

Linear models: regularization

ENSE3 / Grenoble-INP

Parcours Numérique 1A




Florent Chatelain* Olivier Michel*

* GIPSA-lab, Univ. Grenoble Alpes,

2019-2020

Regularization methods

Supplementary materials

-  Prof. A. Ihler short (8mn) and educational video
<https://www.youtube.com/watch?v=s04ZirJh9ds>
-  Wikipedia page
https://en.wikipedia.org/wiki/Regularized_least_squares#Specific_examples
-  Scikit-learn very nice documentation with examples (can stop just before section 1.1.4)
https://scikit-learn.org/stable/modules/linear_model.html

Reminder on Least Squares Estimators (LSE)

Linear regression model

For a sized n training set with p variables (may include the intercept)

$$Y = X\beta + \varepsilon,$$

where

- ▶ $Y \in \mathbb{R}^n$ is the response/output vector,
- ▶ $X \in \mathbb{R}^{n \times p}$ is the data matrix (j th column X^j is the sample vector for j th input variable)
- ▶ $\varepsilon \in \mathbb{R}^n$ is the non-predictible part (noise)
- ▶ $\beta \in \mathbb{R}^p$ are the (unknown) coefficients/weights for the input variables

Least Squares (LS) prediction

For a test data $x \in \mathbb{R}^p$, we predict $\hat{y} = x^T \hat{\beta}$ where the LSE

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

is the LS fit on the training set

Limitations of Least Squares Estimators (LSE)

Problem

When $\text{rank } X < p$, or when X has singular values close to zero, then $X^T X$ is no more invertible, or ill conditioned (eigenvalues close to zero)...

Causes

- ▶ redundant or nearly-collinear predictors, e.g. $X^k \approx aX^l + b$, where X^j is the j th column of X
- ▶ **high dimensional** problem where $p \approx n$ (or $p > n$)

Effects

no single, or stable, solution for $\hat{\beta}$

- ▶ high variance of $\hat{\beta}$ as an eigenvalue λ_i of $X^T X$ is close to zero ($\|\hat{\beta}\| \rightarrow +\infty$ as $\lambda_i \rightarrow 0$),
- ▶ true error rate explodes since a small perturbation in the training set yields a substantially different estimate $\hat{\beta}$ and prediction rule $\hat{y} = x^T \hat{\beta}$

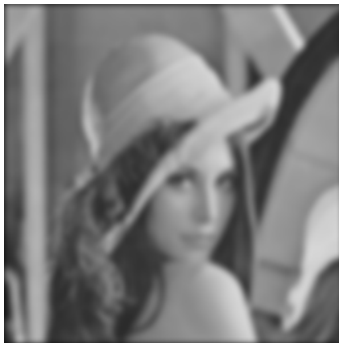
☞ **over-fitting problem**

Instability of LSE: Deconvolution illustration

- ▶ $y \in \mathbb{R}^n$ with $n = 256^2$, $\beta \in \mathbb{R}^p$ with $p = 256^2$,
- ▶ $X \in \mathbb{R}^{n \times p} \leftarrow$ sized $(256^2) \times (256^2)$ matrix...



$\beta \leftarrow$ original image



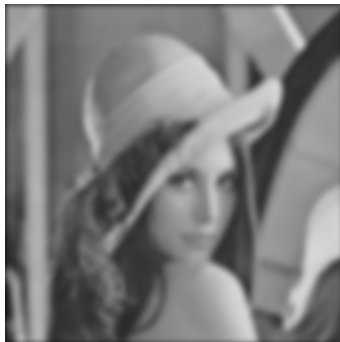
$y = X\beta \leftarrow$ blurred image

Instability of LSE: Deconvolution illustration

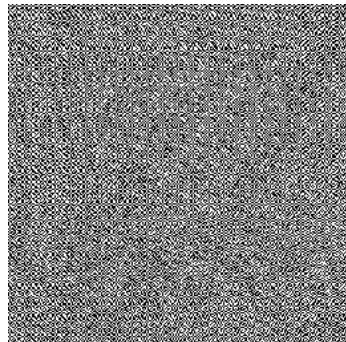
- ▶ $y \in \mathbb{R}^n$ with $n = 256^2$, $\beta \in \mathbb{R}^p$ with $p = 256^2$,
- ▶ $X \in \mathbb{R}^{n \times p} \leftarrow$ sized $(256^2) \times (256^2)$ matrix...



