Statistical Learning – week 3.2

Joris Bierkens

Delft University of Technology, The Netherlands

22 February 2024

Outline

Curse of dimensionality

1. Linear regression
    - Ordinary least squares
    - Ridge regression
    - LASSO

Assignments

- There will be four assignments in total
- These will consist of exercises and problems given after class
- Assignment deadlines are indicated on Brightspace
- Work together (meet up!) in groups of two or three
- Self-enroll in groups on Brightspace
- In your work:
  - clearly show your intermediate steps,
  - motivate your answer,
  - be to the point.
- Prepare clearly legible handwritten work (scanned, e.g. using CamScanner) or LaTeX.
- Submit using Brightspace by the deadline as a single PDF.
- Not adhering to these guidelines will result in a reduced grade.

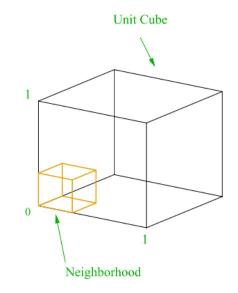
Recap learning objectives lecture 3.1

- Key distinctions: supervised vs unsupervised learning, regression vs classification (G)
- k-nearest neighbours as a simple example of supervised learning
- Probabilistic setting of supervised learning, population model (G)
- Loss functions and risk, residual sum of squares, Bayes estimator (G)
- Mean squared error, bias-variance trade-off (G)
- The use of training- and test-set to estimate risk (G)
- $\Rightarrow$ Curse of dimensionality (G)

Curse of dimensionality

A local method (e.g., k-nearest neighbours) works well if any new input x has many observations $x_1, \ldots, x_n$ in its vicinity.

Example: uniformly distributed points in a hypercube



- Suppose inputs $x_1, \ldots, x_n$ have uniform distribution in the hypercube $[0,1]^p$.
- How many points will lie in the sub-hypercube $[0,0.1]^p$?
- Answer: approximately $n \times (0.1)^p$.
- In order to maintain a fixed ratio of points in any sub-hypercube for growing p, we require n to grow exponentially in p!

Curse of dimensionality: reliable statistical inference becomes increasingly difficult in high dimensions

Function families and regularization

What happens when we minimize empirical risk over all possible functions f?
overfitting

Possible approaches

1. Constrained minimization: choose a suitable family of candidate functions $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{A}\}$.

   We say that $\mathcal{F}$ is a parametric family of functions if it allows the parametrization
   $$\mathcal{F} = \{f_\theta : \theta \in \Theta\},$$
   where $\Theta \subset \mathbb{R}^p$.
2. Adding a penalty for model complexity : regularization.

   As we will see this lecture, this is closely related to constrained optimization.
3. Following a Bayesian approach : will be topic of later lectures.

Linear regression model

- inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$; outputs $y_1, \ldots, y_n \in \mathbb{R}$.
- in linear regression we consider function estimators of the form $f : x \mapsto x^T \boldsymbol{\beta}$, so

$$f \in \mathcal{F} = \{\boldsymbol{x}^T \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

- the parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is called the vector of regression coefficients.
- often we add an intercept and consider

$$\mathcal{F} = \{\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta} : \boldsymbol{\beta}_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p\}$$

- equivalent to take $x_{i,0} = 1$ in the $(p+1)$-dimensional input vectors

$$x_i = (1, x_{i,1}, \ldots, x_{i,p}), \quad i = 1, \ldots, n,$$

with extended regression coefficient vector

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p).$$

Linear regression : ordinary least squares

- matrix notation $\boldsymbol{X} = (x_{ij})$, the jth component of the ith input vector
- $\boldsymbol{X}$ is called the design matrix
- residual sum of squares for observations (y_i)

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2.$$

- if rank $\boldsymbol{X} = p$ then the RSS is minimized by

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

- this is known as the ordinary least squares (OLS) estimator.
- what if rank $\boldsymbol{X} < p$?
  - For example, when $n < p$,
  - or in case of collinearity (dependence between inputs).

