Penalised regression

Ridge, LASSO and elastic net regression

COWIDUR

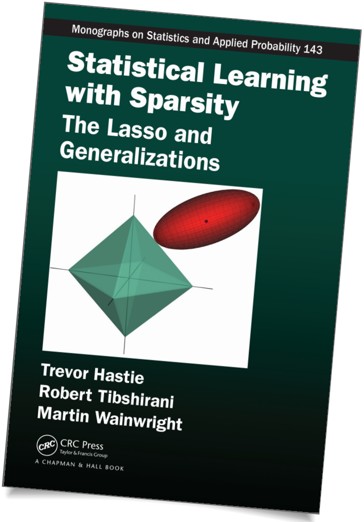
Torben Tvedebrink
tvede@math.aau.dk

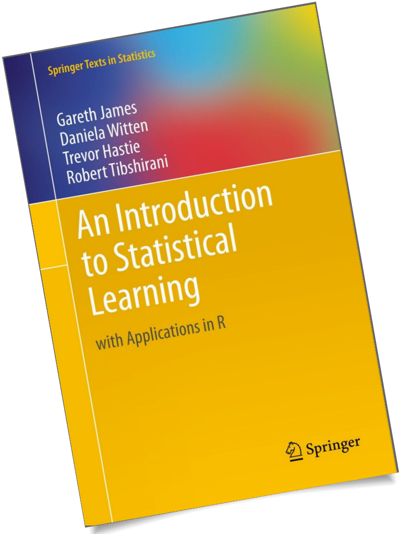
Department of Mathematical Sciences

AALBORG UNIVERSITY
DENMARK

Version: 08/05/2019 11:05

Literature
Freely available online





Bet on sparsity principle

Torben Tvedebrink
tvede@math.aau.dk

Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

When $p \gg n$ (the "short, fat data problem"), two things go wrong:

- The Curse of Dimensionality is acute.
- There are insufficient degrees of freedom to estimate the fullmodel.

However, there is a substantial body of practical experience which indicates that, insome circumstances, one can actually make good statistical inferences and predictions.

Our point of departure

In linear regression we assume that the ith response, y_i, can be modelled using a linear relationship between some covariates and the response with an additive error term with constant variance

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

Torben Tvedebrink
tvede@math.aau.dk

Our point of departure

In linear regression we assume that the ith response, y_i, can be modelled using a linear relationship between some covariates and the response with an additive error term with constant variance

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

If we have observations, $i = 1, \ldots, n > p$, we have that the least squares estimator for β_0 and $\beta = (\beta_1, \ldots, \beta_p)$ is given by

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0, \beta} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2$$

Least squares
On a *budget*

Torben Tvedebrink
tvede@math.aau.dk

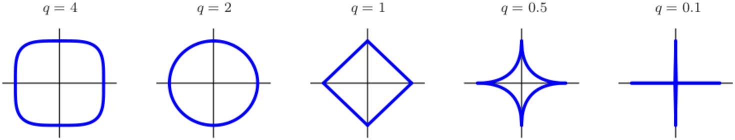
Imagine that we only had a limited *budget* of regression coefficients, t, such that the sum $\sum_{j=1}^{p} h(\beta_j)$ was restricted by t, then the solution should obey this constraint

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2 \quad \text{such that} \quad \sum_{j=1}^{p} h(\beta_j) \leq t$$

Imagine that we only had a limited *budget* of regression coefficients, t, such that the sum $\sum_{j=1}^{p} h(\beta_j)$ was restricted by t, then the solution should obey this constraint

$$\min_{\beta_0,\beta} \sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \quad \text{such that} \quad \sum_{j=1}^{p} h(\beta_j) \leq t$$

Constraint regions for $\sum_{j=1}^{p} h(\beta_j) = |\beta_j|^q \leq 1$.



For all $q < 1$ the contraint region is non-convex.

Penalised regression

Torben Tvedebrink
tvede@math.aau.dk

Least squares
On a *budget*

Imagine that we only had a limited *budget* of regression coefficients, t, such that the sum $\sum_{j=1}^{p} h(\beta_j)$ was restricted by t, then the solution should obey this constraint

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \quad \text{such that} \quad \sum_{j=1}^{p} h(\beta_j) \leq t$$

For

- $h(\beta_j) = |\beta_j|$ we term the regression problem the *LASSO*, and
- $h(\beta_j) = \beta_j^2$ we refer to the problem as *ridge regression*.

