Feature selection and regularised regression

Rebecka Jörnsten, Mathematical Sciences

MSA220/MVE441 Statistical Learning for Big Data

10 th April 2025

CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Goals of modelling

1. **Predictive strength:** How well can we reconstruct the observed data? Has been most important so far.

Goals of modelling

1. **Predictive strength:** How well can we reconstruct the observed data? Has been most important so far.
2. **Model/variable selection:** Which variables are **part of the true model**? This is about uncovering structure to allow for mechanistic understanding.

Feature Selection

Remember ordinary least-squares (OLS)

Consider the model

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

where

- $\mathbf{y} \in \mathbb{R}^n$ is the **outcome**, $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is the **design matrix**, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ are the **regression coefficients**, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the **additive error**

Remember ordinary least-squares (OLS)

Consider the model

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

where

- $\mathbf{y} \in \mathbb{R}^n$ is the **outcome**, $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is the **design matrix**, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ are the **regression coefficients**, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the **additive error**
- **Five basic assumptions** have to be checked

    Underlying relationship is linear (1)

    Zero mean (2), uncorrelated (3) errors with constant variance (4), and no outliers (5). Optimal, but useful, that the errors are (roughly) normally distributed.

Remember ordinary least-squares (OLS)

Consider the model

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

where

- $\mathbf{y} \in \mathbb{R}^n$ is the **outcome**, $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is the **design matrix**, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ are the **regression coefficients**, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the **additive error**
- **Five basic assumptions** have to be checked

  Underlying relationship is linear (1)

  Zero mean (2), uncorrelated (3) errors with constant variance (4), and no outliers (5). Optimal, but useful, that the errors are (roughly) normally distributed.
- **Centring** ($\frac{1}{n}\sum_{l=1}^{n} x_{lj} = 0$) and **standardisation** ($\frac{1}{n}\sum_{l=1}^{n} x_{lj}^2 = 1$) of predictors simplifies interpretation

Remember ordinary least-squares (OLS)

Consider the model

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

where

- $\mathbf{y} \in \mathbb{R}^n$ is the **outcome**, $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is the **design matrix**, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ are the **regression coefficients**, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the **additive error**
- **Five basic assumptions** have to be checked

  Underlying relationship is linear (1)

  Zero mean (2), uncorrelated (3) errors with constant variance (4), and no outliers (5). Optimal, but useful, that the errors are (roughly) normally distributed.
- **Centring** ($\frac{1}{n}\sum_{l=1}^{n} x_{lj} = 0$) and **standardisation** ($\frac{1}{n}\sum_{l=1}^{n} x_{lj}^2 = 1$) of predictors simplifies interpretation
- **Centring** the outcome ($\frac{1}{n}\sum_{l=1}^{n} y_l = 0$) and features removes the need to estimate the intercept

Feature selection as motivation

Analytical solution exists when $\mathbf{X}^\top\mathbf{X}$ is invertible

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Feature selection as motivation

Analytical solution exists when $\mathbf{X}^\top\mathbf{X}$ is invertible

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

The solution can be unstable or impossible to compute if

- there is **high correlation** between predictors, or
- if $p > n$.

Feature selection as motivation

Analytical solution exists when $\mathbf{X}^\top\mathbf{X}$ is invertible

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

The solution can be unstable or impossible to compute if

- there is **high correlation** between predictors, or
- if $p > n$.

Solutions: **Regularisation** or **feature selection**

Filtering for feature selection

- Choose features through pre-processing
  - Features with maximum variance
  - Use only the first k PCA components

Filtering for feature selection

- Choose features through pre-processing
  - Features with maximum variance
  - Use only the first k PCA components
- Examples of other useful measures
  - Use a univariate criterion, e.g. **F-score:** Features that correlate most with the response
  - **Mutual Information:** Reduction in uncertainty about $\mathbf{x}$ after observing y
  - **Variable importance:** Determine variable importance with random forests

Filtering for feature selection

- Choose features through pre-processing
  - Features with maximum variance
  - Use only the first k PCA components
- Examples of other useful measures
  - Use a univariate criterion, e.g. **F-score:** Features that correlate most with the response
  - **Mutual Information:** Reduction in uncertainty about $\mathbf{x}$ after observing y
  - **Variable importance:** Determine variable importance with random forests
- **Summary**
  - **Pro:** Fast and easy
  - **Con:** Filtering mostly operates on single features and is not geared towards a certain method
  - Care with cross-validation and multiple testing necessary

