - (X^TX)^{-1}\right]W(\lambda)^T$$
$$= \sigma^2 W(\lambda)\left[\cancel{(X^TX)^{-1}} + \lambda(X^TX)^{-2} + \lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3} - \cancel{(X^TX)^{-1}}\right]W(\lambda)^T$$
$$= \underline{\sigma^2}\,\underline{W(\lambda)}\left[2\lambda\underline{(X^TX)^{-2}} + \lambda^2\underline{(X^TX)^{-3}}\right]\underline{W(\lambda)^T} > 0$$

all are quadratic form

$$Var(\hat\beta_{ridge}^{(\lambda)}) \le Var(\hat\beta_{OLS})$$

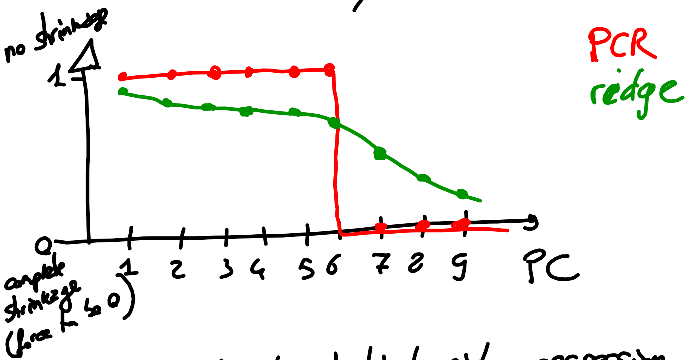
Geometric interpretation related to PCA



$$\hat{\beta}_1^{ols} = \frac{\|y\|}{\|pc1\|} \qquad \hat{\beta}_1^{ridge} = \frac{\|y\|}{\|pc1\| + \lambda}\, \hat{\beta}_1^{ols}$$

$$\hat{\beta}_2^{ols} = \frac{\|y\|}{\|pc2\|} \qquad \hat{\beta}_2^{ridge} = \frac{\|y\|}{\|pc2\| + \lambda}\, \hat{\beta}_2^{ols}$$

- ridge regression projects the response on the principal components
- shrink the low-variance components (likely to be noise) more than the high-variance components (likely to be signal)



The general idea behind ridge regression, is that we minimize a loss function of the form

$$\sum_{i=1}^{n} (y_i - x_i^T\beta)^2 + \lambda \underline{P(\beta)}$$

→ specifically for ridge regression , $P(\beta) = \beta^T\beta = \sum_{j=1}^{p} \beta_j^2$

$P(\beta)$ can but must not be the L_2 norm → we can use the L_1 norm $\sum_{j=1}^{p} |\beta_j|$

from the sum of squares to the sum of absolute values → $\sum_{j=1}^{p} |\beta_j|$

Minimize $\sum_{i=1}^{n} (y_i - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

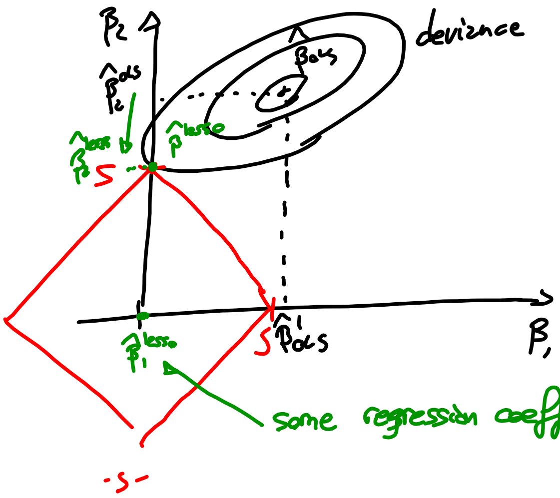
LASSO (Least Angle Shrinkage and Selection Operator)

The most interesting feature of LASSO is that it forces some estimates to be exactly equal to 0 → intrinsic variable selection → for a suitable λ

The same considerations done for λ of ridge regression are valid for LASSO

Using the alternative form for LASSO as well

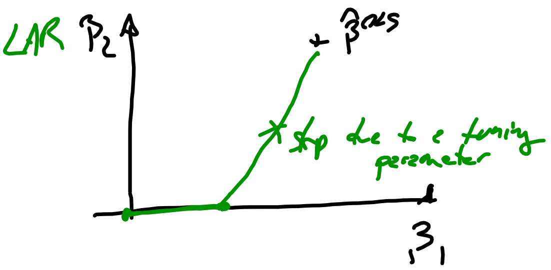
$$\hat{\beta}_{lasso} = \underset{\beta}{\arg\min} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$



