Theory of lasso and beyond linear regression

Pierre-Alexandre Mattei

http://pamattei.github.io – @pamattei
pierre-alexandre.mattei@inria.fr

MSc Data Science

Recap on regularised linear regression

Recap on regularised linear regression

Good old (centered) linear regression

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

If we have a new data point $\mathbf{x}_{\text{new}}$, how do we predict its response?

Good old (centered) linear regression

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

If we have a new data point $\mathbf{x}_{\text{new}}$, how do we predict its response?
We can just use

$$\hat{\mathbf{y}}_{\text{new}} = \mathbb{E}[y|\mathbf{x}_{\text{new}}] = \mathbf{x}_{\text{new}}^T\boldsymbol{\beta}.$$

Good old (centered) linear regression

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

If we have a new data point $\mathbf{x}_{\text{new}}$, how do we predict its response?
We can just use

$$\hat{\mathbf{y}}_{\text{new}} = \mathbb{E}[y|\mathbf{x}_{\text{new}}] = \mathbf{x}_{\text{new}}^T\boldsymbol{\beta}.$$

If we mainly want to do prediction, we're mostly interested in estimating β (and the noise term is not too important to estimate).

OLS for linear regression

Since Legendre anbd Gauss (≈ 1805), the traditional approach to linear regression is through ordinary least squares

$$\hat{\beta}_{\text{OLS}} \in \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\beta||_2^2,$$

which may be interpreted as doing maximum likelihood under the assumption that the noise term is Gaussian.

OLS for linear regression

Since Legendre anbd Gauss ($\approx$ 1805), the traditional approach to linear regression is through ordinary least squares

$$\hat{\beta}_{\text{OLS}} \in \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\beta||_2^2,$$

which may be interpreted as doing maximum likelihood under the assumption that the noise term is Gaussian. If the matrix $\mathbf{X}^T\mathbf{X}$ is invertible, then the problem admits a unique solution:

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

OLS for linear regression

Since Legendre anbd Gauss (≈ 1805), the traditional approach to linear regression is through ordinary least squares

$$\hat{\beta}_{\mathrm{OLS}} \in \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\beta||_2^2,$$

which may be interpreted as doing maximum likelihood under the assumption that the noise term is Gaussian. If the matrix $\mathbf{X}^T\mathbf{X}$ is invertible, then the problem admits a unique solution:

$$\hat{\beta}_{\mathrm{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

In this course, we're concerned by cases where the number of features p is very large. Which leads $\mathbf{X}^T\mathbf{X}$ to be ill-conditioned or non-invertible. This renders OLS impractical.

OLS fails in high dimensions

We have $\text{rank}(\mathbf{X}^T\mathbf{X}) \leq n$, so if $p > n$, the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible!
There is an infinite number of minimisers of the squared error...
We saw last a few simple ways of doing that:

OLS fails in high dimensions

We have $\text{rank}(\mathbf{X}^T\mathbf{X}) \leq n$, so if $p > n$, the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible!
There is an infinite number of minimisers of the squared error...
We saw last a few simple ways of doing that:

- replacing $(\mathbf{X}^T\mathbf{X})^{-1}$ by the Moore-Penrose pseudoinverse $(\mathbf{X}^T\mathbf{X})^\dagger$ ("ridgeless regression" or "Moore-Penrose least squares")

OLS fails in high dimensions

We have $\text{rank}(\mathbf{X}^T\mathbf{X}) \leq n$, so if $p > n$, the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible!
There is an infinite number of minimisers of the squared error...
We saw last a few simple ways of doing that:

- replacing $(\mathbf{X}^T\mathbf{X})^{-1}$ by the Moore-Penrose pseudoinverse $(\mathbf{X}^T\mathbf{X})^{\dagger}$ ("ridgeless regression" or "Moore-Penrose least squares")
- replacing $(\mathbf{X}^T\mathbf{X})^{-1}$ by $(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1}$ with $\lambda > 0$ ("ridge regression", "Tikhonov regularisation", "ℓ_2 regularisation")