$\beta \leftarrow$ original image



$y = X\beta \leftarrow$ blurred image



$\hat{\beta} = (X^T X)^{-1} X^T y \leftarrow$ LS estimate

Due to the bad conditioning of $X^T X$ (e.v. close to zero), the noise (here numerical round-off errors) is multiplied by an almost infinite gain, and the estimated coefficients $\hat{\beta}_j$ explode to $\pm\infty$!

Outline

Regularization and shrinkage methods

Ridge regression

Lasso estimator

Applications

prostate data

Heart diseases data

Regularization: shrinkage

Idea: introducing a little bias in the estimation of β may lead to a substantial decrease in variance and, hence, in the true error rate

Penalized regression

Regularize the estimation problem by introducing a penalization term for β

$$\tilde{\beta} = \arg \min_{\beta} [\text{RSS}(\beta) + \lambda \text{Pen}(\beta)]$$

- ▶ $\text{RSS}(\beta)$ is the *fidelity term* to the training set (replace with the opposite log-likelihood $-\ell(\beta)$ for generalized linear model, e.g. logistic regression)
- ▶ $\text{Pen}(\beta)$ is the *a priori* to regularize the solution,
- ▶ $\lambda > 0$ is the penalization coefficient

Choosing λ : tradeoff between overfitting (small λ) and underfitting (large λ)

- 👉 standard practice is to use cross-validation to estimate an optimal λ for the test error rate

Ridge regression

Penalization in the (squared) ℓ_2 sense:

$$\text{Pen}(\beta) \equiv \beta^T \beta = \|\beta\|_2^2, \quad \leftarrow \text{Tychonov regularization}$$

$\tilde{\beta}$ is thus obtained by minimizing

$$\begin{aligned} \text{RSS}(\beta) + \lambda \text{Pen}(\beta) &= (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta, \\ &= (\beta - (X^T X + \lambda I)^{-1} X^T Y)^T (X^T X + \lambda I) (\beta - (X^T X + \lambda I)^{-1} X^T Y) + \text{Cst}, \end{aligned}$$

Ridge estimator: $\tilde{\beta} = (X^T X + \lambda I)^{-1} X^T Y$

Remark

similar to LSE, with an additional 'ridge' on the diagonal of $X^T X$

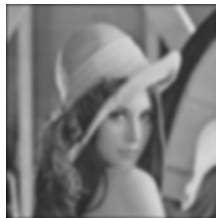
- ▶ $X^T X + \lambda I$ has all its eigenvalues greater than $\lambda > 0$, \leftarrow ensures that $\tilde{\beta}$ is always defined, and stable for large enough λ
- 👉 when $\lambda \rightarrow 0$, then $\tilde{\beta} \rightarrow \hat{\beta}$ (over-fitting risk),
- 👉 when $\lambda \rightarrow +\infty$, then $\tilde{\beta} \rightarrow 0$ (under-fitting)
- ▶ Notebook: [N1_L2_regularization.ipynb](#)

Ridge Regression: deconvolution illustration

- ▶ $y \in \mathbb{R}^n$ with $n = 256^2$, $\beta \in \mathbb{R}^p$ with $p = 256^2$,
- ▶ $X \in \mathbb{R}^{n \times p} \leftarrow$ sized $(256^2) \times (256^2)$ matrix...



