MSA220/MVE440 Statistical Learning for Big Data

Lecture 6 - Regularized regression

Rebecka Jörnsten

Mathematical Sciences
University of Gothenburg and Chalmers University of Technology

Recap from last lecture

We want to fit a regression model to our data using least squares.

$$y = X\beta + \epsilon$$

- y is our $n \times 1$ vector of outcome data
- X is our $n \times p$ design matrix
- ϵ is our $n \times 1$ vector with additive errors.
- For convenience, assume centered and standardized data.

Is this OK?
Reality check: 5 basic assumptions, scatter plots,....
TRANSFORMATIONS! ID EXTREME OUTLIERS!

When p is large or covariates in X are correlated, it is a tricky business to fit regression via OLS.
Why?

- $\min_\beta ||y - X\beta||^2$ has closed form solution
- $(X'X)^{-1}X'y$
- IF the inverse of $X'X$ exists.
- Not true if $p > n$. Inverse unstable if some covariates extremely correlated.

Regularized regression

What do we do?

- Reduce the number of covariates - prefiltering
- PCA of X and use only leading components.
- Partial least squares (more later)
- Regularized regression

Ridge regression

Regularization: To supress variance (due to instability of inverse of $X'X$), be willing to accept some bias!

- Ridge regression:
- $(X'X + \lambda I)^{-1}X'y$
- If $X'X = I$, this estimate $\beta_R = \beta_{OLS}/(1+\lambda)$ so biased but with lower variance
- If Xs are correlated, ridge regression *shrinkage* acts mostly on the directions with lower eigenvalues which correspond to the high variance estimates!
- See regression notes for more on this.

Ridge regression

An alternative formulation of the ridge regression problem through penalized least squares.
We want to minimize

$$||y - X\beta||^2$$

$$\text{subject to } ||\beta||_2^2 \leq \tau$$

I.e., try to minimize least squares but don't let the average β get too big...

- Lagrangian formulation: $\min_\beta \frac{1}{2}||y - X\beta||^2 + \lambda||\beta||_2^2$
- Take derivatives with respect to β
- $-X'(y - X\beta) + \lambda\beta = 0$
- Solution $\beta_R = (X'X + \lambda I)^{-1}X'y$
- Choose λ to make sure condition τ holds or more commonly, choose λ via Cross-validation

Ridge regression

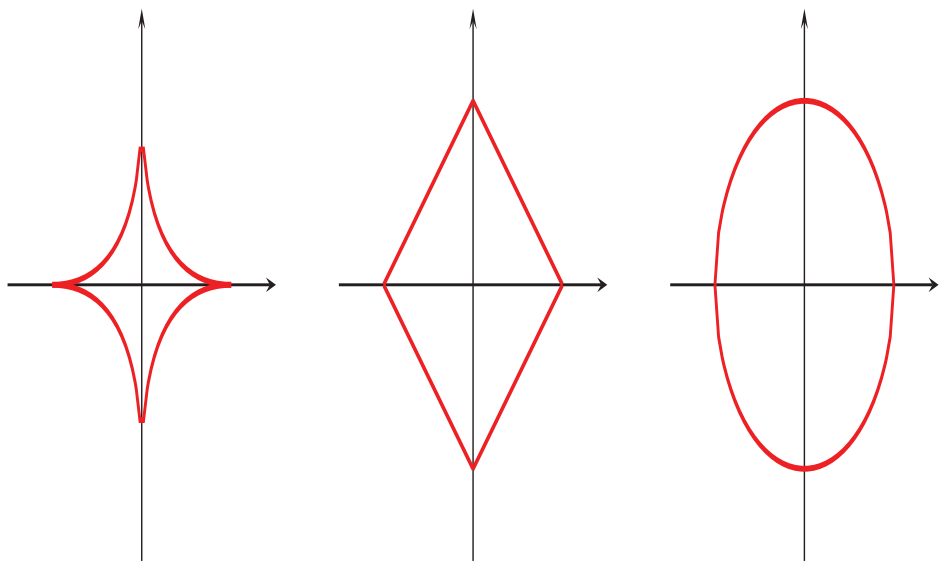
Pros and Cons?

- Pro: easy!
- Pro: can write other types of penalties here as well $\lambda\beta'\Omega\beta$ to penalize βs in a desired way
- Con: bias biggest for large coefficients
- Con: full model always returned since β_R may become very small but never exactly 0.

