Betting on Sparsity with the Lasso. Part I: Linear Regression

Pierre-Alexandre Mattei

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

MSc Data Science

From ridge to lasso

Good old linear regression

The sparse way

From ridge to lasso

Good old linear regression

The sparse way

Good old linear regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mu\mathbf{1}_n + \boldsymbol{\varepsilon}$$

$\mathbf{Y} \in \mathbb{R}^n$ is a vector of n observed responses,
$\mathbf{X} \in \mathcal{M}_{n,p}$ is the design matrix with p input variables,
$\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a stochastic noise term with zero mean and finite variance,
$\mu \in \mathbb{R}$ is the intercept,
Goal: Predicting the value of the response given some new data.

Good old linear regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mu\mathbf{1}_n + \boldsymbol{\varepsilon}$$

$\mathbf{Y} \in \mathbb{R}^n$ is a vector of n observed responses,
$\mathbf{X} \in \mathcal{M}_{n,p}$ is the design matrix with p input variables,
$\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a stochastic noise term with zero mean and finite variance,
$\mu \in \mathbb{R}$ is the intercept,
Goal: Predicting the value of the response given some new data.

Centering the data

We typically assume that the data have been centered, i.e. $\mathbf{Y}$ has zero mean. A nice thing about this is that it allows us to forget about the intercept (which can just be estimated by $\mu = 0$).
Of course, we can still "uncenter" our predictions at the end if we want (as long as we kept the original means). We can just estimate μ by the empirical mean of $\mathbf{Y}$ and add that to our predictions.

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

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

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

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

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

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

Note that, for this to be true, we don't need the noise to be Gaussian (we just needed it to have zero mean).

Good old (centered) linear regression

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

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

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

Note that, for this to be true, we don't need the noise to be Gaussian (we just needed it to have zero mean).
Important consequence: 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

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

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

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. 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

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

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:

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

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...

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...
Exercise: We saw last week two simple ways of doing that. What were they?

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...
Exercise: We saw last week two simple ways of doing that. What were they?

- 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...
Exercise: We saw last week two simple ways of doing that. What were they?

- 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")

Exercise: What are the advantages/drawbacks of the two 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.

Exercise: Find some simple applied examples where it's clear that not using all variables would be better.

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.

Exercise: Find some simple applied examples where it's clear that not using all variables would be better.

In genomics, when there's a causal link between a specific gene and a disease, we certainly won't need all the genome!

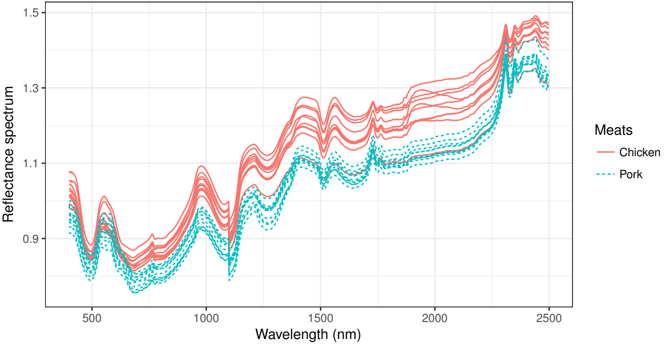
Betting on sparsity

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.

Exercise: Find some simple applied examples where it's clear that not using all variables would be better.

In genomics, when there's a causal link between a specific gene and a disease, we certainly won't need all the genome!

When dealing with spectra, not all wavelengths are useful in general.



What does sparsity actually mean?

A sparse model is a model that willignfully 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.

What does sparsity actually mean?

A sparse model is a model that willignfully 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 willignfully 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. Exercise: How many possible sparsity patterns are there?

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!

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

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!

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

If we use a large enough λ, this will give us sparse solutions. Exercise: what would be the drawbacks of this approach?

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!

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

If we use a large enough λ, this will give us sparse solutions. Exercise: what would be the drawbacks of this approach? 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 10...

Sparsity through ℓ_0 penalisation – links with AIC/BIC

Recall the definitions of AIC/BIC-type penalties

$$\mathrm{AIC} = -2 \times \text{likelihood} + 2 \times \text{nb. of free parameters}$$

$$\mathrm{BIC} = -2 \times \text{likelihood} + \log(n) \times \text{nb. of free parameters}$$

Exercise: can we use this to choose λ in ℓ_0-penalised linear regression?

Sparsity through ℓ_0 penalisation – links with AIC/BIC

Recall the definitions of AIC/BIC-type penalties

$$\text{AIC} = -2 \times \text{likelihood} + 2 \times \text{nb. of free parameters}$$

$$\text{BIC} = -2 \times \text{likelihood} + \log(n) \times \text{nb. of free parameters}$$

Exercise: can we use this to choose λ in ℓ_0-penalised linear regression?