$\beta \leftarrow$ original image



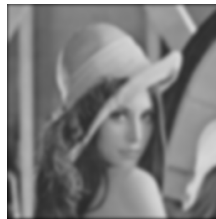
$y = X\beta \leftarrow$ blurred image

Ridge Regression: deconvolution illustration

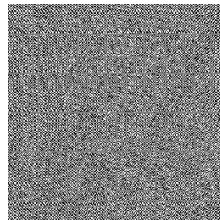
- ▶ $y \in \mathbb{R}^n$ with $n = 256^2$, $\beta \in \mathbb{R}^p$ with $p = 256^2$,
- ▶ $X \in \mathbb{R}^{n \times p} \leftarrow$ sized $(256^2) \times (256^2)$ matrix...



$\beta \leftarrow$ original image



$y = X\beta \leftarrow$ blurred image



$\hat{\beta}_{(X^T X)^{-1} X^T y} \leftarrow$ LS estimate

Ridge Regression: deconvolution illustration

- ▶ $y \in \mathbb{R}^n$ with $n = 256^2$, $\beta \in \mathbb{R}^p$ with $p = 256^2$,
- ▶ $X \in \mathbb{R}^{n \times p} \leftarrow$ sized $(256^2) \times (256^2)$ matrix...



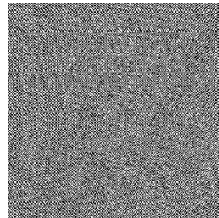
$\beta \leftarrow$ original image



$\tilde{\beta} = (X^T X + \lambda I)^{-1} X^T y \leftarrow$ ridge estimate



$y = X\beta \leftarrow$ blurred image



$\hat{\beta} = (X^T X)^{-1} X^T y \leftarrow$ LS estimate

Regularization by promoting sparsity

Sparse representations/approximations

A representation, or an approximation, is said to be sparse when most of the coefficients are zero

'Bet on Sparsity' principle

Sparsity is a good option in high dimension!

- ▶ if the sparsity assumption does not hold, no method will be able to recover the underlying model in high dimension where $p \approx n$ or $p > n$
- ▶ but if the sparsity assumption holds true, then the parameters can be efficiently estimated by a method that promotes sparsity
- 👉 Occam's razor or KISS (keep it simple, stupid) principles: same idea that simpler models are preferable than more complex ones

Application to the regression problem

choosing a penalization function $\text{Pen}(\beta)$ that promotes the sparsity of β (i.e. with many components $\beta_j = 0$ for $j = 1, \dots, p+1$) \leftarrow Lasso estimator

Lasso ('least absolute shrinkage and selection operator') estimator

Definition

$$\tilde{\beta}_{\text{lasso}} = \arg \min_{\beta} [\text{RSS}(\beta) + \lambda \|\beta\|_1],$$

where $\|\beta\|_1 = \sum_{j=1}^{p+1} |\beta_j|$ is the ℓ_1 norm

- ▶ no analytical expression of $\tilde{\beta}_{\text{lasso}}$
- ▶ but convex optimization problem where very efficient numerical procedures are available to compute $\tilde{\beta}_{\text{lasso}}$

Lasso advantages

Converges to a generally **sparse** solution, i.e. such that $\beta_k = 0$ for a subset of index k

- ☞ the less significant variables are explicitly discarded
- ☞ similar stability than ridge estimator + **variable selection**

▶ **Notebook: `N2_L1_regularization.ipynb`**

Penalization with ℓ_1 and ℓ_2 norms: geometrical interpretation

- Least Squares estimator: $\hat{\beta} = \arg \min \text{RSS}(\beta)$,
- Penalized/Regularized estimator: $\tilde{\beta} = \arg \min (\text{RSS}(\beta) + \lambda \text{Pen}(\beta))$
 $\Leftrightarrow \tilde{\beta} = \arg \min \text{RSS}(\beta)$ under the constraint $\text{Pen}(\tilde{\beta}) \leq s(\lambda)$.

