MA8701 Advanced methods in statistical inference and learning

L9: Lasso-variants for the linear model

Mette Langaas

Lectured 06.02.2023

~~2/5/23~~

Notes on DataAnProj ⇒ moved to wiki + email

(added after class in this colour)

Before we begin

Literature

- [HTW] Hastie, Tibshirani, Wainwrigh: “Statistical Learning with Sparsity: The Lasso and Generalizations”. CRC press. Ebook. Chapter 4.1-4.3,4.6

and for the interested student

- [WNvW] Wessel N. van Wieringen: Lecture notes on ridge regression Chapter 6.6
- [CASI] Efron and Hastie (2016) Chapter 16 (lasso in general)

Goal

The main goal of this part is to

- know about these special versions of the lasso (also in combination with the ridge), and

  and be able to use for data analysis!

- to see which practical data situation these can be smart to use.

  know when to use

Maybe one of these is suitable for the Data analysis project?

Theoretical properties and algorithmic details are not on the reading list.

Lasso and ridge

We have seen that the ridge regression shrinks the regression coefficients (as compared to the least squares solution), while the lasso regression both shrinks and sets some coefficients to zero (model selection).

Why do we need lasso variants

Why lasso variants needed?

assign similar value to these

- for (highly) correlated variables ridge does well, but lasso "kind of randomly" assigns coefficients among those variables

- categorical variables.
  - ordinal: may be ok to use one continuous covariate
  - nominal: k categories

dummy variable coding

$$\begin{array}{c|cccc} & c_1 & c_2 & \cdots & c_n \\ \text{obs } 1 & 1 & 0 & \cdot & 0 \\ 2 & 0 & 1 & 0 \cdots & 0 \\ & 0 & 0 & \cdots & 1 \\ \vdots & & & & \\ N & & & & \end{array}$$

no problem with intercept and k-dummy variables (will be for LS)

lasso: often only <u>some</u> of the dummy variables have nonzero coeffs.

We want all to be in or out!

↑

could also apply to other groupings of covariates

$Y = X\beta + \varepsilon$

l_q regression

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ (Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \quad \text{for } q \geq 0$$

$q=0$

$$\sum_{j=1}^{p} |\beta_j|^0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$$ also referred to as $\|\beta_j\|_0$

→ best subset selection → some true underlying model where some coefs ≠ 0.

$q=1$: lasso, also $\|\beta\|_1$

$q=2$: ridge also $\|\beta\|_2^2 = \sum_{j=1}^{p} |\beta_j|^2$:

Properties

called "bridge" regression for $q > 1$

- $0 \le q \le 1$: not differentiable
- $1 < q < 2$: in between lasso and ridge, but differentiable (and no variable selection property)
- q can be estimated from data, but according to Hastie, Tibshirani, and Friedman (2009) this is "not worth the effort for the extra variance incurred"

$\beta_1^4+\beta_2^4 \le 1$ $\beta_1^2+\beta_2^2 \le 1$ $|\beta_1|+|\beta_2| \le 1$ $\sqrt{|\beta_1|}+\sqrt{|\beta_2|} \le 1$

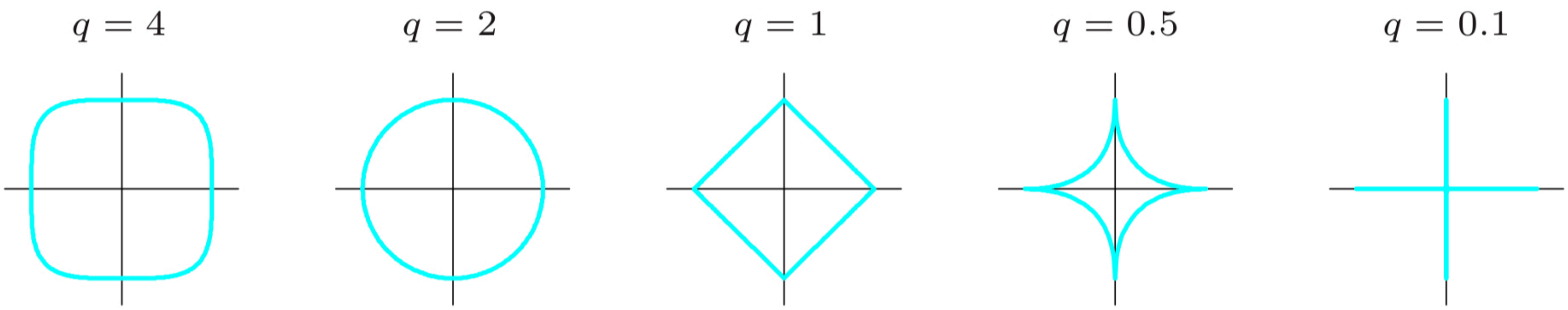


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Figure 1: Figure 3.12 from Hastie, Tibshirani, and Friedman (2009)

corner + edge q = 0.8

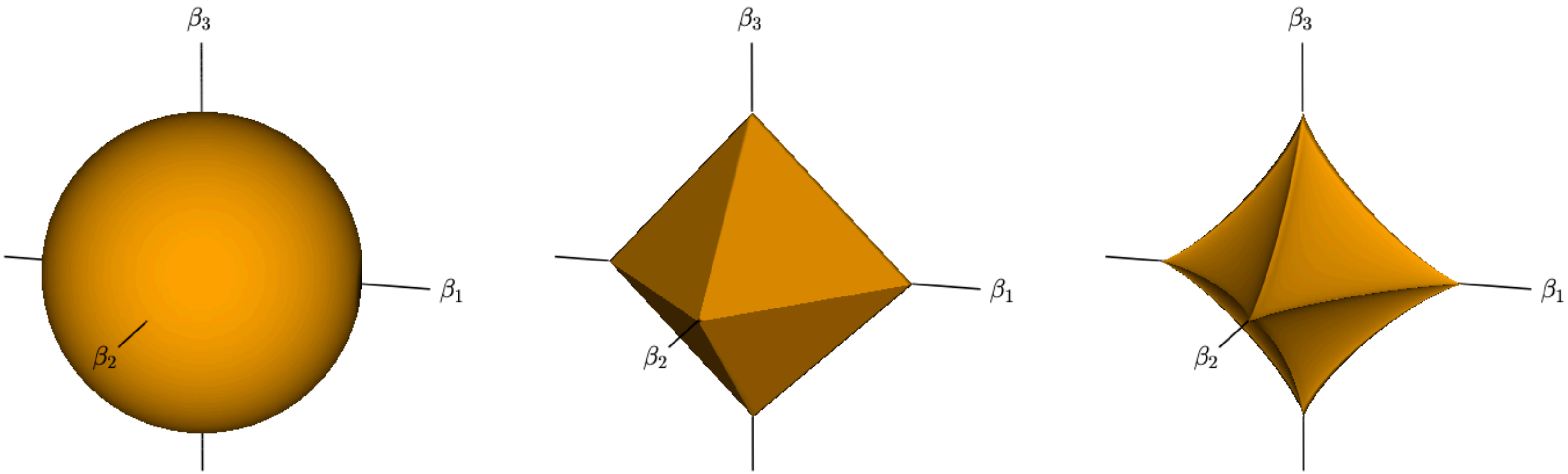


Figure 4.12 *The ℓ_q unit balls in $\mathbb{R}^3$ for $q = 2$ (left), $q = 1$ (middle), and $q = 0.8$ (right). For $q < 1$ the constraint regions are nonconvex. Smaller q will correspond to fewer nonzero coefficients, and less shrinkage. The nonconvexity leads to combinatorially hard optimization problems.*