Lq-penalized regression

We can adress the growing bias and the lack of model interpretability using a different kind of penalty.

- L_0: $\min_\beta ||y - X\beta||^2 + \lambda \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$
- L_q: $\min_\beta ||y - X\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$
- L_1: $\min_\beta ||y - X\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|$
- L_1: $\min_\beta ||y - X\beta||^2 + \lambda \sum_{j=1}^{p} \beta_j^2$



Fraction $q < 1$, $q = 1$ and $q = 2$

Lq-penalized regression

$q = 0$ is the penalty that corresponds to optimal model selection, we only count the number of variables included in the model.
Pro: no bias. Con: since the penalty is non-convex it is very difficult to work with.
$q = 1$ is the smallest q that provides a convex penalty AND has the nice property of performing selection.
Why?

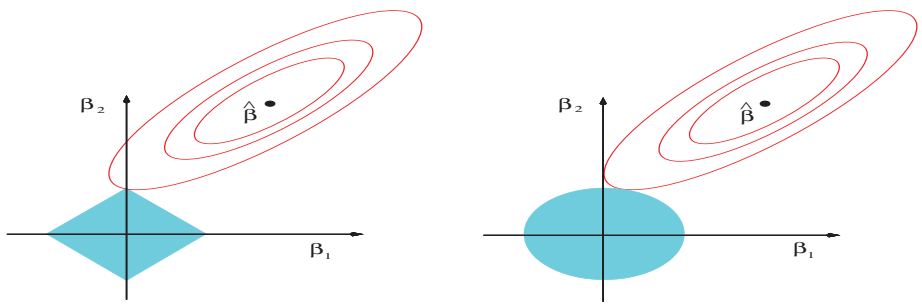


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

Because the $L1$ penalty has "singularities" (points) this makes the selection of solutions at those points more likely.
We will see this by solving the problem mathematically too, but think of this as the penalty region extremes being the most likely to lead to a solution that is optimal for the loss (model fit).

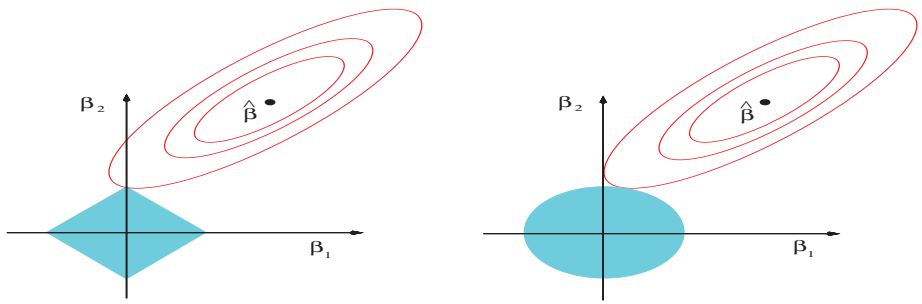


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

L1-penalized regression

$$\frac{1}{2}||y - X\beta||^2 + \lambda||\beta||_1$$

Consider first the special case $X'X = I$.

$$\frac{1}{2}y'y - y'X\beta + \frac{1}{2}\beta'\beta + \lambda||\beta||_1 = ***$$

Take derivatives with respect to β_j:

$$\frac{\partial ***}{\partial \beta_j} = -x_j'y + \beta_j + \lambda\nu_j$$

where

$$\nu_j = \begin{cases} sign(\beta_j) & if \quad \beta_j \neq 0 \\ \{\nu_j : |\nu_j| \leq 1\} & if \quad \beta_j = 0 \end{cases} \tag{1}$$

L1-penalized regression

$$\frac{\partial * **}{\partial \beta_j} = -x_j' y + \beta_j + \lambda \nu_j$$

where

$$\nu_j = \begin{cases} sign(\beta_j) & if \quad \beta_j \neq 0 \\ \{\nu_j : |\nu_j| \leq 1\} & if \quad \beta_j = 0 \end{cases} \tag{2}$$

So if $\beta_j > 0$, this is $\hat{\beta}_j = x_j' y - \lambda$ and if $\beta_j < 0$ this is $\hat{\beta}_j = x_j' y + \lambda$. There is a conflict between the assumed sign and the solution if $|x_j' y| < \lambda$. Note, $x_j' y = \hat{\beta}_j^{LS}$ for this special case $X'X = I$.
Solution:

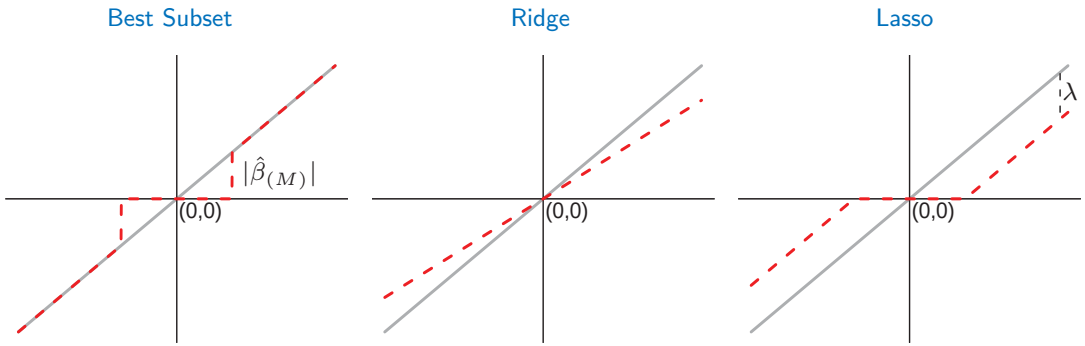
$$\hat{\beta}_j = \begin{cases} \beta_j^{LS} - \lambda & if \quad \beta_j^{LS} > \lambda \\ \beta_j^{LS} + \lambda & if \quad \beta_j^{LS} < -\lambda \\ 0 & if \quad |\beta_j^{LS}| < \lambda \end{cases} \tag{3}$$

This is called the *Soft Thresholding operation*, ST and we write

$$\hat{\beta}_j = ST(x_j' y, \lambda)$$

L1-penalized regression

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



This figure from the book illustrates the estimation of $L0, L1$ and $L2$ in relation to the LS estimate.
Notice the constant bias for $L1$ and growing bias for $L2$.

What about the general case? We can't solve this with a closed-form expression. But there are tons of ways of solving this, numerically and iteratively.
Here I illustrate *coordinate descent* which usually leads to simple steps and can be fast for big problems.

$$\min_{\beta} \frac{1}{2}||y - X\beta||^2 + \lambda||\beta||_1 = **$$

Take derivatives wrt β_j

$$\frac{\partial **}{\partial \beta_j} = -X_j'(y - \sum_{l \neq j} X_l \beta_l - X_j \beta_j) + \lambda \nu_j$$

We can write this as

$$-X_j' r_j + X_j' X_j \beta_j + \lambda \nu_j = 0$$

where r_j is the residual keeping the other $\beta_l, l \neq j$ fixed.
Now we have an expression that looks very much like our special case from before and so the solution is

$$\hat{\beta}_j = \frac{ST(X_j' r_j, \lambda)}{X_j' X_j}$$

If $\lambda = 0$ this is a iterative procedure for estimating LS coefficients which agrees with a simple updating scheme:

$$\hat{\beta}_j = X_j' r_j / X_j' X_j = \beta_j^{old} + X_j' r / X_j' X_j$$

Another popular method for solving constrained optimization problems is ADMM; *alternating directions method of multipliers*. This method decomposes the problem into separate convex optimization parts and then regularizes the solution differences. Here's the original problem:

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

The augmented lagrangian version looks like this:

$$\min_{\beta} \frac{1}{2}||y - X\beta||^2 + \lambda||\theta||_1 + u'(\beta - \theta) + \frac{\rho}{2}||\beta - \theta||^2$$

The lagrangian parameter u controls the differences between the two solutions β and θ and the term with ρ is a tuning term that can help speed up convergence. (Please read S. Boyd's excellent notes and papers on ADMM for more information).

$$\min_{\beta} \frac{1}{2}||y - X\beta||^2 + \lambda||\theta||_1 + u'(\beta - \theta) + \frac{\rho}{2}||\beta - \theta||^2$$