OLS fails in high dimensions

We have $\text{rank}(\mathbf{X}^T\mathbf{X}) \leq n$, so if $p > n$, the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible!
There is an infinite number of minimisers of the squared error...
We saw last a few simple ways of doing that:

- replacing $(\mathbf{X}^T\mathbf{X})^{-1}$ by the Moore-Penrose pseudoinverse $(\mathbf{X}^T\mathbf{X})^\dagger$ ("ridgeless regression" or "Moore-Penrose least squares")
- replacing $(\mathbf{X}^T\mathbf{X})^{-1}$ by $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$ with $\lambda > 0$ ("ridge regression", "Tikhonov regularisation", "ℓ_2 regularisation")
- adding an ℓ_0 penalty to the sum of squared errors ("lasso", "basis pursuit").

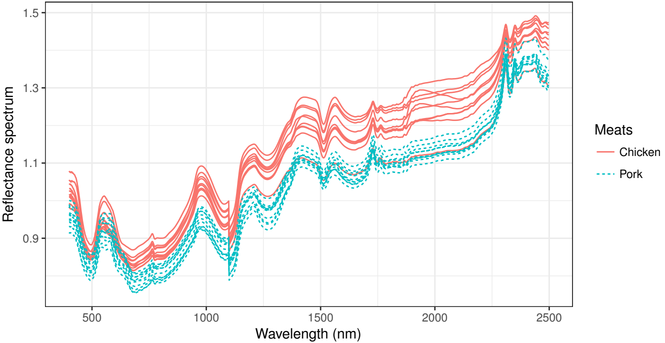
Exercise: What are the advantages/drawbacks of the three methods?

Betting on sparsity

A drawback of both Moore-Penrose and ridge is that they do not lead to a sparse solution. In other words, for them, all p variables are relevant.

Betting on sparsity

A drawback of both Moore-Penrose and ridge is that they do not lead to a sparse solution. In other words, for them, all p variables are relevant.
For example, when dealing with spectra, not all wavelengths are useful in general.



What does sparsity actually mean?

A sparse model is a model that willingfully ignores some of the features of the data at hand. For linear regression, this means that the parameter β will have some coefficients equal to zero.

What does sparsity actually mean?

A sparse model is a model that willingfully ignores some of the features of the data at hand. For linear regression, this means that the parameter $\boldsymbol{\beta}$ will have some coefficients equal to zero.
It will be convient to use the ℓ_0 pseudo-norm of a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, defined as

$$||\boldsymbol{\beta}||_0 = \#\{\beta_j \mid \beta_j \neq 0\} = \text{number of nonzero coefficients of } \boldsymbol{\beta}.$$

What does sparsity actually mean?

A sparse model is a model that willingfully ignores some of the features of the data at hand. For linear regression, this means that the parameter $\boldsymbol{\beta}$ will have some coefficients equal to zero.
It will be convient to use the ℓ_0 pseudo-norm of a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, defined as

$$||\boldsymbol{\beta}||_0 = \#\{\beta_j \mid \beta_j \neq 0\} = \text{number of nonzero coefficients of } \boldsymbol{\beta}.$$

We say that $\boldsymbol{\beta} \in \mathbb{R}^p$ is k-sparse when it contains only k coefficients different from zero. In other words, $\boldsymbol{\beta} \in \mathbb{R}^p$ is k-sparse when $||\boldsymbol{\beta}||_0 = k$. Of course, we need to have $k \in \{0, ..., p\}$. The sparsity pattern of $\boldsymbol{\beta}$ is the subset of $\{1, ..., p\}$ corresponding to nonzero coefficients.

What does sparsity actually mean?

A sparse model is a model that willingfully ignores some of the features of the data at hand. For linear regression, this means that the parameter $\boldsymbol{\beta}$ will have some coefficients equal to zero.
It will be convient to use the ℓ_0 pseudo-norm of a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, defined as

$$||\boldsymbol{\beta}||_0 = \#\{\beta_j \mid \beta_j \neq 0\} = \text{number of nonzero coefficients of } \boldsymbol{\beta}.$$

We say that $\boldsymbol{\beta} \in \mathbb{R}^p$ is k-sparse when it contains only k coefficients different from zero. In other words, $\boldsymbol{\beta} \in \mathbb{R}^p$ is k-sparse when $||\boldsymbol{\beta}||_0 = k$. Of course, we need to have $k \in \{0, ..., p\}$. The sparsity pattern of $\boldsymbol{\beta}$ is the subset of $\{1, ..., p\}$ corresponding to nonzero coefficients.