Figure 2: Figure 4.12 from Hastie, Tibshirani, and Wainwright (2015)

OVERVIEW :

HTW ↓

Lasso $$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2 \right. \quad + \quad \lambda\sum_{j=1}^{p}|\beta_j|$$

$$(y-X\beta)^T(y-X\beta)$$

Ridge $$+ \quad \lambda\sum_{j=1}^{p}(\beta_j)^2$$

HTW ↓

Elasticnet $$\frac{1}{2}(y-X\beta)^T(y-X\beta) \quad + \quad \lambda\sum_{j=1}^{p}\left(\frac{1}{2}(1-\alpha)\beta_j^2+\alpha|\beta_j|\right)$$

Group lasso $$\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\theta_0-\sum_{j=1}^{J}z_{ij}^T\theta_j\right)^2+\lambda\sum_{j=1}^{J}\|\theta_j\|_2$$

Sparse group lasso $$\text{———} \; '' \; \text{———} \quad +\lambda\sum_{j=1}^{J}\left[(1-\alpha)\|\theta_j\|_2+\alpha\|\theta_j\|_1\right]$$

Elastic net

(HTW 4.2) Origin of method: Zou and Hastie (2005)

- Compromise between the ridge and lasso penalty.
- Lasso gives sparsity but does not handle correlated variables well.
- Ridge handles correlated variables well, but is not sparse.

Solution: *elastic net* which handles" coefficients of correlated features together (similar values or all zero).

The penalty used is now weighted sum of the ridge and the lasso penalty.

Elastic net

$$\begin{cases} \alpha = 0 : \text{ridge} \\ \alpha = 1 : \text{lasso} \end{cases}$$

$$\underset{\beta_0, \beta_1}{\text{minimize}} \left\{ \frac{1}{2} (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} \left(\frac{1}{2}(1-\alpha) \beta_j^2 + \alpha |\beta_j| \right) \right\}$$

WNJW: $\frac{1}{2} \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$

parameter constraint for $p=2$:

$$\{(\beta_1, \beta_2)\} \in \mathbb{R}^2 : \lambda_1 \left(|\beta_1| + |\beta_2| \right) + \frac{1}{2} \lambda_2 \left(\beta_1^2 + \beta_2^2 \right) \leq c(\lambda_1, \lambda_2)$$

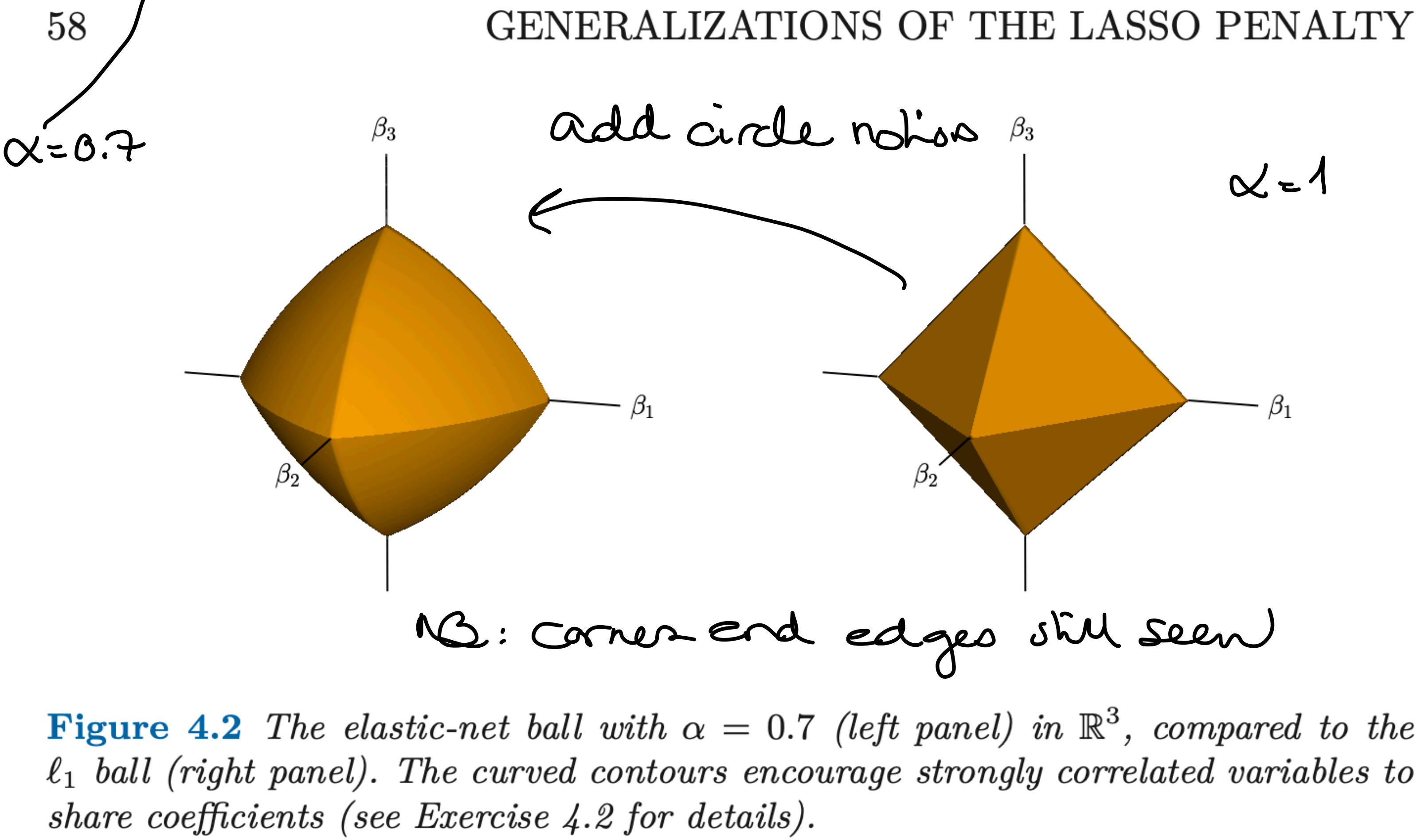
↑ corners

↑ selection property

↑ arch

$$\overset{0.3}{\underbrace{\lambda\cdot\tfrac{1}{2}(1-\alpha)}}\,\beta_j^2 + \overset{0.7}{\underbrace{\lambda\alpha}}\,|\beta_j|$$

$\alpha=0.7$



58 GENERALIZATIONS OF THE LASSO PENALTY

add circle notion

$\alpha=1$

NB: corners and edges still seen

Figure 4.2 *The elastic-net ball with $\alpha = 0.7$ (left panel) in $\mathbb{R}^3$, compared to the ℓ_1 ball (right panel). The curved contours encourage strongly correlated variables to share coefficients (see Exercise 4.2 for details).*

Figure 3: Figure 4.2 from Hastie, Tibshirani, and Wainwright (2015)

What is the elastic net parameter constraint region? Why will this give a variable selection property?

both penalty shrink

many best values for λ, α

NB edge!