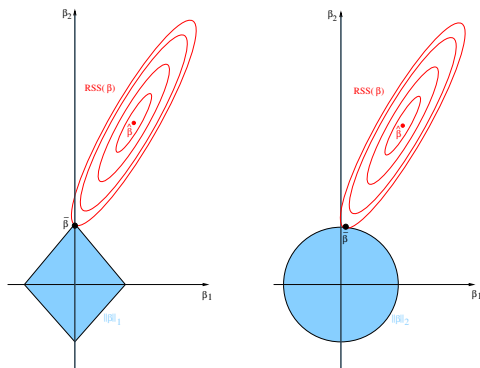


Illustration in dimension $p = 2$: $\beta = (\beta_1, \beta_2)^T$

- red ellipses are the contour plots of RSS
- blue "balls" are the constraint sets for
 lasso: $\text{Pen}(\beta) = \|\beta\|_1 = |\beta_1| + |\beta_2|$ (left),
 ridge: $\text{Pen}(\beta) = \|\beta\|_2^2 = \beta_1^2 + \beta_2^2$ (right).
- LSE $\hat{\beta}$ is the center of the red ellipses
- Penalized LSE $\tilde{\beta}$ is the intersection between red ellipses and blue "ball"
- Here the RSS mainly varies along β_2 , and we get
 $\tilde{\beta}_1 = 0$ for lasso
 (while $\tilde{\beta}_1 \approx 0$ but not zero for ridge)

ℓ_1 norm promotes the sparsity of the estimator: the less significant predictors are explicitly discarded (coeffs β_k are zero) \leftarrow model selection

Scale your data!

- ▶ Linear models (w/o regularization) are invariant under the scaling of the variables: the prediction function is unchanged.
- ▶ Regularized linear models are not due to the penalty term: **scaling of the variables matters!**
- 👉 the variables that have the greatest magnitudes are favoured (same problem for distance based ML methods s.t. K-NN, SVM, ...)

Practical advices

- ▶ If the variables are in different units, scaling each is **strongly recommended**.
- ▶ If they are in the same units, you might or might not scale the variables (depend on your problem)

Usual scaling methods

- ▶ **normalization** in $[0, 1]$: $\tilde{x}_i = \frac{x_i - \min_i}{\max_i - \min_i}$
- ▶ **standardization** to get zero mean and unit variance: $\tilde{x}_i = \frac{x_i - \mu_i}{\sigma_i}$

Outline

Regularization and shrinkage methods

Ridge regression

Lasso estimator

Applications

prostate data

Heart diseases data

Application: prostate data

Stamey et al. (1989) study to examine the association between prostate specific antigen (PSA) and several clinical measures that are potentially associated with PSA in men. Objective is to predict the Log PSA (supervised regression problem) from eight variables

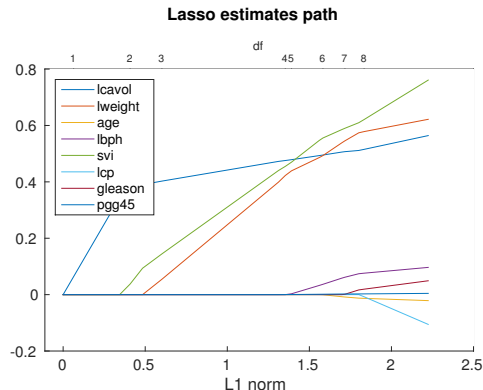
- ▶ lcavol: Log cancer volume
- ▶ lweight: Log prostate weight
- ▶ age: The man's age
- ▶ lbph: Log of the amount of benign hyperplasia
- ▶ svi: Seminal vesicle invasion; 1=Yes, 0=No
- ▶ lcp: Log of capsular penetration
- ▶ gleason: Gleason score
- ▶ pgg45: Percent of Gleason scores 4 or 5

Application : prostate data

Lasso estimate (ℓ_1 -penalization): $\tilde{\beta}(\lambda) = \arg \min_{\beta} \text{RSS}(\beta) + \lambda \|\beta\|_1$,