We now solve the problem by first minimizing wrt β, then θ and then updating u as
$u^{t+1} = u^t + \rho(\beta^{t+1} - \theta^{t+1})$ in iteration $t+1$. See how this controls the shrinkage towards equal solutions β and θ.
The β problem: take derivatives wrt β:

$$-X'(y - X\beta) + u + \rho(\beta - \theta) = 0$$

And therefore $\beta^{t+1} = (X'X + \rho I)^{-1}(X'y + \rho\theta^t - u^t)$ (which looks like a ridge solution update).
The θ problem: take derivatives wrt θ:

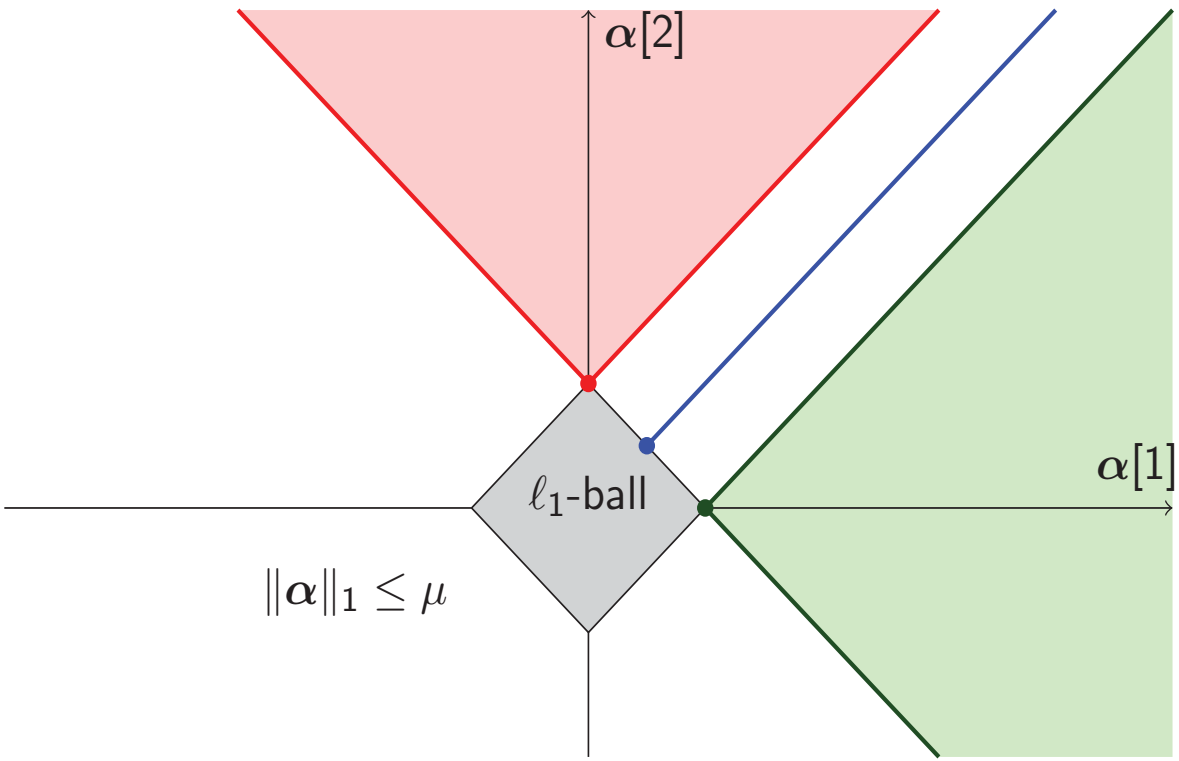
$$\lambda\nu - u^t - \rho(\beta - \theta) = 0$$

which gives us $\theta^{t+1} = ST(\beta^t + u^t/\rho, \lambda/\rho)$

So now you've seen that the L1-penalty induces model selection or *sparsity*. The reason is the non-smooth shape of the penalty region. Here follows some illustration from Julien Mairal's excellent slides.

Why does the ℓ_1-norm induce sparsity?