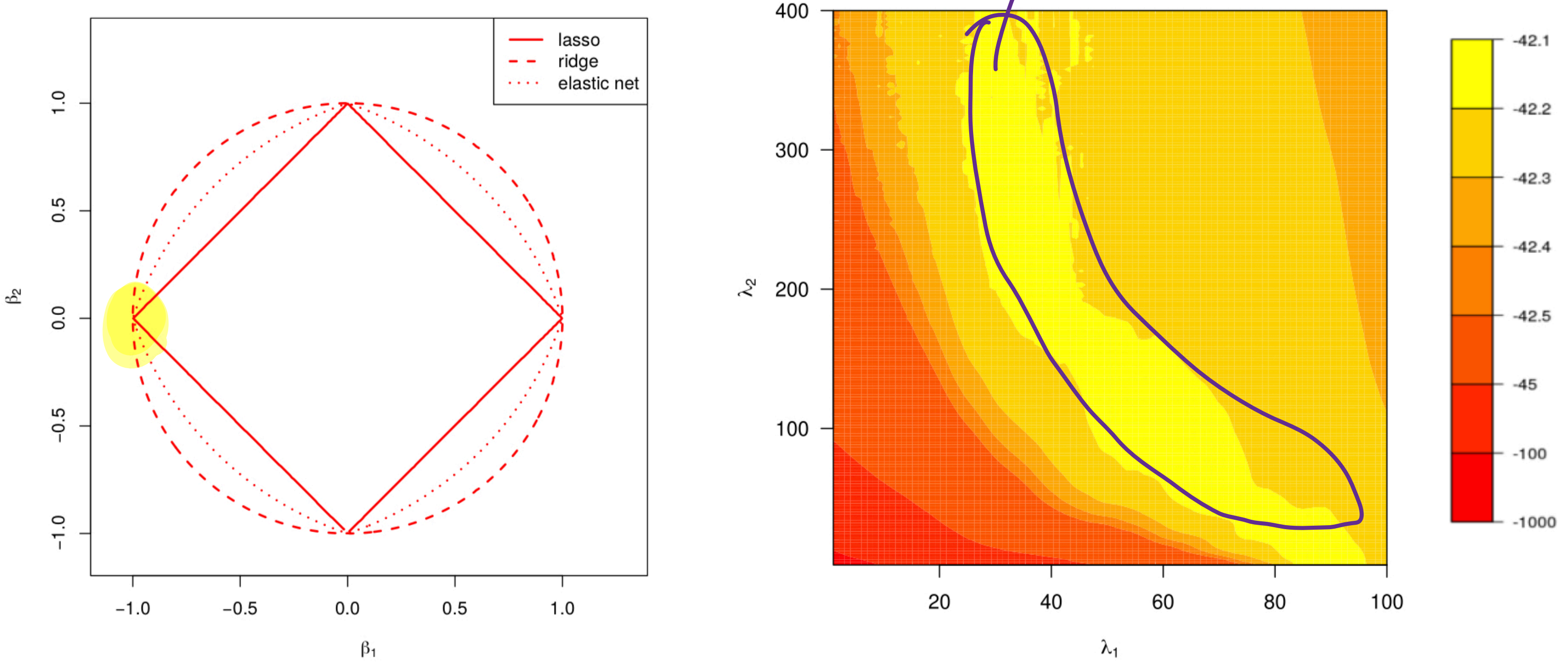


Figure 6.11: The left panel depicts the parameter constraint induced by the elastic net penalty and, for reference, those of the lasso and ridge are added. The right panel shows the contour plot of the cross-validated loglikelihood vs. the two penalty parameters of the elastic net estimator.

Figure 4: Figure 6.11 from Wieringen (2021)

Slightly different parametrization in Wieringen (2021), with $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$.

Now 2 hyperparam to choose!

The figure to the right shows potential problems in selecting the best hyperparameters. Observe that several combination of the two hyperparameters are equally good.

This is the reason for the parameterization with α as a mixing parameter, where the α is assumed to be set by the user, while the λ is found using cross-validation.

However, of cause α is a tuning parameter and need to be set. See for example the Master thesis of Lene Omdal Tillerli Chapter 3.5 and 5.3 for different cross-validation strategies for selecting the two hyperparameters.

Might be challenging to choose both α end λ!

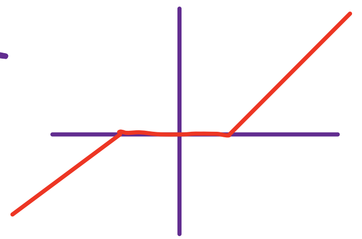
Remember from L8 the cyclic coord. descent?

→ V.s temasidne lin regr

$$\hat\beta_1 = \frac{\sum (x_{ij} - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

```
λ loop
    t = 0, .., convergence
        j = 1, .., p loop
            $r_{ij} = y_i - \sum_{k \neq j} x_{ik} \hat\beta_k^t$
```

$$\frac{\sum x_{ij} r_{ij}}{\sum x_i^2}$$



$$\min_{\beta_j} \frac{1}{2N} \underbrace{\sum_{i=1}^{N} (r_{ij} - x_{ij}\beta_j)^2}_{\hat\beta_{LS_j} \text{ solution to this problem}} + \lambda|\beta_j| \Rightarrow \hat\beta_j^{\text{lasso}} = \underbrace{\text{sign}(\hat\beta_{LS_j})(|\hat\beta_{LS_j}| - \lambda)_+}_{S_\lambda(\hat\beta_{LS_j})}$$

$\hat\beta_{LS_j}$ solution to this problem : $\hat\beta_{LS_j} = \dfrac{\sum_{i=1}^{N} x_{ij} r_{ij}}{\sum x_{ij}^2}$

The elastic net can be solved in a similar way because the added ridge penalty may be absorbed into the sq. loss.

Why do you think this can be solved in such a similar way as for the lasso?

There is a "data augmentation trick" where we can add p 0-reponses with covariates $\sqrt{\lambda(1-\alpha)}I_{pp}$ to perform a ridge regression (Wieringen (2021) 6.8.1).

Details are found in the article in the Journal on Statistical Software on the glmnet Friedman, Hastie, and Tibshirani (2010).

$$\rightarrow \underbrace{\|\tilde{y} - \tilde{X}\beta\|_2^2}_{\text{rewrite}} + \lambda \|\beta\|_1$$

LS $(X^TX)^{-1}X^TY$

ridge $(X^TX + \lambda(1-\alpha)I)^{-1}X^TY$

$$\hat{\beta}_{\text{ridge},j} = \frac{\sum \tilde{y}_i \tilde{x}_{ij}}{\sum \tilde{x}_{ij}^2 + \lambda(1-\alpha)}$$

$$\hat{\beta}_{\text{ridge},j} = \frac{\sum r_{ij} x_{ij}}{\sum x_{ij}^2 + \lambda(1-\alpha)}$$

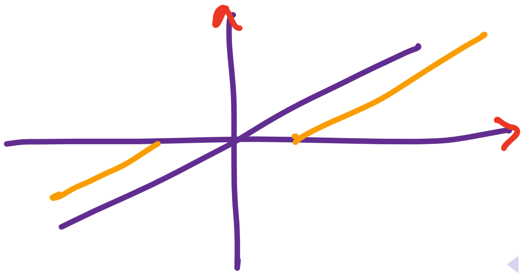
Parameter estimation

The *glmnet*-R package is constructed around the elastic net. Here the cyclic coordinate descent algorithm is used, and compared to the pseudo-algorithm we devised in class in L8, for the step with the update of β_j the soft-threshold solution is slightly modified to (HTW Equation 4.4)

penalty for lasso-part

$$\hat{\beta}_j = \frac{1}{\sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha)} S_{\lambda\alpha}\left(\sum_{i=1}^{N} r_{ij} x_{ij}\right)$$

where (as in L8) the soft thresholding operator is $S_\mu(z) = sign(z)(|z| - \mu)_+$ and the partial residual (as in L8) is $r_{ij} = y_i - \beta_0 - \sum_{k \neq j} x_{ik}\hat{\beta}_k$ (in L8 we used $\tilde{y}$ and not r).