Lasso path: We can plot the estimated variable coeffs $\tilde{\beta}(\lambda)_j$ vs λ , or equivalently vs $\|\tilde{\beta}(\lambda)\|_1$

- For large λ all the coefficients are zeros ($\|\tilde{\beta}(\lambda)\|_1 = 0$)
- When $\lambda \searrow$ then $\|\tilde{\beta}(\lambda)\|_1 \nearrow$: most significant variables sequentially enter the model (non-zero coeffs)

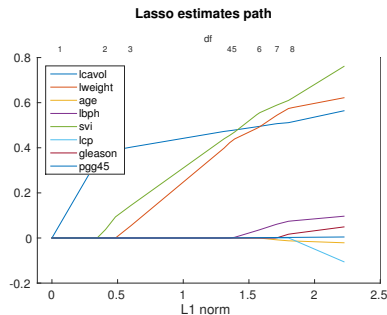
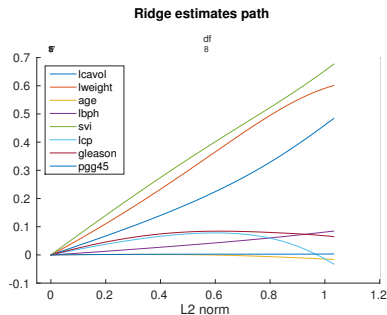


Choosing λ

- large $\|\tilde{\beta}(\lambda)\|_1$ (small λ) \rightarrow overfitting
 - small $\|\tilde{\beta}(\lambda)\|_1$ (large λ) \rightarrow underfitting
 - cross-validation estimation of λ yields $\|\tilde{\beta}(\lambda)\|_1 = 1.06$ ($\lambda = 0.21$)
- \Rightarrow only 3 predictors enter the model to predict PSA: **lcavol**, **svi**, **lweight**

Application : prostate data

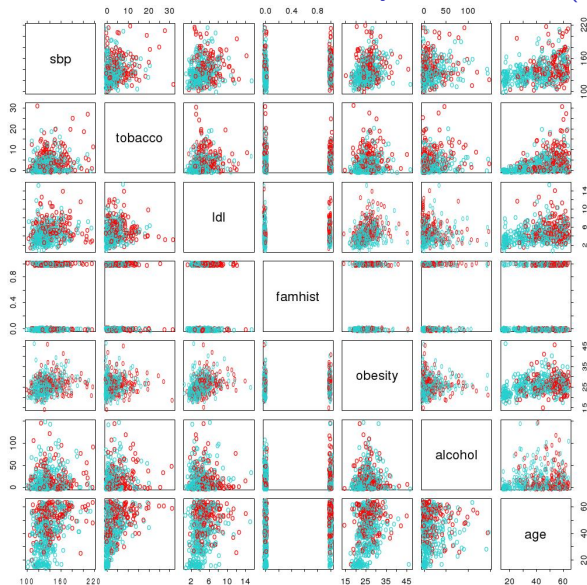
Comparison of ridge and lasso estimators



Path of the penalized coefficients as a function of $\|\tilde{\beta}(\lambda)\|$

- ▶ Ridge estimates are **smooth** functions of λ , with coefficients that are never stuck at zero.
- ▶ Lasso estimates are **piecewise linear** functions, with a kink each time a new variable enter the model
- ▶ **Shrinkage effect**: the larger λ , the more the coefficients are shrunk toward 0 for both penalties
- ▶ For small λ , thus large $\|\tilde{\beta}(\lambda)\|$, both estimator becomes equivalent (convergence toward LSE)

Application: South African coronary heart disease (CHD)



Matrix of the predictor scatterplots

- ▶ each plot \equiv pair of risk factors
- ▶ 160 **cases** / 302 **controls**
- ▶ *ldl*: \sim cholesterol, *sbp*: systolic blood pressure

Application: South African CHD (Cont'd)

Logistic regression fit

	Coefficient	Std. Error	Z score
(Intercept)	-4.130	0.964	-4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	-1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

- A Z score (\equiv Coeff / Std. Error) > 2 in absolute value is significant at the 5% level.