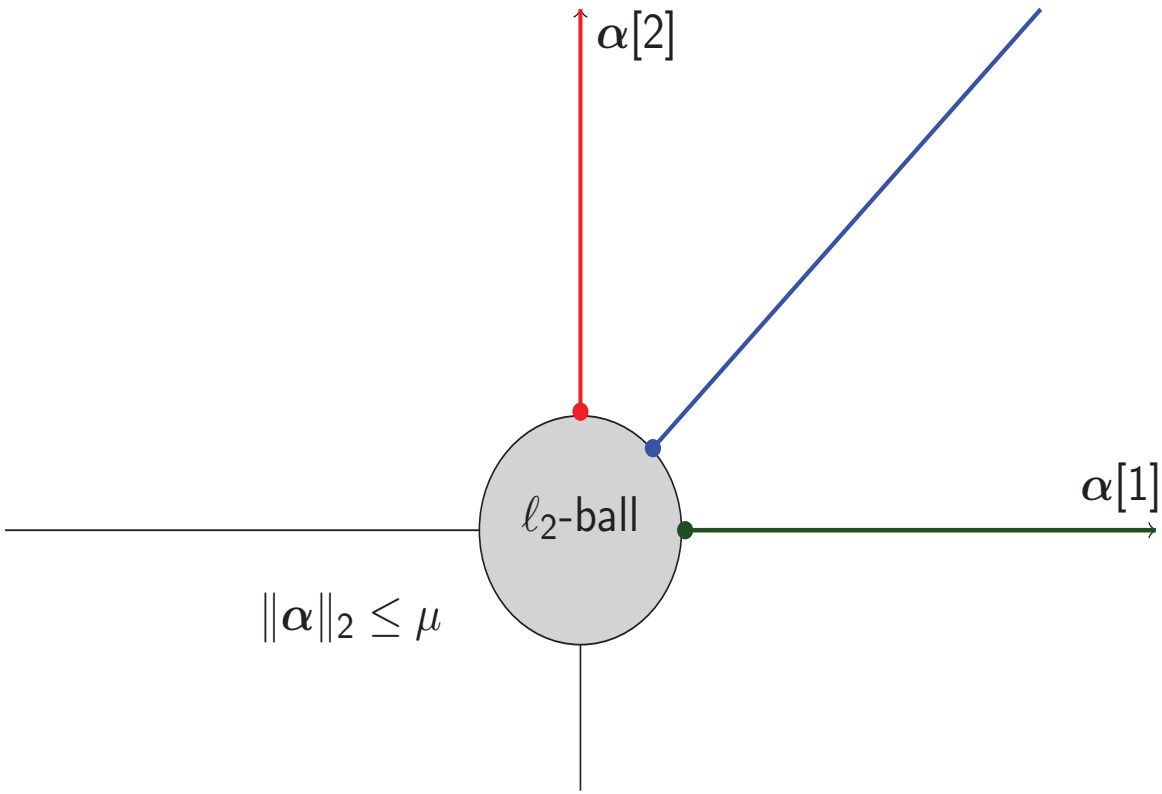
Regularizing with the ℓ_1-norm



The projection onto a convex set is "biased" towards singularities.

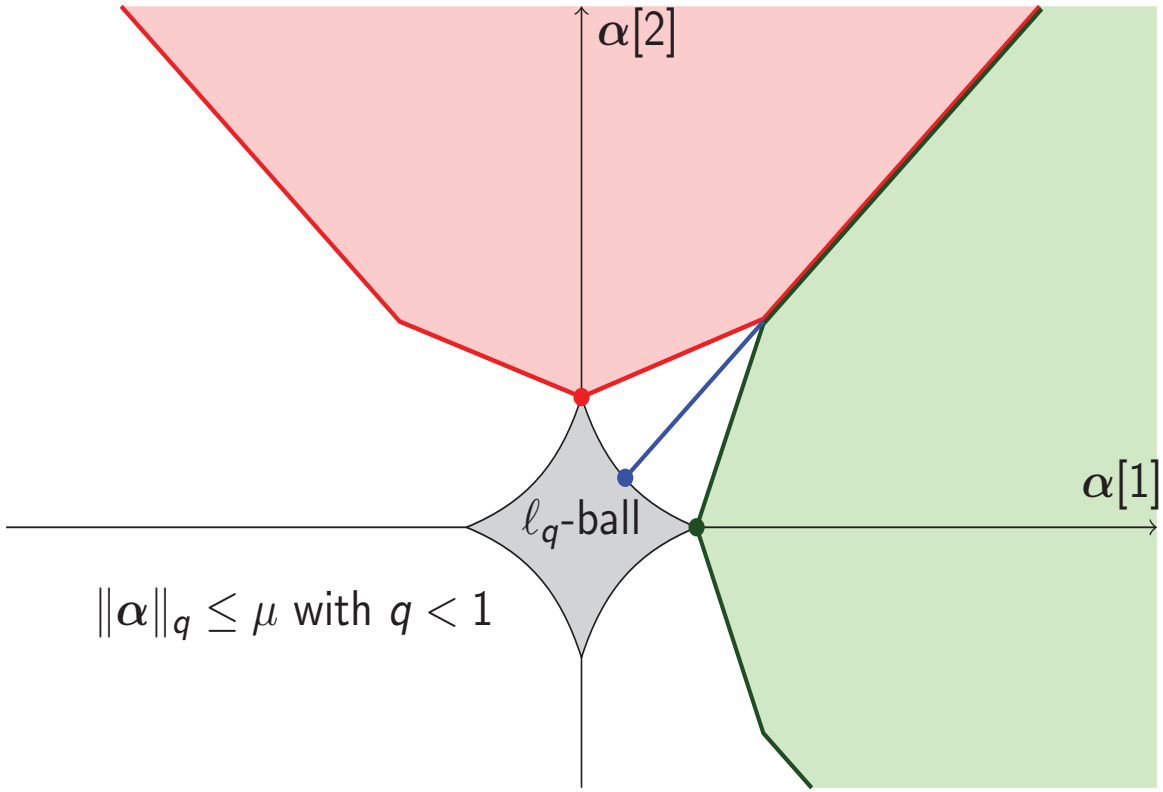
Why does the ℓ_1-norm induce sparsity?