Mathematical intermezzo : Moore-Penrose inverse

- the (compact) singular value decomposition of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is given by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where
  - $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, $\boldsymbol{V} \in \mathbb{R}^{p \times r}$, both with orthonormal columns;
  - $\boldsymbol{D} = \mathrm{diag}(d_1, \dots, d_r)$ has $r \leq \min(p, n)$ positive diagonal elements $d_1 \geq \cdots \geq d_r > 0$: the non-zero singular values of $\boldsymbol{X}$.
- the Moore-Penrose inverse is given by

$$\boldsymbol{X}^+ = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T \in \mathbb{R}^{n \times p},$$

- for $\boldsymbol{y} \in \mathbb{R}^n$, the vector $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^+\boldsymbol{y}$ solves the problem

$$\text{minimize} \quad \|\boldsymbol{\beta}\| \quad \text{subject to} \quad \boldsymbol{\beta} \in \arg\min \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

- Exercise:
  - (a) If $r = p$, then $\boldsymbol{X}^+ = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$;
  - (b) If $r = n$, then $\boldsymbol{X}^+ = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$.

Linear regression : OLS estimator in terms of SVD

- design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$.
- let $r := \text{rank}(\boldsymbol{X})$. What if $r < p$?
- write $\boldsymbol{u}_k$ for the columns of $\boldsymbol{U}$, $\boldsymbol{v}_k$ for the columns of $\boldsymbol{V}$.
- minimizers of $\mathrm{RSS}(\boldsymbol{\beta})$ are given by

$$\hat{\boldsymbol{\beta}} = \sum_{k=1}^{r} d_k^{-1} \boldsymbol{v}_k \boldsymbol{u}_k^T \boldsymbol{y} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \in \ker(\boldsymbol{X}).$$

- the minimum-norm solution $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^+ \boldsymbol{y}$ (with $\boldsymbol{\eta} = 0$) solves the problem

$$\min \|\boldsymbol{\beta}\| \quad \text{subject to} \quad \boldsymbol{\beta} \in \arg\min \mathrm{RSS}(\boldsymbol{\beta}).$$

- in practice: expect numerical issues and high variance for 'large' p.

Linear regression : OLS bias and variance

Now assume that data are generated according to the (population) model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε has zero mean and finite variance σ^2, and $\boldsymbol{\beta}_0$ is the 'true' parameter.

- if rank $\boldsymbol{X} = p$, then the OLS estimator for $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$
- consider inputs $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ to be deterministic; outcomes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are random.
- $\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}_0$: $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}_0$.
- $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$.

Linear regression : MSE for $f(x)$

- prediction of $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}_0$ using $\hat{f}(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$.
- consider inputs $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x})$ to be fixed and known.

Mean Squared Error

$$\mathbb{E}\left[\hat{f}(\boldsymbol{x})\right] = \mathbb{E}\left[\boldsymbol{x}^T\hat{\boldsymbol{\beta}}\right] = \boldsymbol{x}^T\boldsymbol{\beta}_0 = f(\boldsymbol{x}),$$

$$\mathrm{Var}\left(\hat{f}(\boldsymbol{x})\right) = \sigma^2\boldsymbol{x}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x};$$

therefore

$$\begin{aligned}\mathrm{MSE}(\hat{f}(\boldsymbol{x}); f(\boldsymbol{x})) &= \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right] \\ &= \sigma^2\boldsymbol{x}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}.\end{aligned}$$

Linear regression: Expected Prediction Error

$$\mathrm{MSE}(\hat{f}(\boldsymbol{x}); f(\boldsymbol{x})) = \sigma^2 \boldsymbol{x}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}.$$