Example

This example is shown in Figure 4.2 in HTW and reproduced with the R code below.

```
set.seed(8701)
N=100
z1=rnorm(N,0,1); z2=rnorm(N,0,1)
eps=rnorm(N,0,1)

y=3*z1-1.5*z2+2*eps

add=matrix(rnorm(N*6,0,1),ncol=6)
x1=z1+add[,1]/5; x2=z1+add[,2]/5; x3=z1+add[,3]/5
x4=z2+add[,4]/5; x5=z2+add[,5]/5; x6=z2+add[,6]/5

x=as.matrix(data.frame(x1=x1,x2=x2,x3=x3,x4=x4,x5=x5,x6=x6)
```

$Y = 3 \cdot Z_1 - 1.5 \cdot Z_2 + 2\varepsilon$

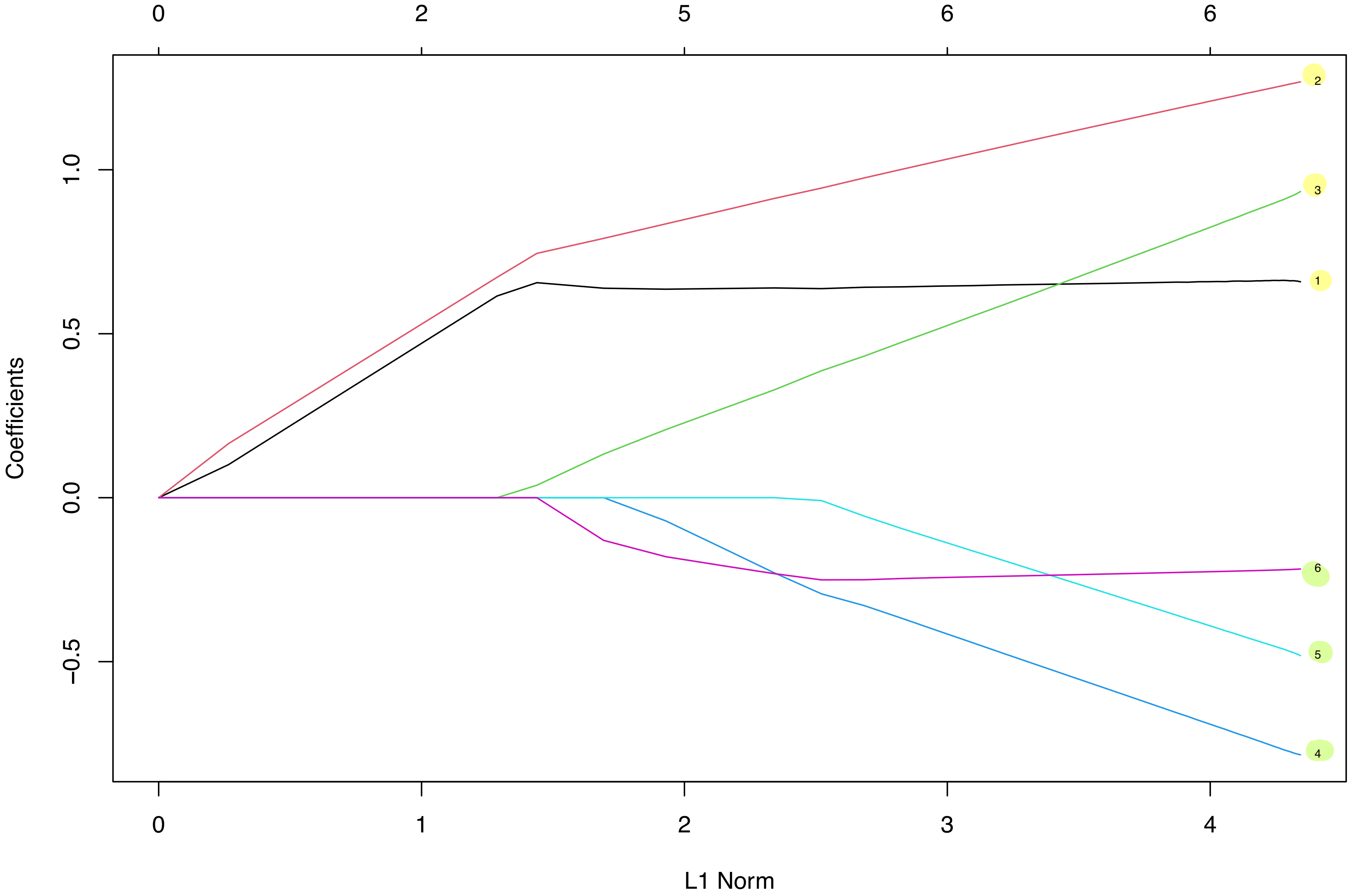
$N(0,1)$ $N(0,1)$ $N(0,1)$

x_1, x_2, x_3 x_4, x_5, x_6

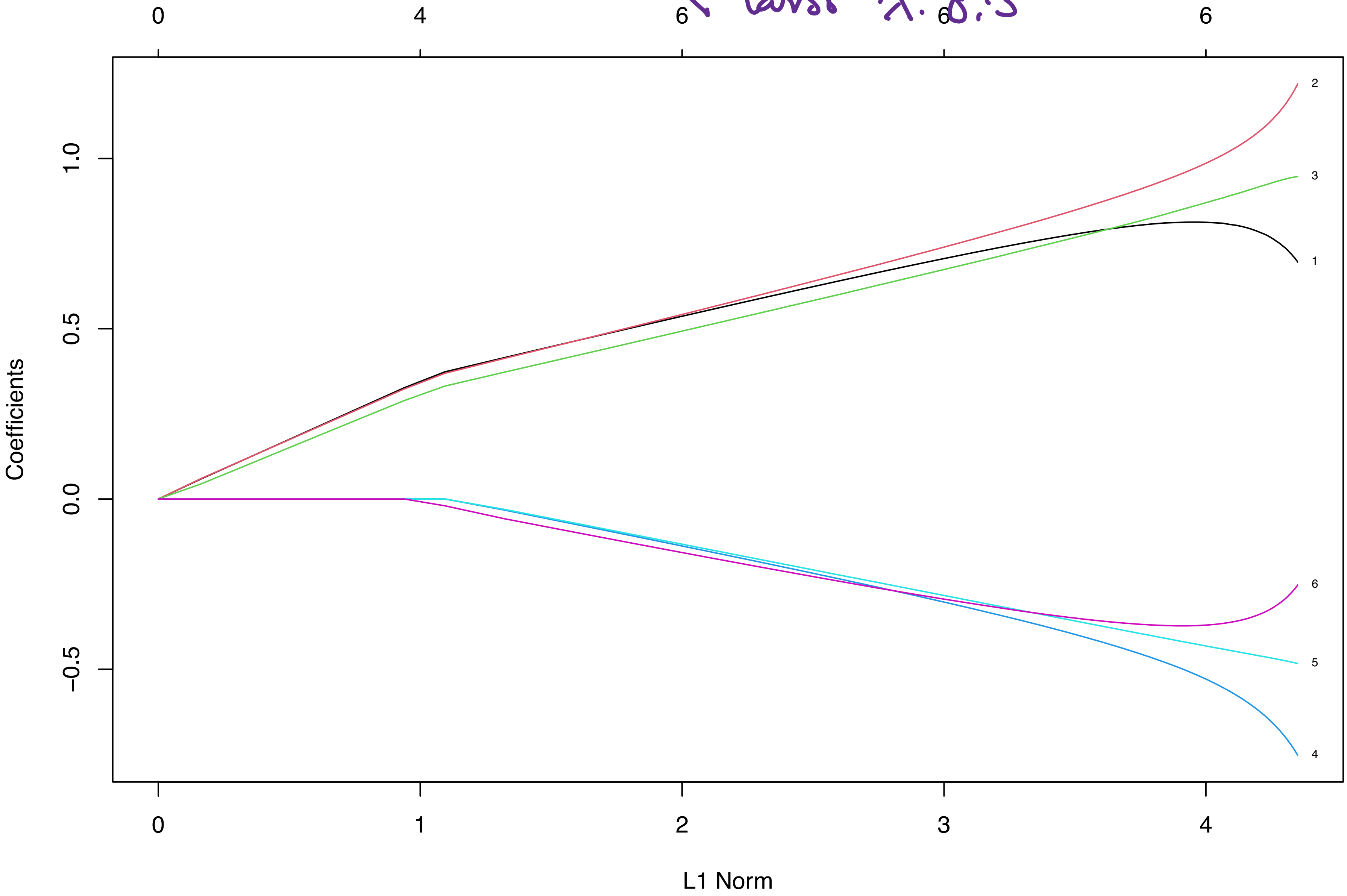
Z_1 + noise Z_2 + noise

$Y = f(x_1, x_2, \ldots x_6)$

LASSO ($\alpha=1$)

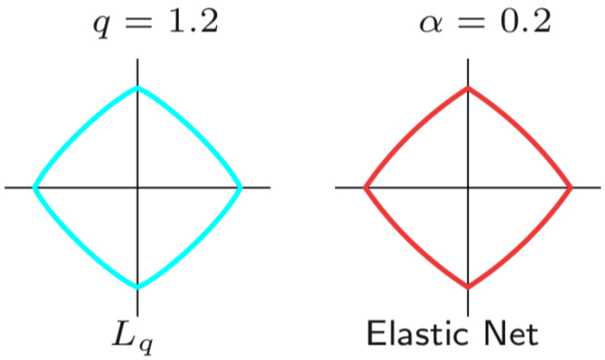


ELASTIC NET $\alpha = 0,3$ $\begin{cases} \text{ridge} \doteq \lambda \cdot 0.7 \\ \text{lasso } \lambda \cdot 0.3 \end{cases}$



Group discussion: Exam 2019 Problem 1c (STK-IN4300, UiO)

Briefly explain *elastic net* and *bridge regression* and explain why despite the corresponding constraints are almo~~u~~st indistinguishable in Figure 3.13 of Hastie, Tibshirani, and Friedman (2009), they provide, in general, quite different models.



$|\beta_j|^{1.2}$

$\lambda \cdot 0.8\,|\beta_j|$

FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j(\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

yet another parametrization ($\alpha\lambda|\beta_j|$ above)

Figure 5: Figure 3.13 from Hastie, Tibshirani, and Friedman (2009)

(and edges)

Elastic net has non-differentiable corners → sparse model because some β's → 0.

Bridge do not have non-diff corners

Exam 2020 Problem 1 (STK-IN4300, UiO)

Consider data simulated with the following setting:

- $\beta_i \sim N(0, 2), i = 1, \ldots, p$
- $X \sim N_p(0, \Sigma)$ where (i)$N_p(\cdot, \cdot)$ denotes a p-dimensional multivariate Gaussian distribution; (ii) 0 is a p-dimensional vector of 0; (iii) Σ is a $p \times p$ matrix with diagonal elements equal to 1 and all other elements equal to 0.9;
- $y = X\beta + \varepsilon$, with $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\varepsilon \sim N(0, 1)$.

a) If you were forced to choose between ridge regression and lasso, which one would you have used to predict y on a test set generated with the same setting? Why? Ridge – due to the corr. variables and if p is large