Regularizing with the ℓ_2-norm



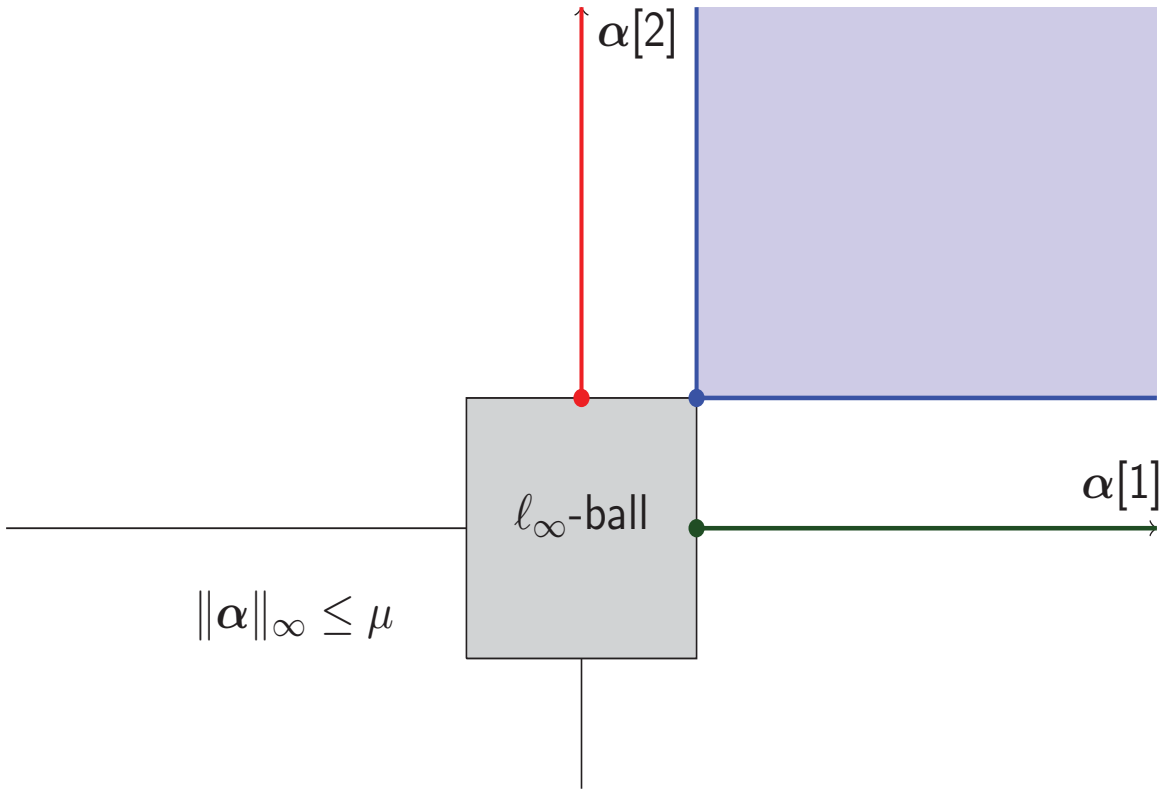
The ℓ_2-norm is isotropic.

