Introduction to statistical learning

3.2 Regularization

V. Lefieux

June 2018

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Table of contents

Introduction

Ridge estimator

LASSO regression

Generalization of Ridge and LASSO

LARS regression

Table of contents

Introduction

Ridge estimator

LASSO regression

Generalization of Ridge and LASSO

LARS regression

Introduction

The regularization methods (e.g Ridge, LASSO, LARS, LOLA) allow to answer several problems:

- Model selection.
- Collinearity correction.
- Cases of "fat matrix": $n < p$ (sparse methods).

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Model

Let consider the regression model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where ε is the error.

Let $(x_{i1}, \ldots, x_{ip}, y_i)_{i \in \{1,\ldots,n\}}$ be n observations of $(X_1, \ldots, X_p, Y)$:

$$\forall i \in \{1, \ldots, n\} : y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \ .$$

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Table of contents

Introduction

Ridge estimator

LASSO regression

Generalization of Ridge and LASSO

LARS regression

Ridge estimator I

The Ridge estimator of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ in the multiple linear model, $\widehat{\boldsymbol{\beta}}^{\text{Ridge}}$, is solution of:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a ℓ^2 penalty term.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Ridge estimator II

It's also the solution of the problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq c \ .$$

Note that:

- $(Y, X_1, \ldots, X_p)$ are centered.
- We exclude the constant in the optimization problem.
- Optimal hyperparameter is obtained by cross-validation.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Matrix form

We can write:

$$\min_{\boldsymbol{\beta}} \left(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\right)^\top \left(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\right) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

where:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbb{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np}. \end{bmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} .$$

The solution is:

$$\widehat{\beta}^{\text{Ridge}} = \left(\mathbb{X}^\top \mathbb{X} + \lambda \, \mathbf{I}_p\right)^{-1} \mathbb{X}^\top \mathbf{Y} \ .$$

It's still a linear estimator in $\mathbf{Y}$.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Singular value decomposition

The singular value decomposition is a factorization M of a (m, n)-dimensional matrix of the form:

$$M = UDV^\top$$

where U (left-singular vectors) and V (right-singular vectors) are (m, m) and (n, n)-dimensional orthogonal matrixes, and D is a "diagonal" (m, n)-dimensional matrix with nonnegative values.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Shrinkage I

Assume that $n > p$.
Singular value decomposition is:

$$\mathbb{X} = UDV^\top \ .$$

where:

- $U\,U^\top = U^\top U = \mathrm{I}_n$,
- $V\,V^\top = V^\top V = \mathrm{I}_p$,
- Diagonal components of D are such that:

$$d_1 \geq \ldots \geq d_p \geq 0 \ .$$

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Shrinkage II

We can write:

$$\mathbb{X}\widehat{\beta}^{\text{Ridge}} = UD\left(D^\top D + \lambda\,\mathrm{I}_p\right)^{-1} D^\top U^\top \mathbf{Y}\ .$$

Let $\left(u^1, \ldots, u^p\right)$ be the columns of U.
Thus columns of UD are $\left(d_1 u^1, \ldots, d_p u^p\right)$ and we have:

$$\mathbb{X}\widehat{\beta}^{\text{Ridge}} = \sum_{j=1}^{p} u^j \left(\frac{d_j^2}{d_j^2 + \lambda^2}\right) {u^j}^\top \mathbf{Y}\ .$$

For $\lambda = 0$, it's the OLS solution:

$$\mathbb{X}\widehat{\beta} = \sum_{j=1}^{p} u^j {u^j}^\top \mathbf{Y}\ .$$

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Shrinkage III

- The j-th component is "constrained" by $\frac{d_j^2}{d_j^2+\lambda^2}$.
- Smallest the coefficients are, more constrained they are.
- If λ is large, components are more constrained.
- Bias increases with λ, variance decreases with λ.
- Watch out for overfitting with small values of λ.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Shrinkage IV