Filtering for feature selection

- Choose features through pre-processing
  - Features with maximum variance
  - Use only the first k PCA components
- Examples of other useful measures
  - Use a univariate criterion, e.g. **F-score:** Features that correlate most with the response
  - **Mutual Information:** Reduction in uncertainty about $\mathbf{x}$ after observing y
  - **Variable importance:** Determine variable importance with random forests
- **Summary**
  - **Pro:** Fast and easy
  - **Con:** Filtering mostly operates on single features and is not geared towards a certain method
  - Care with cross-validation and multiple testing necessary
- Filtering is often more of a pre-processing step and less of a proper feature selection step

Wrapping for feature selection

- **Idea:** Determine the best set of features by fitting models of different complexity and comparing their performance

Wrapping for feature selection

- **Idea:** Determine the best set of features by fitting models of different complexity and comparing their performance
- **Best subset selection:** Try all possible (**exponentially many**) subsets of features and compare model performance with e.g. cross-validation

Wrapping for feature selection

- **Idea:** Determine the best set of features by fitting models of different complexity and comparing their performance
- **Best subset selection:** Try all possible (**exponentially many**) subsets of features and compare model performance with e.g. cross-validation
- **Forward selection:** Start with just an intercept and add in each step the variable that improves fit the most (**greedy algorithm**)

Wrapping for feature selection

- **Idea:** Determine the best set of features by fitting models of different complexity and comparing their performance
- **Best subset selection:** Try all possible (**exponentially many**) subsets of features and compare model performance with e.g. cross-validation
- **Forward selection:** Start with just an intercept and add in each step the variable that improves fit the most (**greedy algorithm**)
- **Backward selection:** Start with all variables included and then remove sequentially the one with the least impact (**greedy algorithm**)

Wrapping for feature selection

- **Idea:** Determine the best set of features by fitting models of different complexity and comparing their performance
- **Best subset selection:** Try all possible (**exponentially many**) subsets of features and compare model performance with e.g. cross-validation
- **Forward selection:** Start with just an intercept and add in each step the variable that improves fit the most (**greedy algorithm**)
- **Backward selection:** Start with all variables included and then remove sequentially the one with the least impact (**greedy algorithm**)
- As discreet procedures, all of these methods **exhibit high variance** (small changes could lead to different predictors being selected, resulting in a potentially very different model)

Embedding for feature selection

- **Embed/include** the feature selection into the model estimation procedure

Embedding for feature selection

- **Embed/include** the feature selection into the model estimation procedure
- Ideally, penalization on the number of included features

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \mathbb{1}(\beta_j \neq 0)$$

However, **discrete optimization problems** are hard to solve

Embedding for feature selection

- **Embed/include** the feature selection into the model estimation procedure
- Ideally, penalization on the number of included features

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \mathbb{1}(\beta_j \neq 0)$$

  However, **discrete optimization problems** are hard to solve
- **Softer regularisation methods** can help

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q^q$$

  where λ is a tuning parameter and $q \geq 1$ or $q = \infty$.

Feature selection

Feature selection can be addressed in multiple ways

- **Filtering:** Remove variables before the actual model for the data is built
  - Often crude but fast
  - Typically only pays attention to one or two features at a time (e.g. F-Score, MIC) or does not take the outcome variable into consideration (e.g. PCA)
- **Wrapping:** Consider the selected features as an additional hyper-parameter
  - computationally very heavy
  - most approximations are greedy algorithms
- **Embedding:** Include feature selection into parameter estimation through penalisation of the model coefficients
  - Naive form is equally computationally heavy as wrapping
  - **Soft-constraints** create biased but useful approximations

Regularised regression

Constrained and regularised regression

The optimization problem

$$\arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_q^q \leq t$$

for $q > 0$ is equivalent to

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q^q$$

when $q \geq 1$.

Constrained and regularised regression

The optimization problem

$$\arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_q^q \leq t$$

for $q > 0$ is equivalent to

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q^q$$

when $q \geq 1$. This is the **Lagrangian** of the constrained problem.

Note: Constraints are convex for all $q \geq 1$ but not differentiable in $\boldsymbol{\beta} = \mathbf{0}$ for $q = 1$.