Torben Tvedebrink
tvede@math.aau.dk

Reasons for abandoning least squares

- The *prediction accuracy* can sometimes be improved because even though least squares has zero bias, its high variance may cause bad prediction ability. Hence, shrinking some coefficients, or setting the *noisy terms* to zero, may improve the accuracy.

Torben Tvedebrink
tvede@math.aau.dk

Reasons for abandoning least squares

- The *prediction accuracy* can sometimes be improved because even though least squares has zero bias, its high variance may cause bad prediction ability. Hence, shrinking some coefficients, or setting the *noisy terms* to zero, may improve the accuracy.
- The second reason is *interpretation*. The fewer terms to interpret the easier it gets.

Torben Tvedebrink
tvede@math.aau.dk

Reasons for abandoning least squares

- The *prediction accuracy* can sometimes be improved because even though least squares has zero bias, its high variance may cause bad prediction ability. Hence, shrinking some coefficients, or setting the *noisy terms* to zero, may improve the accuracy.

- The second reason is *interpretation*. The fewer terms to interpret the easier it gets.

- The third reason being that it fails for *wide* data, i.e. data for which $p \gg n$

Standardisation of $\mathbf{X}$

Torben Tvedebrink
tvede@math.aau.dk

As the *numerical value* of coefficients is sensitive to the scale of the covariates, it is typically preferred to standardise the $\mathbf{X}$ matrix before estimating the coefficients. That is,

$$\sum_{i=1}^{n} x_{ij} = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_{ij}^2 = n$$

Standardisation of $\mathbf{X}$
and centering of $\boldsymbol{y}$

Torben Tvedebrink
tvede@math.aau.dk

As the *numerical value* of coefficients is sensitive to the scale of the covariates, it is typically preferred to standardise the $\mathbf{X}$ matrix before estimating the coefficients. That is,

$$\sum_{i=1}^{n} x_{ij} = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_{ij}^2 = n$$

And in order to discard the intercept, β_0, from the regularisation in the case of linear regression we center the response

$$\sum_{i=1}^{n} y_i = 0$$

The *wide* data problem

Torben Tvedebrink
tvede@math.aau.dk

In the case where $p \gg n$, the least squares estimator is undefined as $(\mathbf{X}^\top \mathbf{X})$ isn't invertible because $\mathbf{X}$ is not of full rank. Hence, $\hat{\beta}^{\text{ols}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$ cannot be evaluated.

The *wide* data problem

Torben Tvedebrink
tvede@math.aau.dk

In the case where $p \gg n$, the least squares estimator is undefined as $(\mathbf{X}^\top\mathbf{X})$ isn't invertible because $\mathbf{X}$ is not of full rank. Hence, $\hat{\beta}^{\text{ols}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ cannot be evaluated.

A solution to this is to add an invertible matrix to $\mathbf{X}^\top\mathbf{X}$ to obtain an invertible matrix. The simplest such candidate is $\lambda\mathbf{I}_p$, for some positive $\lambda \in \mathbb{R}$:

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

which is what is referred to as the ridge regression estimator.

Ridge regression

For the least squares regression problem with a budget on the squared entries of β we have

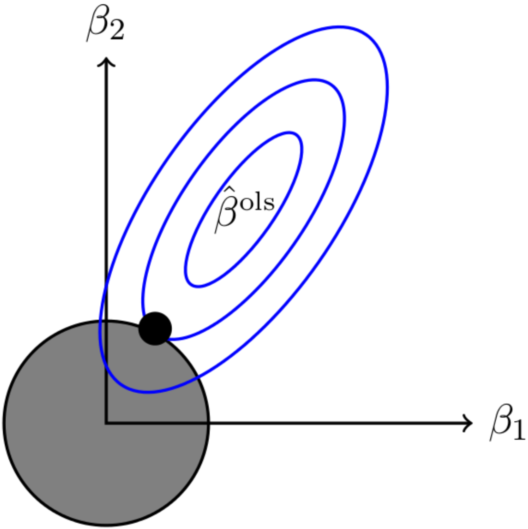
$$\min_{\beta} \sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\big)^2 \quad \text{such that} \quad \sum_{j=1}^{p} \beta_j^2 \leq t.$$

This can also be stated as

$$\min_{\beta} \sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Visual representation of $\hat{\beta}^{\text{ridge}}$

Compared to $\hat{\beta}^{\text{ols}}$ (in two dimensions)



LASSO regression

Now, what happens if we instead of using a squared penalty, β_j^2, uses the absolute penalty, $|\beta|$?