There are 2^p possible sparsity patterns. This means that, for 20 features, we'll already have more than one million patterns... Trying all of them will be impossible in "large p" cases!

Sparsity through ℓ_0 penalisation

Since we have a measure of non-sparsity (the ℓ_0 pseudonorm), we could just use that to obtain sparse solutions!

$$\hat{\boldsymbol{\beta}}_{\ell_0,\lambda} \in \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_0.$$

Sparsity through ℓ_0 penalisation

Since we have a measure of non-sparsity (the ℓ_0 pseudonorm), we could just use that to obtain sparse solutions!

$$\hat{\boldsymbol{\beta}}_{\ell_{\mathbf{0}},\lambda} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_0.$$

If we use a large enough λ, this will give us sparse solutions. However, the loss function is non-differentiable, if we wanted to solve the problem exactly, we would basically need to try all possible 2^p sparsity patterns and compute OLS on them. This is impossible when p becomes bigger than around 50...

Since we can't do ℓ_0 regularisation, what can we do?

Another way of seeing the ℓ_0 regularised problem is as an ℓ_0 constrained problem

$$\text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p, ||\boldsymbol{\beta}||_{\mathbf{0}} \leq k} ||\mathbf{Y} - \mathbf{X}\beta||_2^2.$$

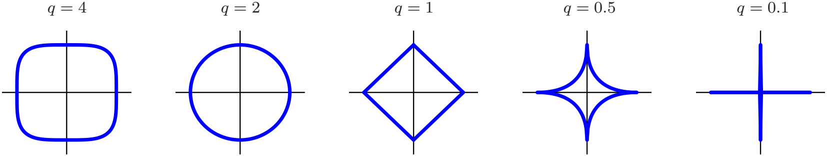
What does the ℓ_0 ball $\{\boldsymbol{\beta} \in \mathbb{R}^p,\ ||\boldsymbol{\beta}||_0 \leq k\}$ look like?

The ℓ_q pseudonorms

One neat way of gaining insight on the ℓ_0 ball is to see it as a limit of ℓ_q balls. For all $q > 0$, let us define

$$||\boldsymbol{\beta}||_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^{1/q}.$$

Note that this measure is not a proper norm unless $q \geq 1$. The ℓ_0 case corresponds to $q \to 0$. Here are various ℓ_q balls (figure from Hastie, Tibshirani & Wainwright, 2015).

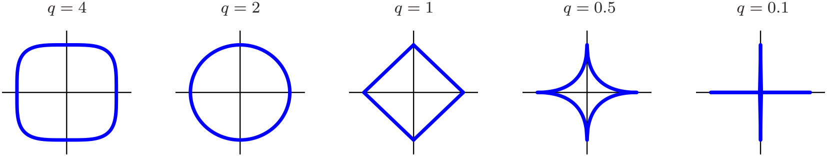


The ℓ_q pseudonorms

One neat way of gaining insight on the ℓ_0 ball is to see it as a limit of ℓ_q balls. For all $q > 0$, let us define

$$||\boldsymbol{\beta}||_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^{1/q}.$$

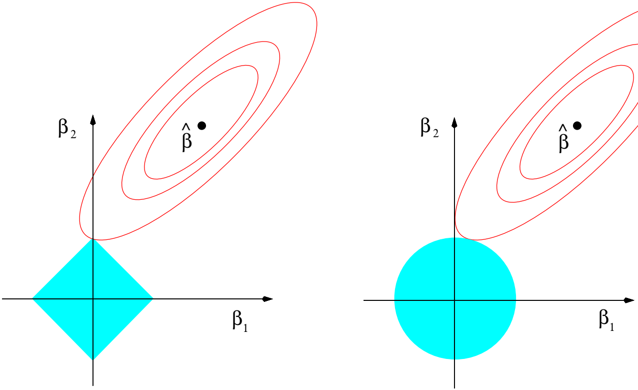
Note that this measure is not a proper norm unless $q \geq 1$. The ℓ_0 case corresponds to $q \to 0$. Here are various ℓ_q balls (figure from Hastie, Tibshirani & Wainwright, 2015).