Ridge regression

For $q = 2$ the constrained problem is **ridge regression** (Tikhonov regularisation)

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$.

Ridge regression

For $q = 2$ the constrained problem is **ridge regression** (Tikhonov regularisation)

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$.

An **analytical solution** exists if $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p$ is invertible

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$$

Ridge regression

For $q = 2$ the constrained problem is **ridge regression** (Tikhonov regularisation)

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$.

An **analytical solution** exists if $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p$ is invertible

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$$

If $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$, then

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \frac{\widehat{\boldsymbol{\beta}}_{\text{OLS}}}{1 + \lambda},$$

i.e. $\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)$ is **biased** but has **lower variance**.

SVD and ridge regression

Recall: The SVD of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

SVD and ridge regression

Recall: The SVD of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

The analytical solution for ridge regression becomes ($n \geq p$)

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$$

SVD and ridge regression

Recall: The SVD of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

The analytical solution for ridge regression becomes ($n \geq p$)

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} \\ &= (\mathbf{V}\mathbf{D}^2\mathbf{V}^\top + \lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y}\end{aligned}$$

SVD and ridge regression

Recall: The SVD of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

The analytical solution for ridge regression becomes ($n \geq p$)

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^\top + \lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{y}
\end{aligned}$$

SVD and ridge regression

Recall: The SVD of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

The analytical solution for ridge regression becomes ($n \geq p$)

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^\top + \lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \sum_{j=1}^{p} \frac{d_j}{d_j^2 + \lambda}\mathbf{v}_j\mathbf{u}_j^\top\mathbf{y}
\end{aligned}$$

SVD and ridge regression

Recall: The SVD of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

The analytical solution for ridge regression becomes ($n \geq p$)

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^\top + \lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \sum_{j=1}^{p} \frac{d_j}{d_j^2 + \lambda}\mathbf{v}_j\mathbf{u}_j^\top\mathbf{y}
\end{aligned}$$

Ridge regression **acts strongest** on principal components with **lower eigenvalues**, e.g. in presence of correlation between features.

Effective degrees of freedom

Recall the **hat matrix** $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ in OLS. The trace of $\mathbf{H}$

$$\mathrm{tr}(H) = \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = \mathrm{tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) = \mathrm{tr}(\mathbf{I}_p) = p$$

is equal to the trace of $\widehat{\boldsymbol{\Sigma}}$ and the **degrees of freedom** for the regression coefficients.

Effective degrees of freedom

Recall the **hat matrix** $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ in OLS. The trace of $\mathbf{H}$

$$\mathrm{tr}(H) = \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = \mathrm{tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) = \mathrm{tr}(\mathbf{I}_p) = p$$

is equal to the trace of $\widehat{\mathbf{\Sigma}}$ and the **degrees of freedom** for the regression coefficients.

In analogy define for ridge regression

$$\mathbf{H}(\lambda) := \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top$$

Effective degrees of freedom

Recall the **hat matrix** $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ in OLS. The trace of $\mathbf{H}$

$$\mathrm{tr}(H) = \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = \mathrm{tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) = \mathrm{tr}(\mathbf{I}_p) = p$$

is equal to the trace of $\widehat{\boldsymbol{\Sigma}}$ and the **degrees of freedom** for the regression coefficients.

In analogy define for ridge regression

$$\mathbf{H}(\lambda) := \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top$$

and

$$\mathrm{df}(\lambda) := \mathrm{tr}(\mathbf{H}(\lambda)) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda},$$

the **effective degrees of freedom**.

Lasso regression

For $q = 1$ the constrained problem is known as the **lasso**

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- Smallest q in penalty such that constraint is still convex
- Produces **sparse solutions** (many coefficients exactly equal to zero) and therefore performs **feature selection**

Intuition for the penalties (I)

