Regression overparameterization

Jean-Michel Marin

University of Montpellier
Faculty of Sciences

HAX912X - 2024/2025

1. Context

2. Ridge regression

3. The LASSO method

4. Dimension reduction by manufacturing new regressors

Context

We're in a situation where there's a very large number of explanatory variables (regressors)

Eventually, there are more regressors than observations ($n < p$)

- If $n < p$, $(X^{\mathsf{T}}X)$ is not invertible and LSE cannot be used
- If $n > p$ but close to p, LSE has low predictive power

Context

Objectives

1. Find a (biased) estimator with good predictive power
2. Estimate to 0 the β_j that are zero

You have to accept a bias to get a better prediction

Regularized / penalized regression !

Ridge regression

We minimize $\sum_{i=1}^{n}(y_i - x_i\beta)^2$ under constraint $\sum_{j=1}^{p}\beta_j^2 \leqslant \gamma$

This is equivalent to minimizing

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a certain $\lambda > 0$ which depends on γ (Lagrangian)

Ridge regression

The X columns must be standardized (centered + reduced) systematically for Ridge regression

The ridge estimator is such that

$$\hat{\beta}^R \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

The result is

$$\hat{\beta}^R = \left(X^T X + \lambda I_p \right)^{-1} X^T y$$

Ridge regression

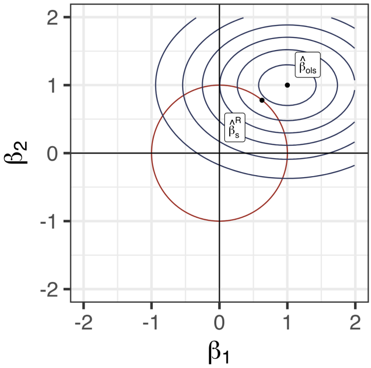
$\hat{\beta}^R$ is a biased estimator of β but with a much smaller RMSE (variance + square of the bias) than the LSE when n is close to p and λ is correctly chosen

If $p > n$, $\hat{\beta}^R$ is well defined

The λ parameter can be chosen by cross-validation

Ridge regression

The predictive power of $\hat{\beta}^R$ is good, but it doesn't lead to simpler models, no estimated coefficient will be zero



The LASSO method

We minimize $\sum_{i=1}^{n}(y_i - x_i\beta)^2$ under constraint $\sum_{j=1}^{p}|\beta_j| \leqslant \gamma$

This is equivalent to minimizing

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

for a certain $\lambda > 0$ which depends on γ (Lagrangian)

The LASSO method

The LASSO estimator is such that

$$\hat{\beta}^L \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} .$$

If $n \geqslant p$, there is a solution, but it's not explicit. Nevertheless, there are very efficient optimization algorithms to solve this problem

When $p > n$, if the solution to the optimization problem is unique, then it will give a non-zero coefficient to at most n regressors

The LASSO method

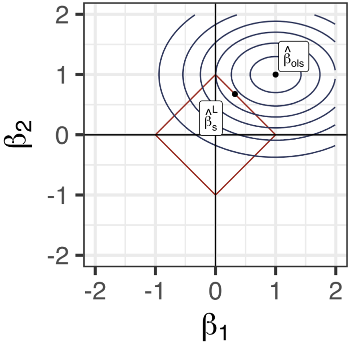
The $\mathbb{X}$ columns must be standardized (centered + reduced) systematically for LASSO

This is also good practice for any type of regression

The λ parameter can be chosen by cross-validation

The LASSO method

The irregularity of penalization means that many coefficients are zero ; LASSO can be used for variable selection



Dimension reduction by manufacturing new regressors

Principal component regression

A PCA transforms X into a matrix $\tilde{X} = XW$ whose columns are orthonormal.

The principal components, those with the greatest inertia, are placed first, only the first q principal components are kept

$$H = XW_q$$

$X \in \mathscr{M}_{n\times p} \longrightarrow \mathscr{M}_{n\times q}$ with $q < p$

Partial Least Square regression