Assuming that the noise is Gaussian with known variance σ^2, the likelihood is, up to a constant, $-||\mathbf{Y} - \mathbf{X}\beta||_2^2/(2\sigma^2)$, therefore

$$-2 \times \text{likelihood} = \frac{1}{\sigma^2}||\mathbf{Y} - \mathbf{X}\beta||_2^2,$$

which means that AIC will correspond to $\lambda = 2/\sigma^2$, and BIC to $\lambda = \log(n)/\sigma^2$.

Sparsity through ℓ_0 penalisation – links with AIC/BIC

Recall the definitions of AIC/BIC-type penalties

$$\mathrm{AIC} = -2 \times \text{likelihood} + 2 \times \text{nb. of free parameters}$$

$$\mathrm{BIC} = -2 \times \text{likelihood} + \log(n) \times \text{nb. of free parameters}$$

Exercise: can we use this to choose λ in ℓ_0-penalised linear regression?

Sparsity through ℓ_0 penalisation – links with AIC/BIC

Recall the definitions of AIC/BIC-type penalties

$$\text{AIC} = -2 \times \text{likelihood} + 2 \times \text{nb. of free parameters}$$

$$\text{BIC} = -2 \times \text{likelihood} + \log(n) \times \text{nb. of free parameters}$$

Exercise: can we use this to choose λ in ℓ_0-penalised linear regression?

Assuming that the noise is Gaussian with known variance σ^2, the likelihood is, up to a constant, $-||\mathbf{Y} - \mathbf{X}\beta||_2^2/(2\sigma^2)$, therefore

$$-2 \times \text{likelihood} = \frac{1}{\sigma^2}||\mathbf{Y} - \mathbf{X}\beta||_2^2,$$

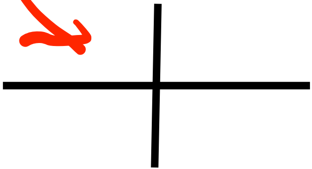
which means that AIC will correspond to $\lambda = 2/\sigma^2$, and BIC to $\lambda = \log(n)/\sigma^2$.

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

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

What does the ℓ_0 ball $\{\boldsymbol{\beta} \in \mathbb{R}^p, \ ||\boldsymbol{\beta}||_0 \leq k\}$ look like? Exercise: draw it for $p = 2$ and $k = 1$.



$p = 2$

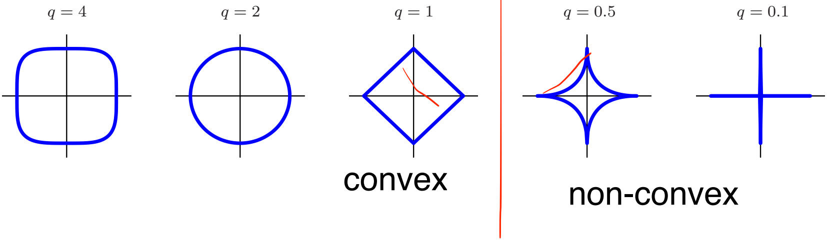
$k = 1$

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).

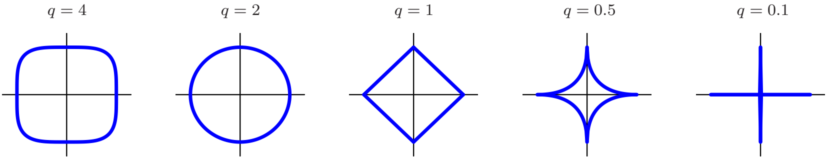


The ℓ_q pseudonorms

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

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

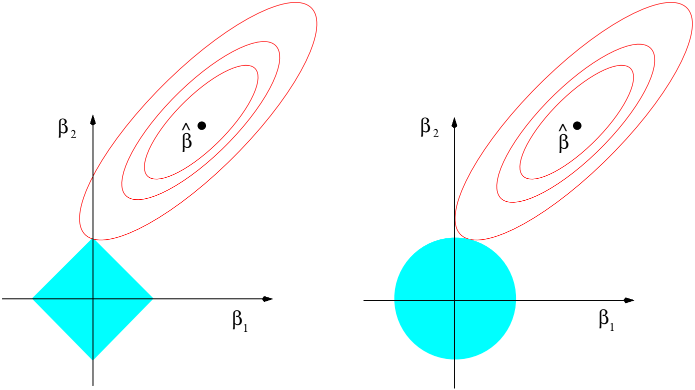
Note that this measure is not a proper norm unless $q \geq 1$. The ℓ_0 case corresponds to $q \to 0$. Here are various ℓ_q balls (figure from Hastie, Tibshirani & Wainwright, 2015).



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 $\boldsymbol{\beta}$. Exercise : 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 $\boldsymbol{\beta}$. Exercise : What about the penalty?

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

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

Some properties of the lasso

We'll give more details about some of these later in the course:

- convex problem, with very fast possible optimisation
- $\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda}$ is sparse (the larger the λ, the sparser)
- $\hat{\boldsymbol{\beta}}_{\text{lasso},\lambda}$ will contain at most $\min(n, p)$ nonzero coefficients
- "easy" to generalise beyond linear regression