some regression coefficients estimates are forced to be 0

Advantages
- shrinkage
- automatic variable selection

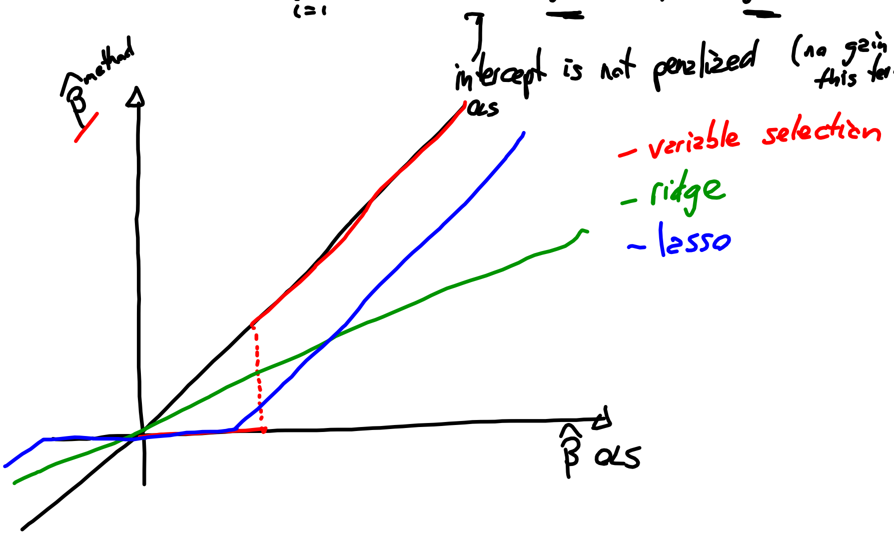
Disadvantages
- no close form for $\hat{\beta}^{lasso}$ due to the non-differentiability of the L_1 loss
- we need to rely on numerical computations
  - ↳ in practice, very good algorithms based on LAR, of which lasso is a special case



cor$(y - X\beta, x_1)$
cor$(y - X\beta, x_2)$

$$\sum_{i=1}^{n} \Big(y_i - \beta_0 + \sum_{j=1}^{p} x_j \beta_j \Big) + \sum_{j=1}^{p} |\beta|$$

intercept is not penalized (no gain in variance by penalizing this term)



Prediction of quantitative variables

Back to the initial problem, predict y using x

$y = f(x)$ → until now: parametric approach

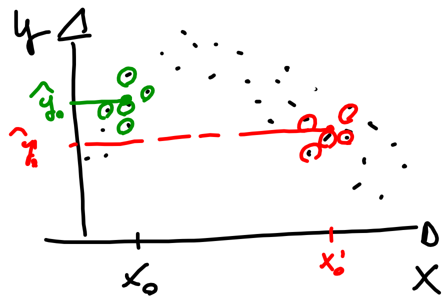
$$y = f(x;\beta)$$

family of functions indicized by β

$$\hat{y} = f(x;\beta)\big|_{\beta=\hat{\beta}}$$

- in the class of parametric functions find the best by selecting the best $\hat{\beta}$
- simple, easy to compute

Alternative: do not restrict to a parametric form, base our estimation on the data only (plus some regularity conditions) → NON-PARAMETRIC APPROACH



$$\hat{y}_0 = \hat{f}(x_0)$$

$$K=5 \quad = \frac{1}{5}\sum_{x_i \in N_5(x_0)} y_i$$

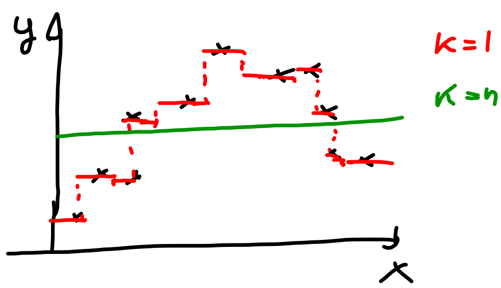
$$\hat{y}_0 = \frac{1}{K}\sum_{x_i \in N_K(x_0)} y_i$$

where $N_K(x_0)$ denotes the set of the K closest points to x_0

We assume that the values of the closest points are similar to the one of interest, and we base our estimates on those response → simple mean

Basically, we are assuming that $f(x)$ is constant close to x_0 $\quad f(x) = \beta_0$

K is the tuning parameter, that tells how many neighbours to include in the estimate
→ smaller value, more complex model, until the extreme $K=1$, in which each x_0 is estimated though its closest neighbour
→ larger values, simpler model, until the extreme $K=n$, in which x_0 is estimated with all the observations: $f(x_0) = \bar{y}$



- complexity is inverse of K
- how to choose K
  - CROSS-VALIDATION