Key idea of the lasso: replace the discrete, non-differentiable, non-convex ℓ_0 ball by a more regular object, like an ℓ_q ball.

From ridge to lasso

We already studied the ℓ_2 case, it's just ridge! But ridge does not give sparsity.... To get sparsity, we need sharp edges on the ball (figure from Hastie, Tibshirani & Wainwright, 2015).



From ridge to lasso

All ℓ_q balls have sharp edges when $q < 1$, and all of them would lead to sparse solutions.

From ridge to lasso

All ℓ_q balls have sharp edges when $q < 1$, and all of them would lead to sparse solutions.

Keep in mind that we're doing this for computational reasons: we want something fast and cheap.
This leads to the desideratum of having a convex optimisation problem. Of course, the squared error is convex in $\boldsymbol{\beta}$. What about the penalty?

From ridge to lasso

All ℓ_q balls have sharp edges when $q < 1$, and all of them would lead to sparse solutions.

Keep in mind that we're doing this for computational reasons: we want something fast and cheap.
This leads to the desideratum of having a convex optimisation problem. Of course, the squared error is convex in $\boldsymbol{\beta}$. What about the penalty?

The only convex ℓ_q ball is the ℓ_1 ball, which justifies the choice of $q = 1$. This leads to the lasso estimate

$$\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda} \in \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1.$$

Recap on regularised linear regression

What does it mean to do lasso theory?

There are many goals one could have in mind. Here are a few examples, on the fundamental side...

- estimation: assuming the "true model" is really sparse, does the lasso solution recover the true sparsity pattern

What does it mean to do lasso theory?

There are many goals one could have in mind. Here are a few examples, on the fundamental side...

- estimation: assuming the "true model" is really sparse, does the lasso solution recover the true sparsity pattern
- support recovery: assuming the "true model" is really sparse, does the lasso solution recover the true sparsity pattern

What does it mean to do lasso theory?

There are many goals one could have in mind. Here are a few examples, on the fundamental side...

- estimation: assuming the "true model" is really sparse, does the lasso solution recover the true sparsity pattern
- support recovery: assuming the "true model" is really sparse, does the lasso solution recover the true sparsity pattern
- predictive performance: does the lasso solution give better predictions than other (linear) predictions?

What does it mean to do lasso theory?

There are many goals one could have in mind. Here are a few examples, on the fundamental side...

- estimation: assuming the "true model" is really sparse, does the lasso solution recover the true sparsity pattern
- support recovery: assuming the "true model" is really sparse, does the lasso solution recover the true sparsity pattern
- predictive performance: does the lasso solution give better predictions than other (linear) predictions?

An important question is whether or not one has to make asymptotic assumptions (i.e. n, p, or both go to infinity) in order to have theoretical results. Non-asymptotic results are usually more realistic and useful, but they are harder to derive and sometimes uglier/more cryptic. All these results can be useful to know under which assumptions the lasso "works", and when it is relevant to practically use it.

What does it mean to do lasso theory?

... and on a more practical side...

- optimisation: is the lasso solution unique? what are the guarantees that we will find it in reasonable time?
- degrees of freedom: can we derive AIC/BIC-type criteria for the lasso?
- properties of $\hat{\beta}_{\text{lasso}}$: are we sure that some coefficients will be zero? how many of them?
- can we use theoretical insight to improve the lasso, or design faster algorithms for it?

Brief history of lasso theory

Important theoretical work on the lasso has been one the biggest trends in stats of the period 2000-2020.

Brief history of lasso theory

Important theoretical work on the lasso has been one the biggest trends in stats of the period 2000-2020.

One of the first important papers that started it was *Asymptotics for lasso-type estimators*, by Knight & Fu (*Annals of Statistics*, 2000). Then, many prominent researchers form stats/proba/optim communities tackled the issues of the previous slides, even including "famous outsiders" like Terence Tao (Candès & Tao, *Annals of Statistics*, 2007).