b) Would your choice have been the same if you ignored the first information on β ? Why? Lasso, maybe – better on sparsity if not true

c) Do you think that elastic net could have been a better choice in the situation of point (b)? Why? Handles corr. var. better + also shrinks

What is p? Ridge because of corr.

Group lasso

(HTW Section 4.3.1)

Now we aim at fixing the following problem with the lasso: if we have a factor and have used dummy variable coding, then the lasso may only choose to select some of the dummy variables to be in the model, and the lasso solution also is dependent on how the dummy variable encoding is done (choosing different contrasts will produce different solutions). Other application might have groups of genes in pathways, where those can be handled together.

The solution is to use a penalty that can be seen kind of intermediate to L_1 and squared L_2:

Group Lasso

J groups with p_j covariates in each group

$Z_j \in \mathbb{R}^{p_j}$

$\theta_j \in \mathbb{R}^{p_j}$

$$\{(y_i, \overset{\overset{\text{vec}}{\downarrow}}{z_{i1}}, z_{i2}, \dots, z_{iJ})\} \quad i=1,..,N$$

$$\underset{\substack{\theta_0, \theta_j \\ \uparrow \\ j=1,..,J}}{\text{minimize}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \theta_0 - \underbrace{\sum_{j=1}^{J} \underbrace{z_{ij}^T \theta_j}_{\text{sum}}}_{\text{sum}} \right)^2 + \lambda \underbrace{\sum_{j=1}^{J} \| \theta_j \|_2}_{\sqrt{\theta_1^T\theta_1} + \dots + \sqrt{\theta_J^T \theta_J}} \right\}$$

$$\sqrt{\theta_1^T\theta_1} \to \sqrt{\beta_1^2 + \beta_2^2 + \dots + \beta_{p_1}^2}$$

If $p_j = 1$: $\sqrt{\theta^T\theta} = \sqrt{\beta^2} = |\beta|$

$$\beta_1^2 + \beta_2^2 \le t \qquad \|\beta\|_2^2$$

$$\sqrt{\beta_1^2 + \beta_2^2} \le \sqrt{t} \qquad \|\beta\|_2$$

What does this new (unsquared) L_2 penalty do?

- All groups with one variable ends up with lasso L_1 penalty because: when $p_j = 1$ then $||\theta_j||_2 = |\theta_j|$, and thus the L_1 lasso penalty is used for singelton groups.
- All groups with more than two variables end up with the square root of the ridge penalty. since the penalty is $\sqrt{\sum_{j \in J} \beta_j^2}$ for all elements of this group J.