Must be interpreted with caution!

- systolic blood pressure (sbp) is not significant!
- nor is obesity (conversely, < 0 coefficient)!
- result of the **strong correlations** between the predictors: **over-fitting** issue !

Application: South African CHD (Cont'd) with greedy selection procedure

Model selection: greedy backward procedure

To prevent from over-fitting, find the variables that are sufficient for explaining the CHD outputs

- ▶ drop the least significant predictor, and refit the model
- ▶ repeat until no further terms can be dropped ← **backward selection**

Logistic regression fit with backward model selection procedure

	Coefficient	Std. Error	Z score
(Intercept)	-4.204	0.498	-8.45
tobacco	0.081	0.026	3.16
ldl	0.168	0.054	3.09
famhist	0.924	0.223	4.14
age	0.044	0.010	4.52

Interpretations

- ▶ Tobacco is measured in total lifetime usage in kilograms, with a median of 1kg for the controls and 4.1kg for the cases
- ▶ An increase of 1kg \Rightarrow increase of the CHD proba of $\exp(0.081) = 1.084$ or 8.4% (confidence interval at 95% [1.03, 1.14])

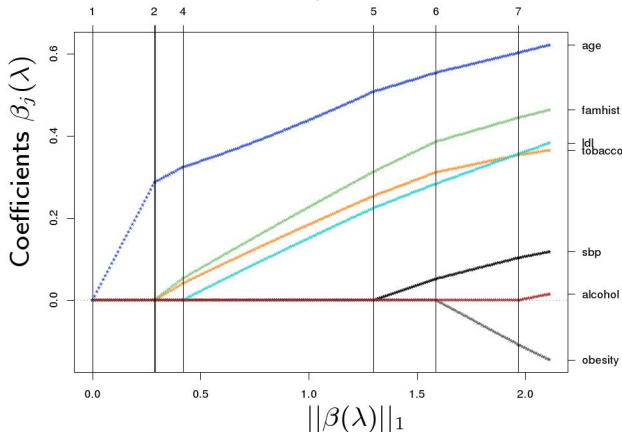
Application: South African CHD (Cont'd) with lasso selection procedure

Model selection: ℓ_1 penalization (Lasso type method)

$$\tilde{\beta}(\lambda) = \arg \min_{\beta} -\ell(\beta) + \lambda \|\beta\|_1,$$

→ function of λ where less significant variables are explicitly discarded

Path of the des coefficients ℓ_1 -penalized coefficients as a function of $\|\hat{\beta}(\lambda)\|_1$



Choosing λ

- ▶ large $\|\tilde{\beta}(\lambda)\|_1$ (small λ) → over-fitting
- ▶ small $\|\tilde{\beta}(\lambda)\|_1$ (large λ) → under-fitting
- ▶ $0.43 \leq \|\tilde{\beta}(\lambda)\|_1 \leq 1.3$ → 4 same predictors than backward selection procedure

Notebook:


`N3_LR_heart_diseases_SA.ipynb`

Conclusions on Regularization for linear models

Regularization procedures are essential tools for data analysis, especially for big datasets involving many predictors, to

- ▶ prevent for over-fitting,
- ▶ better interpret the relations between the variables,
- ▶ improve the prediction performance

Shrinkage procedures

- ▶ ℓ_2 (ridge) regularization promotes the **simplicity**: shrink all the coefficients toward 0
- ▶ ℓ_1 (lasso) regularization promotes the **simplicity+sparsity**: shrink all the coefficients toward 0 + coefficients of non-significant enough variables exactly equal to 0
- ▶ useful to capture the main effects and to interpret the relations between the variables
- ▶  concepts that extend to non-linear methods, e.g. neural nets