- estimator $\hat{f}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}; D_n)$ where $D_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$
- let y be new (random) observation associated with input $\boldsymbol{x}$.
- recall the Expected Prediction Error,

$$\begin{aligned}\mathrm{EPE}[\hat{f}](\boldsymbol{x}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &= \mathbb{E}_{\boldsymbol{x},y} \mathbb{E}_{D_n} \left[(y - \hat{f}(x))^2 \mid \boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right] \\ &= \mathrm{MSE}(\hat{f}(\boldsymbol{x}); f(\boldsymbol{x})) + \text{noise} = \sigma^2 \left(1 + \boldsymbol{x}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}\right).\end{aligned}$$

- now consider $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x})$ to be random as well.
- assume $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ is non-singular.

$$\mathrm{EPE}(\hat{f}) = \mathbb{E}_{(\boldsymbol{x},y)} \mathbb{E}_{D_n} \left[(y - \hat{f}(\boldsymbol{x}))^2\right] \sim \sigma^2 \left(1 + \frac{p}{n}\right) \quad (n \to \infty).$$

The Gauss-Markov theorem

In the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$$

assume $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{I}_n$

Notation

Suppose $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{R}^{p \times p}$ are symmetric. We write $\boldsymbol{P} \succeq \boldsymbol{Q}$ if $\boldsymbol{P} - \boldsymbol{Q}$ is positive semidefinite, i.e.,

$$\boldsymbol{a}^T(\boldsymbol{P} - \boldsymbol{Q})\boldsymbol{a} \geq 0 \quad \text{for all} \quad \boldsymbol{a} \in \mathbb{R}^p.$$

Gauss-Markov theorem

Let $\hat{\boldsymbol{\beta}}$ denote any unbiased linear estimator: $\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$, with $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y}$. Then $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) \succeq \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})$ where $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is the ordinary least squares estimator.

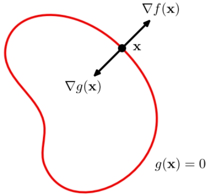
Does this mean that the OLS-estimator gives the smallest possible MSE?

Crash course: Lagrange multipliers (1/2)

[B06], Appendix E

Constrained minimization with equality constraints

$$\begin{aligned} \min \quad & f(\boldsymbol{x}), & f : \mathbb{R}^d \to \mathbb{R} \\ \text{subject to (s.t.)} \quad & \boldsymbol{g}(\boldsymbol{x}) = \mathbf{0} & \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^k. \end{aligned} \qquad (*)$$



- Consider $k = 1$, for simplicity
- At any point $\boldsymbol{x}$ on the hypersurface $g(\boldsymbol{x}) = 0$, $\nabla g(\boldsymbol{x})$ is orthogonal to the hypersurface.
- At any local optimum $\boldsymbol{x}$, $\nabla f(\boldsymbol{x})$ is orthogonal to the hypersurface.

Theorem

A necessary condition for $\boldsymbol{x}_* \in \mathbb{R}^d$ to be a minimum of (*) is that there is a $\boldsymbol{\lambda}^\star \in \mathbb{R}^k$ such that $(x^\star, \boldsymbol{\lambda}^\star)$ is a stationary point of the Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) := f(\boldsymbol{x}) + \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{x}).$$

Crash course: Lagrange multipliers (2/2)

Example

$$\begin{aligned} \min \quad & x_1^2 + x_2^2 \\ \text{s.t.} \quad & x_1 + x_2 = 1. \end{aligned}$$

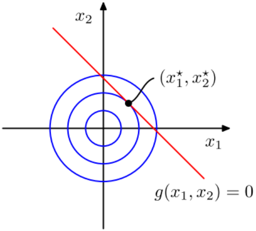
$$\mathcal{L}(x, \lambda) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1).$$

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \lambda) = \begin{bmatrix} 2x_1 + \lambda \\ 2x_2 + \lambda \end{bmatrix} = 0.$$

$$x_1 = -\lambda/2, \quad x_2 = -\lambda/2$$

$$\nabla_{\lambda} \mathcal{L}(\boldsymbol{x}, \lambda) = x_1 + x_2 - 1 = 0.$$

$$\lambda = -1, \quad x_1 = 1/2, \quad x_2 = 1/2.$$



Ridge regression

- consider situation with large p.
- partial explanation for high MSE: no penalty for large values of $\hat{\boldsymbol{\beta}}_i$.
- solution: shrink coefficients of $\hat{\boldsymbol{\beta}}$ by introducing a penalty term.
- example of regularization