Assume the OLS solution $\boldsymbol{\beta}_{\text{OLS}}$ exists and set

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\text{OLS}}$$

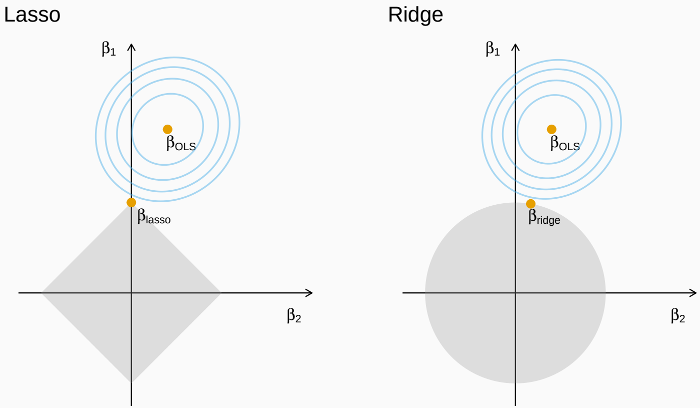
it follows for the **residual sum of squares (RSS)** that

$$\begin{aligned}
\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 &= \|(\mathbf{X}\boldsymbol{\beta}_{\text{OLS}} + \mathbf{r}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\
&= \|(\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{OLS}}) - \mathbf{r}\|_2^2 \\
&= (\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{OLS}})^\top \mathbf{X}^\top \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{OLS}}) - 2\mathbf{r}^\top \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{OLS}}) + \mathbf{r}^\top \mathbf{r}
\end{aligned}$$

which is an **ellipse** (at least in 2D) centred on $\boldsymbol{\beta}_{\text{OLS}}$.

Intuition for the penalties (II)

The least squares RSS is minimized for $\boldsymbol{\beta}_{\mathrm{OLS}}$. If a constraint is added ($\|\boldsymbol{\beta}\|_q^q \leq t$) then the RSS is minimized by the closest $\boldsymbol{\beta}$ possible that fulfills the constraint.



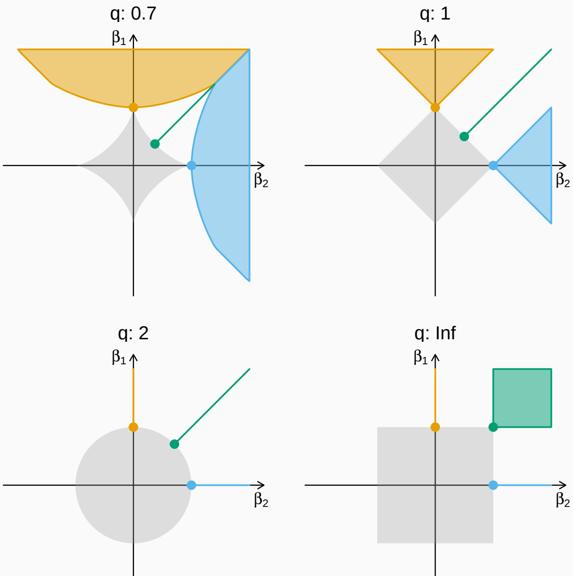
The blue lines are the contour lines for the RSS.

Intuition for the penalties (III)

Depending on q the different constraints lead to different solutions. If $\boldsymbol{\beta}_{\mathrm{OLS}}$ is in one of the coloured areas or on a line, the constrained solution will be at the corresponding dot.

Sparsity only for $q \leq 1$

Convexity only for $q \geq 1$



Computational aspects of the Lasso (I)

What estimates does the lasso produce?

Computational aspects of the Lasso (I)

What estimates does the lasso produce?

Target function

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Computational aspects of the Lasso (I)

What estimates does the lasso produce?

Target function

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Special case: $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$. Then

$$\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 = \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \underbrace{\mathbf{y}^\top\mathbf{X}}_{=\boldsymbol{\beta}_{\mathrm{OLS}}^\top}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 = g(\boldsymbol{\beta})$$

Computational aspects of the Lasso (I)

What estimates does the lasso produce?

Target function

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Special case: $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$. Then

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 = \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \underbrace{\mathbf{y}^\top\mathbf{X}}_{=\boldsymbol{\beta}_{\mathrm{OLS}}^\top}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 = g(\boldsymbol{\beta})$$

How do we find the solution $\widehat{\boldsymbol{\beta}}$ in presence of the **non-differentiable** penalisation $\|\boldsymbol{\beta}\|_1$?

Computational aspects of the Lasso (II)

For $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ the target function can be written as

$$\arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{p} -\beta_{\mathrm{OLS},j}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|$$

Computational aspects of the Lasso (II)

For $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ the target function can be written as

$$\arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{p} -\beta_{\mathrm{OLS},j}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|$$

This results in p **uncoupled** optimization problems.