Brief history of lasso theory

Important theoretical work on the lasso has been one the biggest trends in stats of the period 2000-2020.

One of the first important papers that started it was *Asymptotics for lasso-type estimators*, by Knight & Fu (*Annals of Statistics*, 2000). Then, many prominent researchers form stats/proba/optim communities tackled the issues of the previous slides, even including "famous outsiders" like Terence Tao (Candès & Tao, *Annals of Statistics*, 2007).

A good overview of these results is provided in Chapter 11 of Hastie, Tibshirani & Wainwright (2015).

The kind of theory we'll study today

We will show a simple theorem that allows us to understand where the zeros in the lasso solution come from.

The kind of theory we'll study today

We will show a simple theorem that allows us to understand where the zeros in the lasso solution come from.

We will state without proof a few interesting (and usually much harder to prove) results. Some more fundamental, some more practical.

Tibshirani's simple theorem

Theorem (Tibshirani, 1996). If $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, then

$$\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda} = \text{sign}(\hat{\boldsymbol{\beta}}_{\text{OLS}})(|\hat{\boldsymbol{\beta}}_{\text{OLS}}| - \lambda)^+,$$

where $x^+ = \max(x, 0) = \text{ReLU}(x)$.

[1] https://stats.stackexchange.com/questions/323234/how-is-the-lasso-orthogonal-design-case-solution-derived

Tibshirani's simple theorem

Theorem (Tibshirani, 1996). If $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, then

$$\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda} = \text{sign}(\hat{\boldsymbol{\beta}}_{\text{OLS}})(|\hat{\boldsymbol{\beta}}_{\text{OLS}}| - \lambda)^+,$$

where $x^+ = \max(x, 0) = \text{ReLU}(x)$.

A proof can be found on stats stackexchange![1]

[1] https://stats.stackexchange.com/questions/323234/how-is-the-lasso-orthogonal-design-case-solution-derived

Tibshirani's simple theorem

Theorem (Tibshirani, 1996). If $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, then

$$\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda} = \text{sign}(\hat{\boldsymbol{\beta}}_{\text{OLS}})(|\hat{\boldsymbol{\beta}}_{\text{OLS}}| - \lambda)^+,$$

where $x^+ = \max(x, 0) = \text{ReLU}(x)$.

A proof can be found on stats stackexchange![1]

The function $S : x \mapsto \text{sign}(x)(|x| - \lambda)^+$ is called the soft-thresholding operator, and is present a lot in sparse contexts.

[1] `https://stats.stackexchange.com/questions/323234/how-is-the-lasso-orthogonal-design-case-solution-derived`

A few more complex results

There are a lot of result that show that, under some (unfortunately strong) assumtions on $\mathbf{X}$, we can have consistency of $\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda}$ or of its support when both n and p are large but the model is assumed to be truly sparse. See chapter 11 of of Hastie, Tibshirani & Wainwright (2015) for more on this.

A few more complex results

There are a lot of result that show that, under some (unfortunately strong) assumtions on $\mathbf{X}$, we can have consistency of $\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda}$ or of its support when both n and p are large but the model is assumed to be truly sparse. See chapter 11 of of Hastie, Tibshirani & Wainwright (2015) for more on this.

There are also results that pretty much do not need any assumption. One of these is the one we will see in the next slide, that in particular does not assume that the "true model" is linear.

An "assumption-free" result

We just assume that $\mathbf{x}$ and y are two random variables, and that all features are bounded by some constant C. We do not assume that $x \mapsto \mathbb{E}[y|x]$ is linear.

[2] various versions due to Greenshtein, Ritov, Juditsky, Nemirovski, and Wasserman, see more details on `https://normaldeviate.wordpress.com/2013/10/03/assumption-free-high-dimensional-inference/`

An "assumption-free" result

We just assume that $\mathbf{x}$ and y are two random variables, and that all features are bounded by some constant C. We do not assume that $x \mapsto \mathbb{E}[y|x]$ is linear.