Ridge regression

In the model $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$, the ridge regression objective function is

$$R(\boldsymbol{\beta}) = \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}_{\mathrm{RSS}(\boldsymbol{\beta})} + \underbrace{\lambda\|\boldsymbol{\beta}\|^2}_{\text{penalty}}, \quad \lambda > 0.$$

The ridge estimator for $\boldsymbol{\beta}$ is given by the minimizer,

$$\hat{\boldsymbol{\beta}}_{\mathrm{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Alternative problem formulation:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|^2 \leq t.$$

Ridge regression as a shrinkage method

recall the singular value decomposition $\boldsymbol{X} = \boldsymbol{UDV}^T$.

Ridge regression in terms of SVD

$$\begin{aligned}\hat{\beta}_{\text{ridge}} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T y \\ &= \boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_r)^{-1}\boldsymbol{D}^T\boldsymbol{U}^T\boldsymbol{y}.\end{aligned}$$

the fitted vector of outcomes is

$$\begin{aligned}\boldsymbol{X}\hat{\beta}_{\text{ridge}} &= \boldsymbol{UD}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_r)^{-1}\boldsymbol{D}^T\boldsymbol{U}^T\boldsymbol{y} \\ &= \sum_{j=1}^{r} \frac{d_j^2}{d_j^2 + \lambda}\boldsymbol{u}_j\boldsymbol{u}_j^T\boldsymbol{y}.\end{aligned}$$

$X^TX = (UDV^T)^T(UDV^T) = VD^TU^TUDV^T = VD^2V^T$

现在，我们可以用这个表达式来替换岭回归解中的 X^TX:

$\hat{\beta}_{\text{ridge}} = (VD^2V^T + \lambda I_p)^{-1}X^Ty$

注意到岭回归中 λI_p 可以表示为 $V\lambda IV^T$ 因为 $VV^T = I$。于是，我们得到:

$\hat{\beta}_{\text{ridge}} = (VD^2V^T + V\lambda IV^T)^{-1}X^Ty$

由于 V 是正交的，我们可以将 V 和 V^T 放到括号外面:

$\hat{\beta}_{\text{ridge}} = V(D^2 + \lambda I)^{-1}V^TX^Ty$

再次使用奇异值分解 $X = UDV^T$，我们有 $X^T = VDU^T$。因此，

$\hat{\beta}_{\text{ridge}} = V(D^2 + \lambda I)^{-1}V^TVDU^Ty$

由于 $V^TV = I$，它简化为:

$\hat{\beta}_{\text{ridge}} = V(D^2 + \lambda I)^{-1}DU^Ty$

$\rightarrow$ ridge regression shrinks the directions with small singular values d_j^2 relatively more

MSE of ridge regression

The MSE of ridge regression can be explicitly computed (exercise) to be

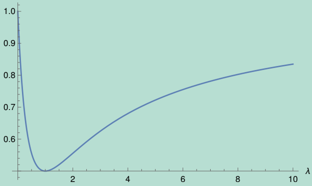
$$\begin{aligned}
&\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}; \boldsymbol{\beta}_0) \\
&= \lambda^2 (\boldsymbol{\beta}_0)^T \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-2} \boldsymbol{\beta}_0 + \sigma^2 \,\mathrm{tr}\left[\boldsymbol{X}^T \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-2}\right].
\end{aligned}$$

Example

Assume for simplicity $\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{I}_p$. Then

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}) = \frac{\lambda^2 \|\boldsymbol{\beta}_0\|^2 + p\sigma^2}{(1+\lambda)^2},$$

minimized at $\lambda = p\sigma^2 / \|\boldsymbol{\beta}_0\|^2$.