Computational aspects of the Lasso (II)

For $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ the target function can be written as

$$\arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{p} -\beta_{\mathrm{OLS},j}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|$$

This results in p **uncoupled** optimization problems.

- **If** $\beta_{\mathrm{OLS},j} > 0$, then $\beta_j > 0$ to minimize the target
- **If** $\beta_{\mathrm{OLS},j} \leq 0$, then $\beta_j \leq 0$

Computational aspects of the Lasso (II)

For $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ the target function can be written as

$$\arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{p} -\beta_{\mathrm{OLS},j}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|$$

This results in p **uncoupled** optimization problems.

- **If** $\beta_{\mathrm{OLS},j} > 0$, then $\beta_j > 0$ to minimize the target
- **If** $\beta_{\mathrm{OLS},j} \leq 0$, then $\beta_j \leq 0$

Each case results in

$$\widehat{\beta}_{\mathrm{lasso},j} = \mathrm{sign}(\beta_{\mathrm{OLS},j})(|\beta_{\mathrm{OLS},j}| - \lambda)_+ = \mathrm{ST}(\beta_{\mathrm{OLS},j}, \lambda),$$

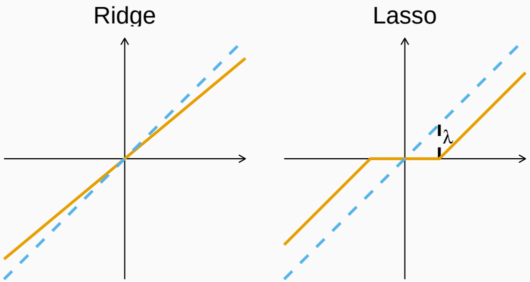
where

- $x_+ = x$ if $x > 0$ or 0 otherwise,
- and ST is called the **soft-thresholding operator**

Relation to OLS estimates

Both ridge regression and the lasso estimates can be written as functions of $\boldsymbol{\beta}_{\text{OLS}}$ if $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$.

$$\beta_{\text{ridge},j} = \frac{\beta_{\text{OLS},j}}{1+\lambda} \quad \text{and} \quad \widehat{\beta}_{\text{lasso},j} = \text{sign}(\beta_{\text{OLS},j})(|\beta_{\text{OLS},j}| - \lambda)_+$$



Visualisation of the transformations applied to the OLS estimates.

Shrinkage and effective degrees of freedom

When λ is fixed, the **shrinkage** of the lasso estimate $\boldsymbol{\beta}_{\text{lasso}}(\lambda)$ compared to the OLS estimate $\boldsymbol{\beta}_{\text{OLS}}$ is defined as

$$s(\lambda) = \frac{\|\boldsymbol{\beta}_{\text{lasso}}(\lambda)\|_1}{\|\boldsymbol{\beta}_{\text{OLS}}\|_1}$$

Note: $s(\lambda) \in [0, 1]$ with $s(\lambda) \to 0$ for increasing λ and $s(\lambda) = 1$ if $\lambda = 0$

Shrinkage and effective degrees of freedom

When λ is fixed, the **shrinkage** of the lasso estimate $\boldsymbol{\beta}_{\text{lasso}}(\lambda)$ compared to the OLS estimate $\boldsymbol{\beta}_{\text{OLS}}$ is defined as

$$s(\lambda) = \frac{\|\boldsymbol{\beta}_{\text{lasso}}(\lambda)\|_1}{\|\boldsymbol{\beta}_{\text{OLS}}\|_1}$$

Note: $s(\lambda) \in [0, 1]$ with $s(\lambda) \to 0$ for increasing λ and $s(\lambda) = 1$ if $\lambda = 0$

Recall: For ridge regression define

$$\mathbf{H}(\lambda) := \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^\top$$

and

$$\text{df}(\lambda) := \text{tr}(\mathbf{H}(\lambda)) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda},$$

the **effective degrees of freedom**.