Well – we obtain the LASSO

$$\min_{\beta} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 \quad \text{such that} \quad \sum_{j=1}^{p} |\beta_j| \leq t.$$

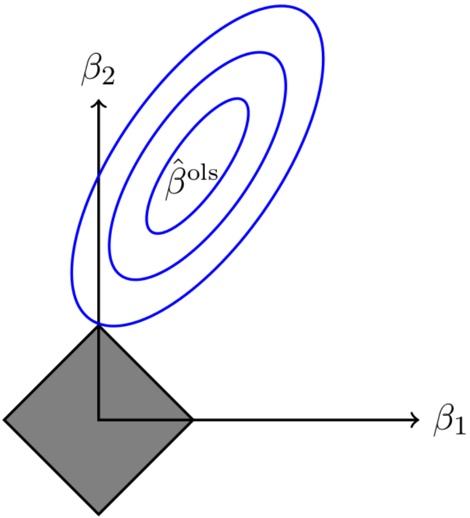
and again an equivalent form

$$\min_{\beta} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Torben Tvedebrink
tvede@math.aau.dk

Visual representation of $\hat{\beta}^{\text{lasso}}$

Compared to $\hat{\beta}^{\text{ols}}$ (in two dimensions)



LASSO solution

Comparison to Least Sqaures solution

With a standardized predictor, the LASSO solution is a soft-thresholded version of the ordinary least-squares (OLS) estimate $\hat{\beta}$

$$\hat{\beta}_j = \begin{cases} \hat{\beta}_j^{(\mathrm{OLS})} + \lambda, & \hat{\beta}_j^{(\mathrm{OLS})} < -\lambda \\ 0, & -\lambda \leq \hat{\beta}_j^{(\mathrm{OLS})} \leq \lambda \\ \hat{\beta}_j^{(\mathrm{OLS})} - \lambda, & \hat{\beta}_j^{(\mathrm{OLS})} > \lambda. \end{cases}$$

Torben Tvedebrink
tvede@math.aau.dk

LASSO solution

Comparison to Least Sqaures solution

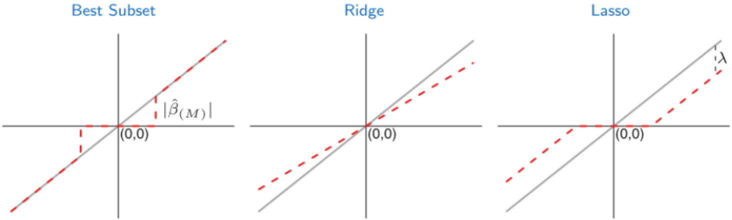
With a standardized predictor, the LASSO solution is a soft-thresholded version of the ordinary least-squares (OLS) estimate $\hat{\beta}$

$$\hat{\beta}_j = \begin{cases} \hat{\beta}_j^{(\mathrm{OLS})} + \lambda, & \hat{\beta}_j^{(\mathrm{OLS})} < -\lambda \\ 0, & -\lambda \leq \hat{\beta}_j^{(\mathrm{OLS})} \leq \lambda \\ \hat{\beta}_j^{(\mathrm{OLS})} - \lambda, & \hat{\beta}_j^{(\mathrm{OLS})} > \lambda. \end{cases}$$

This relationship also holds (in a slightly modified way) in case where the $\hat{\beta}^{(\mathrm{OLS})}$ do not exists.

Torben Tvedebrink
tvede@math.aau.dk

Soft thresholding
Modifications of the OLS estimates (if they exists)



Torben Tvedebrink
tvede@math.aau.dk

Elastic Net

The best from two worlds?

A downside with the Lasso is that it may have difficulties when several variables are collinear, such that linear combinations of them are hard to distinguish.

In such a case the Ridge Regression is better as it will typically form an average of the variables. Hence, for stable selection of variables in this case Ridge Regression may be preferred.

However, Ridge Regression seldom sets any parameters to zero, i.e. no variable selection which is what we would like in the end...

Torben Tvedebrink
tvede@math.aau.dk

Elastic Net
The best from two worlds?

The solution to the problem is Elastic Net, which incorporates both the Lasso and Ridge penalties in a convex way:

$$\min_{\beta} \sum_{i=1}^{2} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 + \lambda \sum_{j=1}^{p} \{\alpha|\beta_j| + (1-\alpha)\beta_j^2\},$$

where α is yet another tuning parameter deciding the amount of Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$) penalty that goes into the solution.

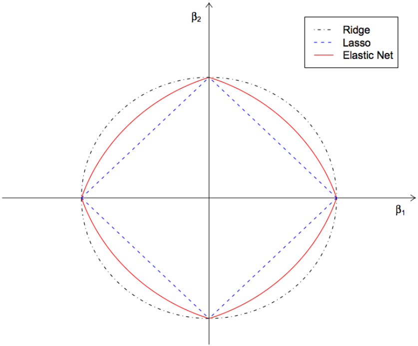
Both α and λ are selected based on cross-validation.