$\theta_1 = (\beta_1, \beta_2)$

$\theta_2 = \beta_3$

↑ lasso

all lasso

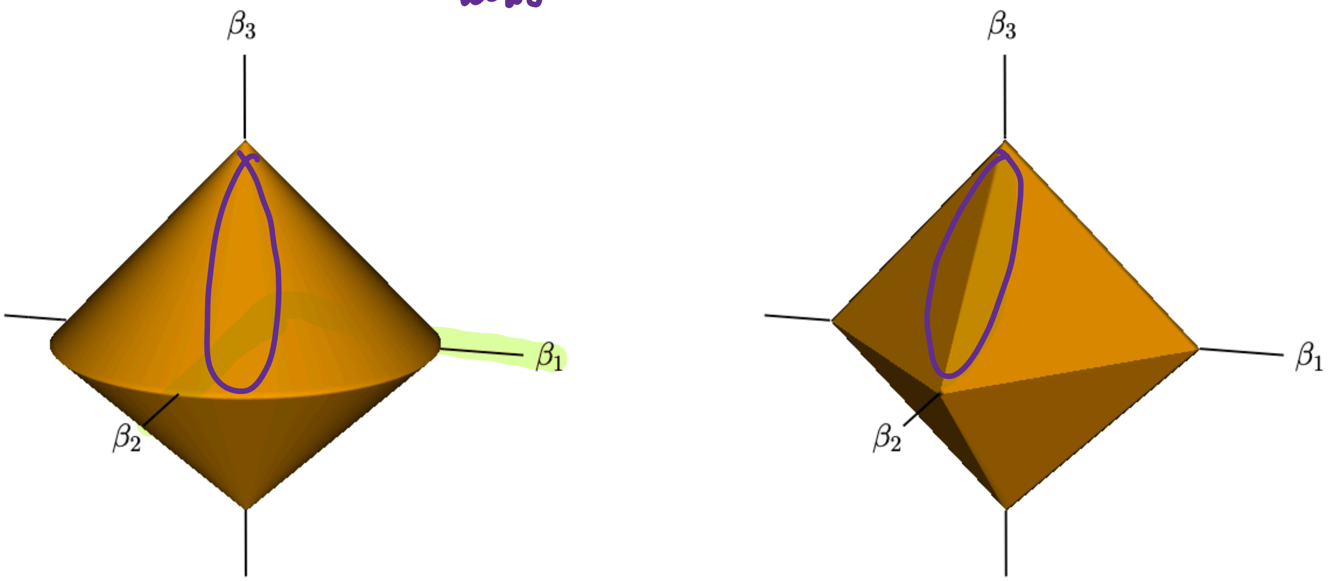


Figure 4.3 *The group lasso ball (left panel) in* $\mathbb{R}^3$*, compared to the* ℓ_1 *ball (right panel). In this case, there are two groups with coefficients* $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ *and* $\theta_2 = \beta_3 \in \mathbb{R}^1$.

Figure 6: Figure 4.3 from Hastie, Tibshirani, and Wainwright (2015)

Observe that the penalty is the same for all groups, independent of the group size- but it is common to also include the group size in the penalty (HTW does not, WNvW does).

HTW Exercise 4.4: the penalty term ensures that the coefficients in a group sum to zero given that there is an intercept term in the model.

Comment: some results are for orthogonal design matrices for a group. But, this will only happen if we have a balanced design, with the same number of observations for each level of a categorical variable group. This is very seldom the case in observational data, but in Design of Experiments this may happen for example in 2^k designs.

Dropp

Parameter estimation

The coordinate descent algorithm may be modified to a block coordinate descent version. The step to update $\hat{\theta}_j$ in the coordinate descent cyclic algorithm is

$$\hat{\theta}_j = (Z_j^T Z_j + \frac{\lambda}{||\hat{\theta}_j||_2} I)^{-1} Z_j^T r_j$$

where as earlier r_j is a partial residual.
If the designmatrix Z_j is ortogonal this is simplified to

$$\hat{\theta}_j = (1 \frac{\lambda}{||Z_j^T r_j||})_+ Z_j^T r_j$$

For non-ortogonal design matrices iterative methods are used.

Sparse group lasso

(HTW Section 4.3.2, WNvW Section 6.8.3)
The group lasso (with the Euclidean penalty) is now joined by the L_1 penalty. This is kind of similar to the elastic net now the squared L_2 penalty is replaced by L_2 penalty.

$$\text{minimize}_{\theta_0,\theta}\{\sum_{i=1}^{N}(y_i-\theta_0-\sum_{j=1}^{J} z_{ij}\theta_j)^2+\lambda\sum_{j=1}^{J}[(1-\alpha)\overset{\substack{\text{old group lasso}\\ \downarrow}}{||\theta_j||_2}+\alpha\overset{\substack{\text{new term}\\ \downarrow}}{||\theta_j||_1}]\}$$