Non-convex sparsity-inducing penalties



Why does the ℓ_1-norm induce sparsity?

Regularizing with the ℓ_∞-norm



The ℓ_∞-norm encourages $|\boldsymbol{\alpha}[1]| = |\boldsymbol{\alpha}[2]|$.

The point is that we can be a bit more adventurous about constructing penalty regions in order to generate a desired type of sparsity.

Group-penalized regression

Let's say there's a natural grouping of the variables: e.g. factor levels of a categorical level, source, etc. We want to include a group of variables and not select them separately.
We can achieve this by using a *group penalty* instead. Consider the case of K groups of variables:

$$\min_{\beta} \frac{1}{2}||y - X\beta||^2 + \lambda \sum_{k=1}^{K} ||\beta_k||_2$$

where $||\beta_k||_2 = \sqrt{\sum_{j \in k} |\beta_j|^2}$, i.e. we penalize the average β value within each group.

What's the effect of this penalty? We use coordinate descent at the group level to find out. Take derivatives wrt to group k

$$-X_k'(y - \sum_{j \ not \in k} X_j\beta_j - X_k\beta_k) + \lambda\nu_k = -X_k'(r_k - X_k\beta_k) + \lambda\nu_k$$

where r_k is the partial residual where we hold all β not in group k fixed and

$$\nu_k = \begin{cases} \frac{\beta_k}{||\beta_k||_2} & if \quad ||\beta_k||_2 \neq 0 \\ \{\nu : ||\nu||_2 \leq 1\} & if \quad ||\beta_k||_2 = 0 \end{cases} \tag{4}$$

If $||\beta_k||_2 = 0$ we get $-X_k' r_k + \lambda\nu = 0$
and so the condition for $||\beta_k||_2 = 0$ is $||X_k r_k||_2 \leq \lambda$
If $||\beta_k||_2 \neq 0$ we get $-X_k r_k + X_k' X_k \beta_k + \lambda \frac{\beta_k}{||\beta_k||_2}$
which starts to take on the familiar form...
and so the solution is

$$\beta_k = (X_k' X_k + \frac{\lambda}{||\beta_k||_2} I)^{-1} X_k r_k$$

if $||X_k r_k|| > \lambda$ and 0 otherwise.
Notice, this is like a soft threshold of a ridge estimate. The penalty for group k depends on the size of the previous iteration estimates. So this induces a group-sparsity. Why?

Group LASSO Ball

- If $D_k \in \mathbb{R}^{N \times 1}$, then $||X||_{2,1} = ||X||_1$
- That is, if all the groups are singletons, the optimization problem reduces to LASSO.
- Group LASSO ball shares attributes of both ℓ_2 and ℓ_1 balls.

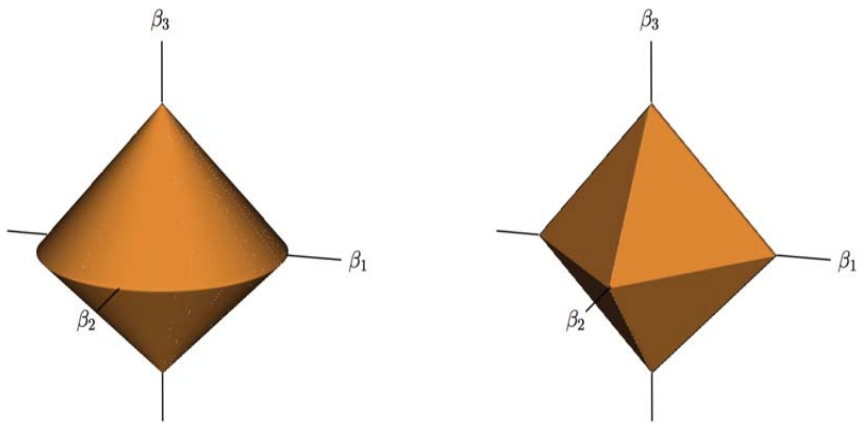


Figure 2: Group LASSO LASSO

From a talk by Hojjat Akhondi-Asl, UCL.