Torben Tvedebrink
tvede@math.aau.dk

Elastic Net

The best from two worlds?

In the Figure below we see the three types of regularisation discussed above. The shape of the Elastic Net solution area depends on α - the closer to 1 the more square it is, and the closer to 0 the more spherical.



A brief history of LASSO algorithms

And practical limits (in terms of number of covariates, p)

As mentioned earlier, the lasso penalty lacks a closed form solution in general.

As a result, optimisation algorithms must be employed to find the minimising solution

The historical efficiency of algorithms to fit lasso models can be summarized as follows:

Year	Algorithm	Operations	Practical limit
1996	QP†	$O(n^2 p)$	~ 100
2003	LARS‡	$O(np^2)$	$\sim 10,000$
2008	Coordinate descent	$O(np)$	$\sim 1,000,000$

†: Quadratic Programming

‡: Least Angle Regression

Torben Tvedebrink
tvede@math.aau.dk

The LASSO penalises each β_j coefficient individually by assessing the correlation between the partial residuals and the explanatory variable.

However, in the case of regression involving factors, the usual dummy variable encoding implies that the different derived dummy variables are penalised individually.

This causes some problems as we prefer that *all* dummy variables are set to zero, i.e. *all* levels of the factor are insignificant.
– This why we in ordinary regression use `anova(lm(...))` to test for significance of factors and not the individual t-tests reported in `summary(lm(...))`.

Torben Tvedebrink
tvede@math.aau.dk

Group LASSO
Adjusting the penalty

SLS use θ for the group LASSO in order to avoid confusion between the LASSO with penalty on the individual β parameters. Hence, we may reformulate the minimisation problem as

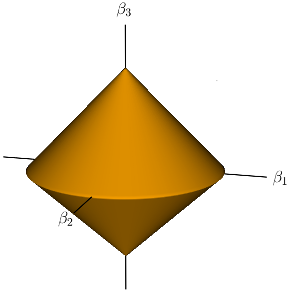
$$\min_{\theta_0,\theta}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{j=1}^{J}z_{ij}^{\top}\theta_j\right)^2+\lambda\sum_{j=1}^{J}\|\theta_j\|_2\right\},$$

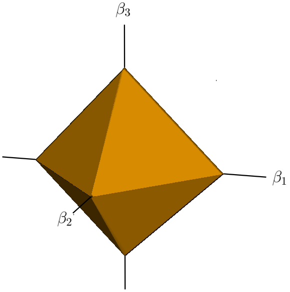
where $\|\theta_j\|_2=\sqrt{\sum_{k=1}^{p_j}\theta_{jk}^2}$ is the ℓ_2-norm.

For $p_j=1$ we have that $\|\theta_j\|_2=\sqrt{\theta_{j1}^2}=|\theta_{j1}|$, which is just the LASSO penalty.

For $p_j>1$, the ℓ_2-penalty will imply that either $\theta_j=0$ or non-zero.

The group LASSO ball
In $\mathbb{R}^3$



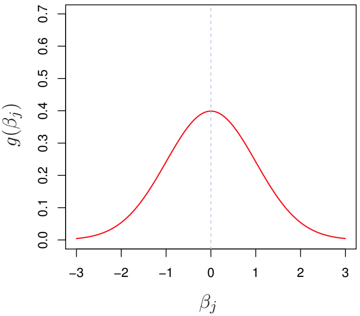


Left: Group LASSO; $\theta_1 = (\beta_1, \beta_2)$ and $\theta_2 = \beta_3$

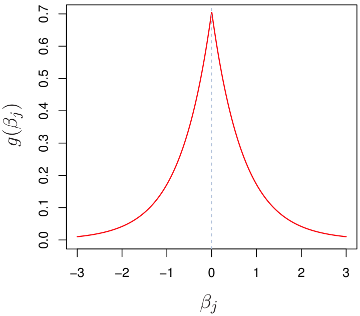
Right: LASSO; $\theta_j = \beta_j$

Torben Tvedebrink
tvede@math.aau.dk

Normal prior
Ridge regression



Laplacian prior
LASSO



Interpretation:
For the LASSO we *a priori* expect more parameters (due to the peaked nature of the Laplace prior) to be zero than for the normal ridge regression prior.

Bootstrap

Yet another handy application of the resampling technique

The bootstrap procedure may be one of the most important contributions to modern statistics – simple and yet powerful.