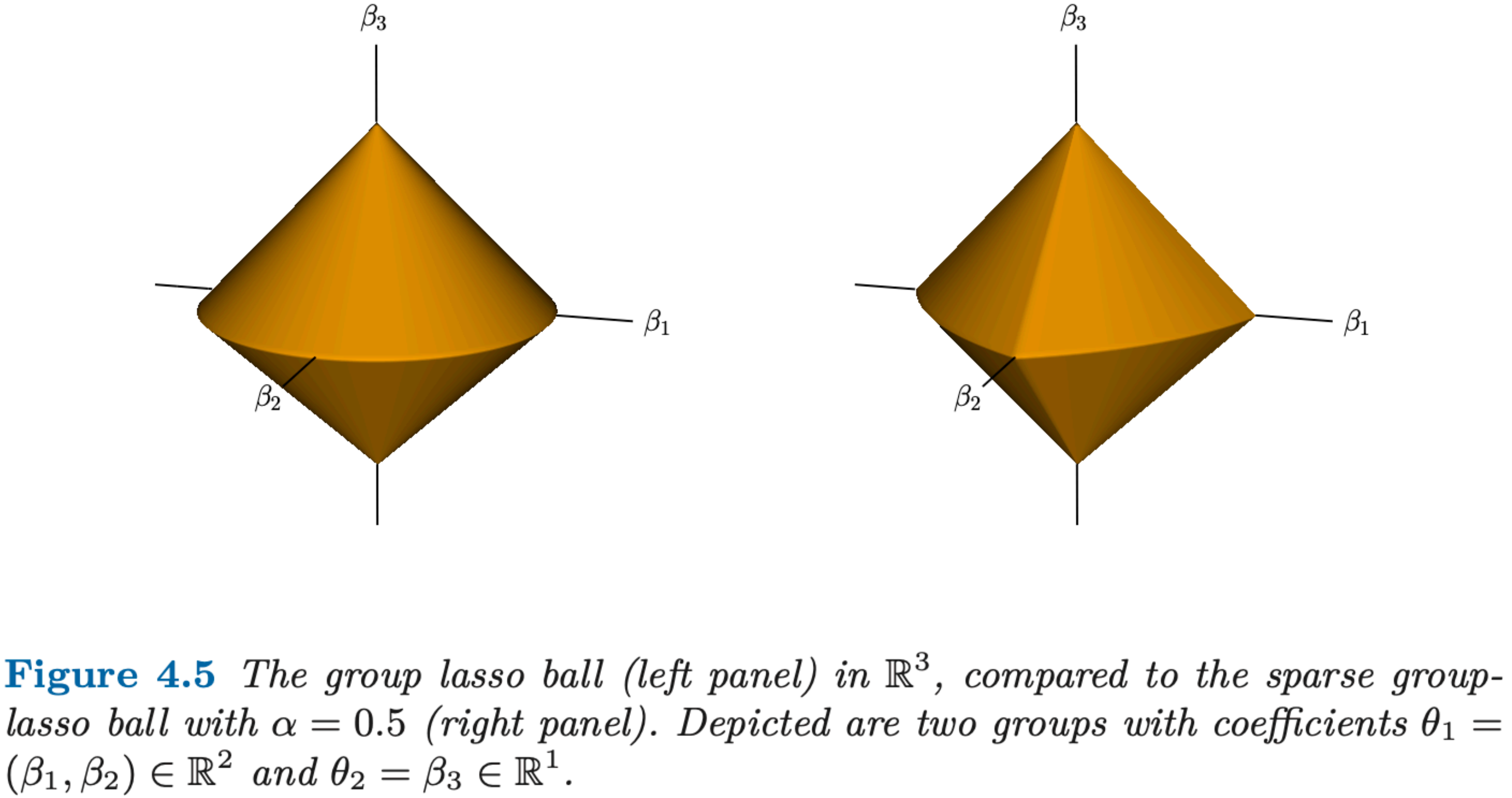


Figure 4.5 *The group lasso ball (left panel) in* $\mathbb{R}^3$*, compared to the sparse group-lasso ball with* $\alpha = 0.5$ *(right panel). Depicted are two groups with coefficients* $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ *and* $\theta_2 = \beta_3 \in \mathbb{R}^1$.

Figure 7: Figure 4.5 from Hastie, Tibshirani, and Wainwright (2015)

Parameter estimation

Again, a version of cyclic block-wise coordinate descent can be used.

The case when Z_j is not orthogonal requires more work than for orthonal versions.

Again, as for the elastic net, tuning the two hyperparametres may have several values being equally good.

Group discussion: Exam 2019 Problem 1b (STK-IN4300, UiO)

Consider the following version of the sparse group lasso:

$$\text{minimize}_{\beta_0,\beta}\{||y-\beta_0\vec{1}-\sum_{l=1}^{L}X_l\beta_l||_2^2+(1-\alpha)\lambda\sum_{l=1}^{L}\sqrt{p_l}||\beta_j||_2+\alpha\lambda||\beta||_1\}$$

where $\vec{1}$ denotes an N-dimensional vector of 1s, $\lambda \geq 0$ and $0 \geq \alpha \geq 1$. Answer the following questions:

- Why does β_0 only appear in the first term? – no need to shrink or set to 0
- What happens when $\alpha = 0$ and $\alpha = 1$, respectively?
- Briefly describe the concept of "bet on sparsity".

group lasso

lasso

Overlap group lasso

(HTW Section 4.3.3)

This is an extension to allow for a covariate to belong to more than one group.

The overlap group lasso "replicates a variable" in whatever group it is a member of, and then fits the group lasso to the problem.

The overlap group lasso can be used to ensure that interactions between covariates are only part of the model if the main effects of the covariates are in the model. See example HTW 4.3 for details.

$\{X_1,X_2\}$ $\{X_3\}$ $\{X_1,X_2\}$ $\{X_2,X_3)$

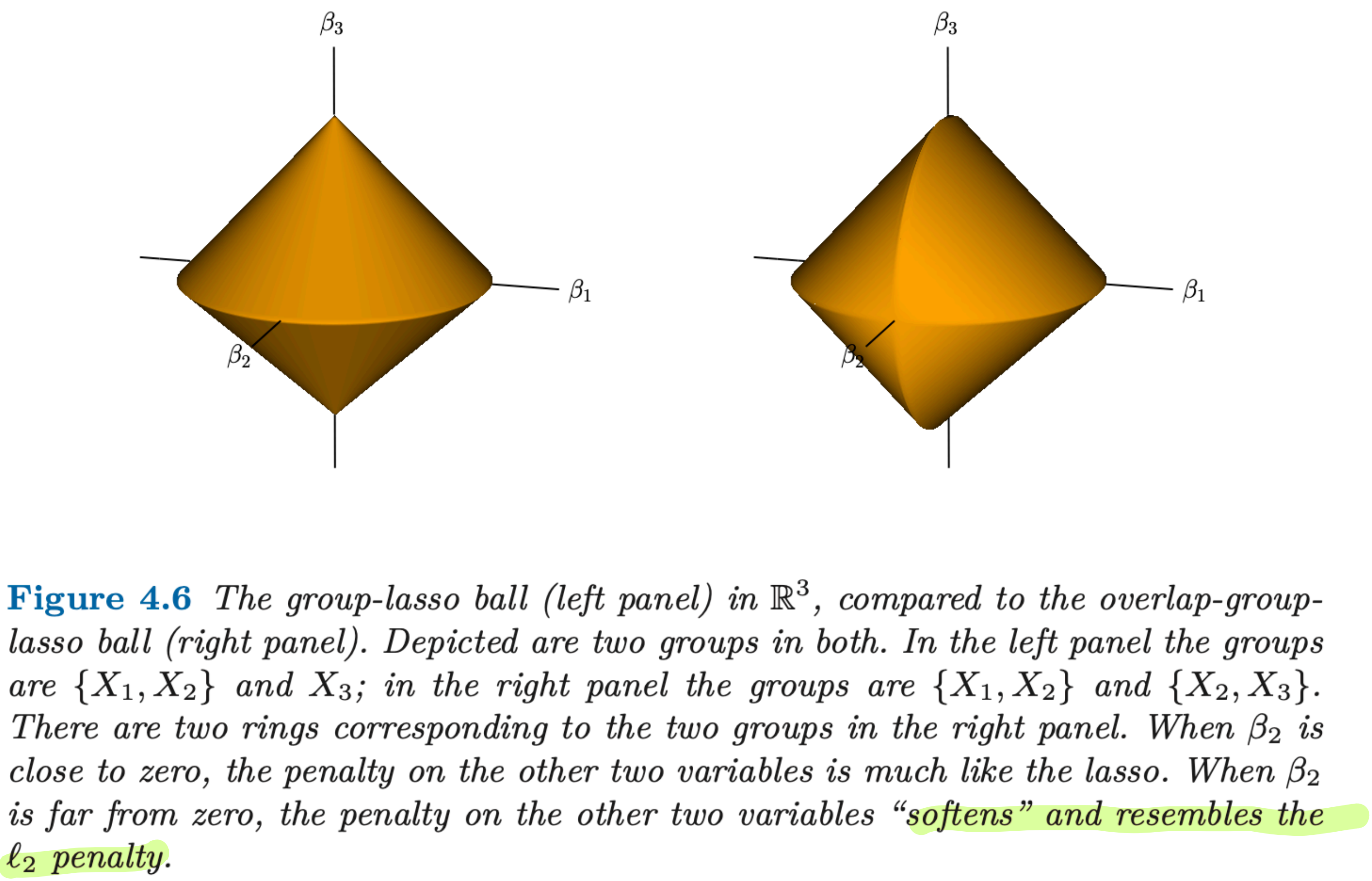


Figure 4.6 *The group-lasso ball (left panel) in* $\mathbb{R}^3$*, compared to the overlap-group-lasso ball (right panel). Depicted are two groups in both. In the left panel the groups are* $\{X_1, X_2\}$ *and* X_3*; in the right panel the groups are* $\{X_1, X_2\}$ *and* $\{X_2, X_3\}$*. There are two rings corresponding to the two groups in the right panel. When* β_2 *is close to zero, the penalty on the other two variables is much like the lasso. When* β_2 *is far from zero, the penalty on the other two variables "softens" and resembles the* ℓ_2 *penalty.*

Figure 8: Figure 4.6 from Hastie, Tibshirani, and Wainwright (2015)

Non-convex penalties

(HTW Section 4.5, WNvW Section 6.9)

We have looked at the l^q penalty formula in the start of L9:

$$\text{minimize}_{\beta_0,\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|^q\}, q \leq 0$$

Observe that if $0 \leq q \leq 1$ is non-convex.