In this plot $p = 1, \sigma^2 = 1, \|\boldsymbol{\beta}_0\| = 1$

Variable selection and the LASSO

- ridge regression: penalty term $\|\boldsymbol{\beta}\|^2$ shrinks every parameter.
- can we recover a sparse coefficient vector?
- the penalty function $\|\boldsymbol{\beta}\|_0 := \sum_{j=1}^{p} \mathbb{1}_{\beta_j \neq 0}$ counts the number of non-zero parameters.
- problem: the optimization problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_0$$

  - is non-convex
  - requires systematically checking all combinations of non-zero β_j: combinatorial problem.
- alternative: the LASSO[1] optimization problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_1}_{=\sum_{j=1}^{p} |\beta_j|} .$$

[1] Least Absolute Shrinkage and Selection Operator

The LASSO

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

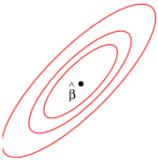
- convex optimization target: enables efficient computation
- alternative formulation

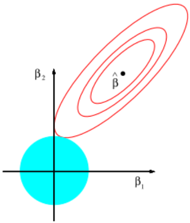
$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_1 \leq t.$$

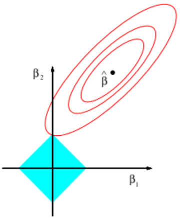
- yields sparse estimators

Ridge vs LASSO

OLS | **Ridge regression** | **LASSO**







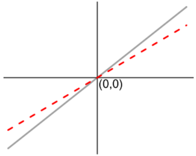
suppose $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$:

OLS: $\hat{\beta}_j$

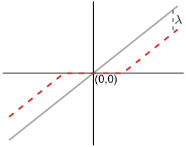
Ridge regression: $\hat{\beta}_j/(1+\lambda)$

LASSO: $\mathrm{sgn}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$

Ridge

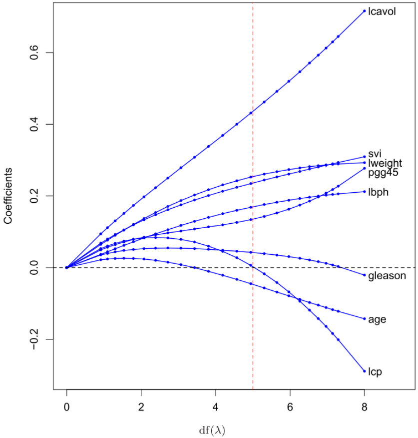


Lasso



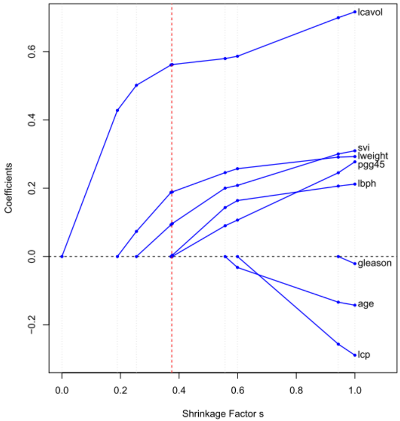
Ridge vs LASSO

Ridge regression



$$\text{df}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$$

LASSO



$$s = \frac{t}{\|\hat{\beta}_{\text{OLS}}\|_1}$$

Learning objectives lecture 3.2

- Parametric models (G)
- The linear regression model and the least squares estimator
- Linear regression in the over-parametrized case; Moore-Penrose inverse
- The Gauss-Markov theorem
- Refresher: Lagrange Multipliers (G)
- Regularized linear regression: ridge regression and the LASSO

Assignment and exercises

- First part of Assignment 1 is available on Brightspace (under Content - week 3.2)
- Exercises in lecture notes:
  - 1.8, 1.15 – 1.17, 1.19, 1.22, 1.24
  - 3.7, 3.12, 3.17 – 3.19