If variables are centered, the variance-covariance matrix is:

$$\frac{1}{n}\mathbb{X}^\top\mathbb{X} = \frac{1}{n}VD^\top DV^\top \ .$$

We know that:

$$D^\top D = \mathrm{diag}\left(d_1^2, \ldots, d_p^2\right) \ .$$

Let $\left(v^1, \ldots, v^p\right)$ be the colums of V.
One can show that $\mathbb{X}v^j$ is j-th principal component of $\mathbb{X}$, with variance $\frac{d_j^2}{n}$.
The last principal components are more constrained that the first ones (large values for d_j).
The degrees of freedom is defined by:

$$\mathrm{df}\left(\lambda\right) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda^2} \ .$$

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Table of contents

Introduction

Ridge estimator

LASSO regression

Generalization of Ridge and LASSO

LARS regression

Shrinkage I

The LASSO estimator (Least Absolute Shrinkage and Seletion Operator) of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ in the multiple linear model, $\widehat{\boldsymbol{\beta}}^{\text{LASSO}}$, is solution of:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \geq 0$ is a ℓ^1 penalty term.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Shrinkage II

It's also the solution of the problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq c \ .$$

Note that:

- $(Y, X_1, \ldots, X_p)$ are centered.
- We exclude the constant in the optimization problem.
- Optimal hyperparameter is obtained by cross-validation.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Properties

- It isn't a linear estimator in $\mathbf{Y}$.
- There is no explicit formula for the LASSO estimator.
- If $\lambda \to +\infty$, $(\beta_j)_{j\in\{1,\dots,p\}}$ tend towards 0.
- If $\mathbb{X}$ is such that $\mathbb{X}^\top\mathbb{X} = \mathsf{I}_p$, the solution of:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + 2\lambda \sum_{j=1}^{p} |\beta_j|$$

is explicit:

$$\forall j \in \{1,\dots,p\} : \widehat{\beta}_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(\left|\widehat{\beta}_j\right| - \lambda\right)\mathbb{1}_{|\widehat{\beta}_j|\geq\lambda}$$

where $\widehat{\beta}_j$ is the OLS estimator ("soft thresholding" of the OLS estimator).

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Table of contents

Introduction

Ridge estimator

LASSO regression

Generalization of Ridge and LASSO

LARS regression

Criterion ℓ^q

Ridge and LASSO estimators are particular cases of the following minimization problem:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

where $\lambda \geq 0$ is a penalty term (obtained by cross validation) and $q \in \mathbb{R}^+$.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Elastic Net regularization

The Elastic Net regularization combines the ℓ^1 and ℓ^2 penalties of the Ridge and LASSO methods:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

where $\lambda \geq 0$ and $\alpha \in]0,1[$ are penalty terms (obtained by cross validation).

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Table of contents

Introduction

Ridge estimator

LASSO regression

Generalization of Ridge and LASSO

LARS regression

Principle

From Hastie et al. (2009):

Forward stepwise regression builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in the active set, and then updates the least squares fit to include all the active variables. Least angle regression uses a similar strategy, but only enters "as much" of a predictor as it deserves.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

Algorithm

After standardizing the predictors (zero mean and unit variance):

1. Initialization: $\mathbf{e} = \mathbf{Y} - \bar{y}\mathbf{1}_n$, $\boldsymbol{\beta} = 0$.
2. Find the most correlated predictor X_j with residual e.
3. "Move" β_j towards $\mathrm{corr}\,(x_j, e)$ until another X_k is more correlated with the current residual.
4. "Move" β_j and β_k towards a direction defined by the OLS estimation of the residual on the (X_j, X_k) until another predictor X_l is more correlated with the current residual.
5. Continue in this way until all predictors have been considered.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References

References

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning. Data Mining, inference, and prediction*. Springer Series in Statistics. Springer, 2 edition.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity. The Lasso and generalizations*. Monographs on Statistics & Applied Probability (143). CRC. Chapman & Hall.

Introduction
Ridge estimator
LASSO regression
Generalization of Ridge and LASSO
LARS regression
References