For l^0 we aim for best subset selection and need to investigate 2^p possible models. This is not easy for $p > 40$.

The Smoothly Clipped Absolute Deviation SCAD method is an alternative to the l^q.

Adaptive lasso

(HTW Section 4.6, WNvW Section 6.8.4)
Origin: Zou (2006)
The aim is to fit models that are even sparser than the lasso. The method uses a so-called *pilot estimate* $\tilde{\beta}$:

$$\text{minimize}_{\beta_0,\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|\}$$

where $w_j = 1/|\tilde{\beta}_j|^\nu$ includes the pilot estimated, and given this pilot estimate the criterion i convex in β. The value of ν makes this an approximation to the l^q penalty where $q = 1 - \nu$.

Since the pilot estimate needs to be found first, this can be seen as a two-step procedure.

If $p < N$ then the least squares estimator can be used as the pilot estimate, and for larger p the ridge or lasso estimate may be used.

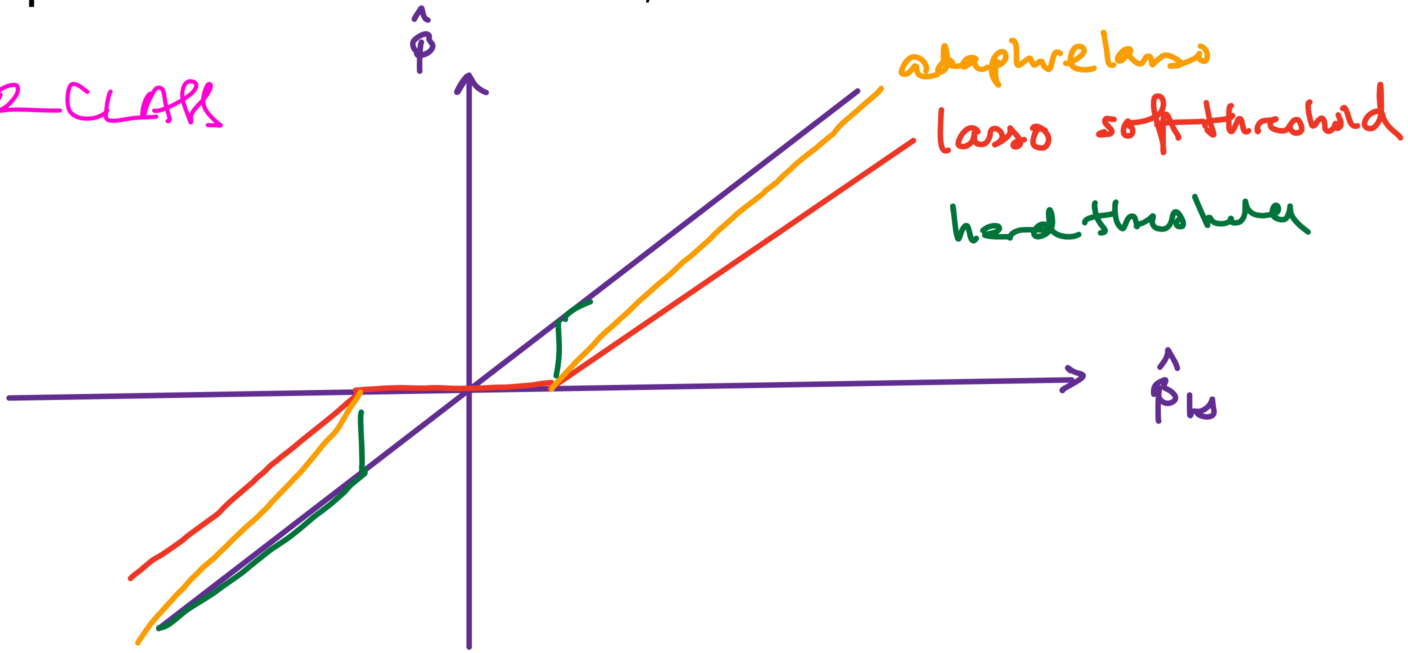
If $\tilde{\beta}_j = 0$ then the penalty of the jth element of the coefficient vector is infinity and the adaptive lasso estimate for the coeffisient will be zero.

For the orthogonal design, the adaptive lasso can be written as:

$$\hat{\beta}(\lambda) = sign(\hat{\beta}_{\mathsf{LS},j})(|\hat{\beta}_{\mathsf{LS},j}| - \frac{\lambda}{2\hat{\beta}_{\mathsf{LS},j}})_+$$

Compare this with the lasso, the difference is the last denominator.

AFTER CLASS



Unlike the lasso (according to Zou (2006)) the adaptive lasso is found to fulfill the *oracle property*.

According to Zou (2006), for an oracle procedure δ then $\beta(\delta)$ has the following properties:

- It identifies the right (correct) subset model, $\{j : \beta_j \neq 0\} = A$
- "Has the optimal estimation rate" $\sqrt{N}(\hat{\beta}(\delta)_A - \beta^*_A) \to_d N(0, \Sigma)$

If stepwise selection is used to find the active set, it can be trapped in local minima.

The continuous shrinkage property of the lasso is know to improve the prediction accuracy of the method (bias-variance trade-off).

The adaptive lasso can be estimated using the LARS algorithm of Efron et al (2004) (not covered in this course, but presented in Hastie, Tibshirani, and Friedman (2009) Section 3.4.4).

Back to forward stepwise model selection

If the aim is to minimize the squared loss with the l^0 penalty, the *forward stepwise model* method for model selection is efficent and “hard to beat”.

The forward stepwise model selection is a greedy algorithm.

- build a model sequentially by adding one variable at a time.
- At each step the best variable to include in the active set is identified and
- then the LS-fit is (re)computed for all the active variables.

This is an algorithm and not an optimization problem, and the theoretical properties of the algorithm “are less well understood” (HTW page 86).

A never ending story?

There seems to always be something that can be improved upon, and there are several lasso variantes that we have not discussed.

Other variants include

- The fused lasso (HTW Section 4.5)
- The random lasso

Group discussion

Choose one of the lasso/ridge variants we have covered in L8-L9 and write down:

- which variation on the classic lasso penalty is used (write down the penalty part of the minimization problem)
- make a drawing of the penalty (comparable to the sphere for ridge and the diamond for lasso)
- in which practical data analysis situation is this variation used (e.g. when many correlated variables are present, when the covariates have a natural group structure, ...)
- how can the parameter estimates be found?
- anything else you found interesting?