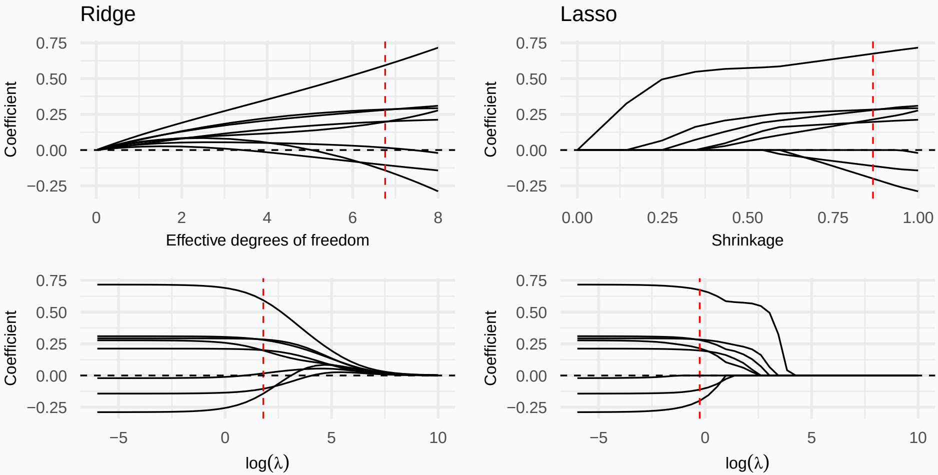
Prostate cancer dataset

Data to examine the correlation between the level of a prostate cancer-specific substance and a number of clinical measurements in men who just before partial or full removal of the prostate in patients.

- $n = 67$ samples
- A continuous response on the log-scale
- $p = 8$ features
  - e.g. log cancer volume, log prostate weight or age of patient

Regularisation paths for varying λ

Red dashed lines indicate the λ selected by cross-validation



- In the general case, i.e. $\mathbf{X}^\top \mathbf{X} \neq \mathbf{I}_p$, there is no explicit solution.

Notes on the lasso

- In the general case, i.e. $\mathbf{X}^\top \mathbf{X} \neq \mathbf{I}_p$, there is no explicit solution.
- Numerical solution possible, e.g. with **coordinate descent** where each β_j is updated separately with the remaining β_i with $i \neq j$ fixed

Notes on the lasso

- In the general case, i.e. $\mathbf{X}^\top \mathbf{X} \neq \mathbf{I}_p$, there is no explicit solution.
- Numerical solution possible, e.g. with **coordinate descent** where each β_j is updated separately with the remaining β_i with $i \neq j$ fixed
- As for ridge regression, **estimates are biased**

Notes on the lasso

- In the general case, i.e. $\mathbf{X}^\top\mathbf{X} \neq \mathbf{I}_p$, there is no explicit solution.
- Numerical solution possible, e.g. with **coordinate descent** where each β_j is updated separately with the remaining β_i with $i \neq j$ fixed
- As for ridge regression, **estimates are biased**
- **Degrees of freedom** are equal to the number of non-zero coefficients

Potential caveats of the lasso (I)

- **Sparsity of the true model:**
  - The lasso only works if the data is generated from a sparse process.
  - However, a dense process with many variables and not enough data or high correlation between predictors can be unidentifiable either way

Potential caveats of the lasso (I)

- **Sparsity of the true model:**
  - The lasso only works if the data is generated from a sparse process.
  - However, a dense process with many variables and not enough data or high correlation between predictors can be unidentifiable either way
- **Correlations:** Many non-relevant variables correlated with relevant variables can lead to the selection of the wrong model, even for large n

Potential caveats of the lasso (I)

- **Sparsity of the true model:**
  - The lasso only works if the data is generated from a sparse process.
  - However, a dense process with many variables and not enough data or high correlation between predictors can be unidentifiable either way
- **Correlations:** Many non-relevant variables correlated with relevant variables can lead to the selection of the wrong model, even for large n
- **Irrepresentable condition:** Split $\mathbf{X}$ such that $\mathbf{X}_1$ contains all **relevant variables** and $\mathbf{X}_2$ contains all **irrelevant variables**. If

$$|(\mathbf{X}_2^\top \mathbf{X}_1)(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}| < 1 - \eta$$

for some $\eta > 0$ then the lasso is (almost) guaranteed to pick the true model

Potential caveats of the lasso (II)

In practice, both the **sparsity of the true model** and the **irrepresentable condition** cannot be checked.

- Assumptions and domain knowledge have to be used

Take-home message

- Filtering and wrapping methods useful for feature selection in practice but can be unprincipled or have high variance
- Regularised regression can help in numerically unstable situations (such as in ridge regression)
- The lasso can in addition perform variable selection