We may use the bootstrap as a non-parametric alternative to the Bayesian LASSO in order to assess the coefficient variability.

By *permuting the data* the parameter estimates may differ substantially. Hence, in order to capture this *repetition* of the experiment over and over again, resampling the data with replacement reflects this uncertainty.

Torben Tvedebrink
tvede@math.aau.dk

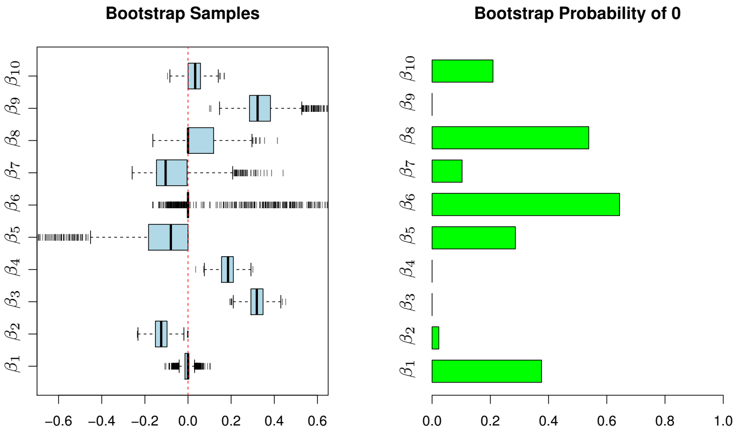


Figure 6.4 in SLS based on 1,000 bootstrap samples for fixed $\hat{\lambda}_{\text{CV}}$. *Left:* Parameter estimates $\hat{\beta}^*$; *Right:* "Significance" of each covariate.

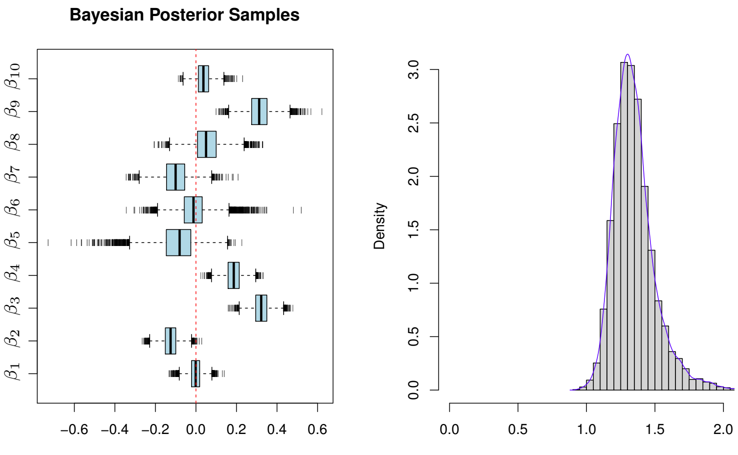


Figure 6.3 in SLS based on 10,000 samples from the posterior distributions (*left*) and $\|\beta\|_1$ (*right*).

Torben Tvedebrink
tvede@math.aau.dk

Example
Diabetes – parametric Bootstrapped parameters

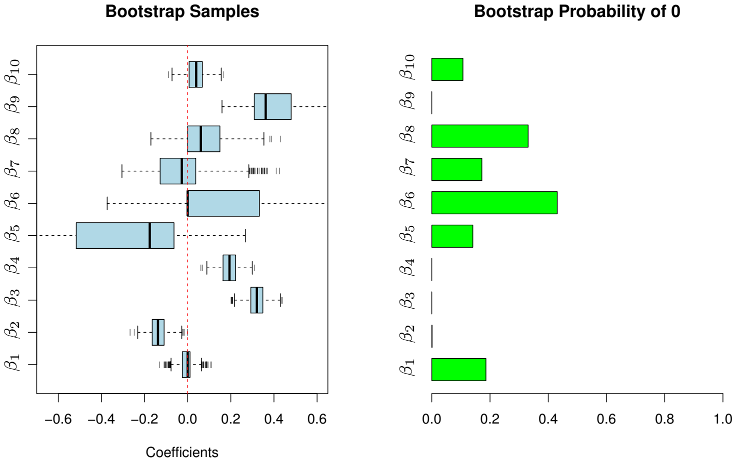


Figure 6.7 in SLS. $\boldsymbol{y}^*$ is sampled from the estimated model with $(\hat{\beta}, \hat{\sigma}^2)$ from the full data, and estimated for $(\boldsymbol{y}^*, \mathbf{X})$.