For a parameter $\boldsymbol{\beta}$, we define the predictive error

$$R(\boldsymbol{\beta}) = \mathbb{E}_{\mathbf{x},y}\left[(y - \mathbf{x}^T\boldsymbol{\beta})^2\right].$$

[2] various versions due to Greenshtein, Ritov, Juditsky, Nemirovski, and Wasserman, see more details on `https://normaldeviate.wordpress.com/2013/10/03/assumption-free-high-dimensional-inference/`

An "assumption-free" result

We just assume that $\mathbf{x}$ and y are two random variables, and that all features are bounded by some constant C. We do not assume that $x \mapsto \mathbb{E}[y|x]$ is linear.

For a parameter $\boldsymbol{\beta}$, we define the predictive error

$$R(\boldsymbol{\beta}) = \mathbb{E}_{\mathbf{x},y}\left[(y - \mathbf{x}^T \boldsymbol{\beta})^2\right].$$

Note that computing $R(\boldsymbol{\beta})$ is impossible because we only have a finite data set. Let us consider

$$\boldsymbol{\beta}^* \in \operatorname{argmin}_{||\boldsymbol{\beta}||_1 \leq L} R(\boldsymbol{\beta}).$$

The parameter $\boldsymbol{\beta}^*$ does not correspond to a "true model", but is the best sparse linear predictor that we could have computed with an infinite data set.

Theorem.[2] The lasso does almost as good as $\boldsymbol{\beta}^*$:

$$R(\hat{\boldsymbol{\beta}}_{\text{lasso},L}) \leq R(\boldsymbol{\beta}^*) + \sqrt{\frac{8C^2L^4}{n}\log\left(\frac{2p^2}{\delta}\right)}.$$

[2] various versions due to Greenshtein, Ritov, Juditsky, Nemirovski, and Wasserman, see more details on `https://normaldeviate.wordpress.com/2013/10/03/assumption-free-high-dimensional-inference/`

A few "useful in the real-world" results

- convex problem, with very fast possible optimisation
- essentially, the problem has a unique solution
- $\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda}$ is sparse (the larger the λ, the sparser)
- $\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda}$ will contain at most $\min(n, p)$ nonzero coefficients

Exercise: Why are these useful?

The elastic net

One potential issue of the lasso is that the solution has at most $\min(n, p)$ nonzero coefficients. That feels quite arbitrary, and a bit weird and unwanted.

A solution was proposed by Zou and Hastie (JRSSB, 2005), it's a simple extension of the lasso called the elastic net.

$$\hat{\boldsymbol{\beta}}_{\text{enet},\lambda_1,\lambda_2} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||_2^2.$$

Exercise: For which values of λ_1, λ_2 do we recover ridge, or lasso?

A few (appealing) properties of the elastic net

$$\hat{\boldsymbol{\beta}}_{\text{enet},\lambda_1,\lambda_2} \in \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||_2^2.$$

- The solution can be sparse, but can also have as many nonzero as it wants
- The problem is strictly convex
- There is a grouping effect: if two features are very similar the elastic net is quite likely to select them both, not the lasso

Exercise: Any apparent drawback?

Recap on regularised linear regression

Beyond linear regression

Let's say we have a loss function $\ell(\boldsymbol{\beta})$ that we want to minimise, and we want to find sparse solutions. A simple way to borrow the elastic-net/lasso ideas is to look at

$$\hat{\boldsymbol{\beta}}_{\text{enet},\lambda_1,\lambda_2} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta}) + \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||_2^2.$$

Exercise: Can you think of few examples?

Beyond linear regression

Let's say we have a loss function $\ell(\boldsymbol{\beta})$ that we want to minimise, and we want to find sparse solutions. A simple way to borrow the elastic-net/lasso ideas is to look at

$$\hat{\boldsymbol{\beta}}_{\text{enet},\lambda_1,\lambda_2} \in \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta}) + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2.$$

Exercise: Can you think of few examples?

- classification: $\ell(\boldsymbol{\beta})$ is the cross-entropy, or the negative likelihood
- clustering: $\ell(\boldsymbol{\beta})$ is k-means loss, or the negative likelihood of a Gaussian mixture model

We will now see another quite different example where the ideas of sparse modelling are very useful: collaborative filtering, aka recommender systems.