Elastic net

Lasso struggles when covariates are correlated and tends to pick only one of them even if both are related to the outcome. We can form groups of correlated variables and run group-lasso (Howard Bondell and others) or we can let the data decide for us and "helping" a bit by altering the penalty as follows:

$$\min_{\beta} \frac{1}{2}||y - X\beta||^2 + (1-\alpha)\lambda\frac{1}{2}||\beta||_2^2 + \alpha\lambda||\beta||_1$$

As you can see, this uses both an $L1$ and an $L2$ penalty on β. This penalty strategy is called the *elastic net*.

Elastic net

We go through the machinery again and arrive at a soft-thresholding solution

$$\beta_j = \frac{ST(X_j' r_j, \lambda\alpha)}{X_j' X_j + \lambda(1-\alpha)}$$

What tends to happen is that the bigger you make the $L2$ penalty (small α) the more elastic net with add groups of variables together into the model (see class demo).

We return for a moment to the group-penalty. What if we don't want the whole group but only "encourage" it but use fewer variables if possible?
We impose an additional sparsity constraint on the individual parameters too!
Consider the case of K groups of variables:

$$\min_{\beta} \frac{1}{2}||y - X\beta||^2 + \lambda(1-\alpha)\sum_{k=1}^{K}||\beta_k||_2 + \lambda\alpha||\beta||_1$$

This looks a bit like elastic net actually!

What's the effect of this penalty? We use coordinate descent at the group level to find out. Take derivatives wrt to group k

$$-X_k'(r_k - X_k\beta_k) + (1-\alpha)\lambda\nu + \alpha\eta$$

where r_k is the partial residual where we hold all β not in group k fixed and

$$\nu = \begin{cases} \frac{\beta_k}{||\beta_k||_2} & \text{if} \quad ||\beta_k||_2 \neq 0 \\ \{\nu : ||\nu||_2 \leq 1\} & \text{if} \quad ||\beta_k||_2 = 0 \end{cases} \tag{5}$$

$$\eta_j = \begin{cases} sign(\beta_j) & \text{if} \quad \beta_j \neq 0 \\ \{\eta : |\eta| \leq 1\} & \text{if} \quad \beta_j = 0 \end{cases} \tag{6}$$

If $||\beta_k||_2 = 0$ we get $-X_k' r_k + \lambda(1-\alpha)\nu + \alpha\lambda\eta = 0$
and so the condition for $||\beta_k||_2 = 0$ is $ST(X_k r_k, \alpha\lambda) \leq (1-\alpha)\lambda$
If $||\beta_k||_2 = 0$ and $\beta_j = 0$ we get the condition $||X_j r_j|| \leq \alpha\lambda$.
If $||\beta_k||_2 \neq 0$ and $\beta_j \neq 0$ we get

$$-X_j r_j + X_j' X_j \beta_j + \lambda(1-\alpha)\frac{\beta_k}{||\beta_k||_2} + \lambda\alpha\eta$$

which starts to take on the familiar form...
and so the solution is

$$\beta_j = \frac{ST(X_j' r_j, \lambda\alpha)}{X_j' X_j + (1-\alpha)\lambda/||\beta_k||_2}$$

Summary

- Structured penalty to induce desired sparsity
- Simple coordinate descent works for big problems but perhaps not the most efficient
- ADMM is good for complex penalties since one can decompose the problem in layers
- LARS: least angle regression and resolution paths. Deriving solutions for all values of λ in one go. Idea: forward stepwise regression such that add the variable *most correlated with current residuals* until another variable takes on that role.

Summary

How well do the L1-penalized methods perform?

- Biased estimates $\rightarrow$ *adaptive lasso, SCAD* next lecture
- If $\lambda = o(n)$, then $\beta_{l1-pen} \rightarrow \beta_{true}$ as $n \rightarrow \infty$
- If $\lambda \propto n^{1/2}$ L1-pen has a non-zero probability of identifying the true model (model selection consistency) (Knight and Fu, 2000)
- BUT if many of the non-relevant variables are correlated with the relevant variables, L1-pen regression may fail to select the true model even if n is large.
- We need the *Irrepresentable condition* to hold

$$|(X_1'X_1)^{-1}(X_2'X_2)| < 1 - \eta$$

where X_1 are the irrelevant and X_2 the relevant variables. (Zhao and Yu, 2006)