



CentraleSupélec

Statistics and Learning

Lecturers (alphabetic order):

Julien Bect, Gilles Faÿ, Ziad Kobeissi, Laurent Le Brusquet,
Vincent Lescarret, Arshak Minasyan, Arthur Tenenhaus[†] & Xujia Zhu

[†] Course coordinator

Lecture 8/9

Regularization and model selection

Course objectives

- ▶ Introduction to regularization for regression and classification.
- ▶ Estimation of generalization error.
- ▶ Selection of hyperparameter values and model selection.

Lecture outline

- 1 – Regularized regression (or classification): penalization
- 2 – Estimation of the risk (generalization error)
- 3 – Hyper-parameters, model selection
- 4 – Exercises and solutions
- 5 – Appendices

Lecture outline

1 – Regularized regression (or classification): penalization

1.1 – Limitations of “ordinary least squares”

1.2 – Ridge regression

1.3 – LASSO regression

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

Lecture outline

1 – Regularized regression (or classification): penalization

1.1 – Limitations of “ordinary least squares”

1.2 – Ridge regression

1.3 – LASSO regression

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

Limitations of “ordinary least squares”

Recall that \underline{X} has size **#individuals** \times **#variables** ($n \times (p + 1)$).

Critical situations for (ordinary) linear regression:

- ▶ when $\underline{X}^\top \underline{X}$ is singular
- ▶ or poorly conditioned

Typical cases

- 1 when the number of variables is large ($p + 1 > n$),
- 2 when there are strong correlations between explanatory variables.

Limitations of “ordinary least squares”

Recall that \underline{X} has size $\# \text{individuals} \times \# \text{variables}$ ($n \times (p + 1)$).

Critical situations for (ordinary) linear regression:

- ▶ when $\underline{X}^T \underline{X}$ is singular
- ▶ or poorly conditioned

Typical cases

- 1 when the number of variables is large ($p + 1 > n$),
- 2 when there are strong correlations between explanatory variables.

Limitations of “ordinary least squares”

Recall that \underline{X} has size $\# \text{individuals} \times \# \text{variables}$ ($n \times (p + 1)$).

Critical situations for (ordinary) linear regression:

- ▶ when $\underline{X}^\top \underline{X}$ is singular
- ▶ or poorly conditioned

Typical cases

- 1 when the number of variables is large ($p + 1 > n$),
- 2 when there are strong correlations between explanatory variables.

Example: $p > n$

	A	B	C	D	E	F	G	H	I	J	K	
1	ID	air_time1	disp_index1	gmrt_in_air1	gmrt_on_paper1	max_x_extension1	max_y_extension1	mean_acc_in_air1	mean_acc_on_paper1	mean_gmrt1	mean_jerk_in_air1	mean
2	id_1	5160	1.25E-05	1.21E+02	8.69E+01	957	6601	3.62E-01	2.17E-01	1.04E+02	5.18E-02	2.
3	id_2	51980	1.60E-05	1.15E+02	8.34E+01	1694	6998	2.73E-01	1.45E-01	9.94E+01	3.98E-02	1.1
4	id_3	2600	1.03E-05	2.30E+02	1.73E+02	2333	5802	3.87E-01	1.81E-01	2.01E+02	6.42E-02	2.1
5	id_4	2130	1.03E-05	3.69E+02	1.83E+02	1756	8159	5.57E-01	1.65E-01	2.76E+02	9.04E-02	2.
6	id_5	2310	6.86E-06	2.58E+02	1.11E+02	987	4732	2.66E-01	1.45E-01	1.85E+02	3.75E-02	1.1
7	id_6	1920	1.14E-05	2.00E+02	1.10E+02	1548	6260	2.13E-01	1.43E-01	1.55E+02	2.84E-02	1.
8	id_7	6415	1.16E-05	2.77E+02	2.80E+02	1837	13414	6.78E-01	1.93E-01	2.78E+02	1.22E-01	2.
9	id_8	1510	6.94E-06	2.84E+02	1.55E+02	2883	4663	6.69E-01	1.68E-01	2.19E+02	1.23E-01	2.
10	id_9	4860	1.31E-05	2.37E+02	3.09E+02	3171	7348	2.77E-01	2.14E-01	2.73E+02	4.08E-02	2.
11	id_10	6265	1.26E-05	3.82E+02	3.54E+02	5568	12313	1.28E+00	1.93E-01	3.68E+02	2.34E-01	1.
12	id_11	2985	1.27E-05	2.21E+02	9.32E+01	1938	6711	3.67E-01	1.53E-01	1.57E+02	5.66E-02	1.1
13	id_12	1970	1.07E-05	2.31E+02	9.06E+01	1434	5643	2.10E-01	1.44E-01	1.61E+02	3.17E-02	1.1
14	id_13	3890	1.05E-05	1.84E+02	1.46E+02	1528	7011	2.50E-01	1.82E-01	1.65E+02	3.21E-02	2.
15	id_14	1190	8.49E-06	3.48E+02	1.98E+02	1739	7297	1.89E-01	1.59E-01	2.73E+02	2.50E-02	1.
16	id_15	2900	1.14E-05	3.05E+02	1.31E+02	1214	8202	9.80E-01	1.28E-01	2.18E+02	1.85E-01	1.1
17	id_16	4955	1.19E-05	3.07E+02	2.09E+02	1652	8863	7.13E-01	1.80E-01	2.58E+02	1.34E-01	2.1
18	id_17	5655	1.01E-05	1.25E+02	1.20E+02	1336	6170	4.82E-01	1.32E-01	1.22E+02	8.48E-02	1.1
19	id_18	12980	1.05E-05	1.65E+02	6.96E+01	11195	7222	4.13E-01	1.61E-01	1.17E+02	7.09E-02	1.1
20	id_19	6265	1.11E-05	7.99E+01	1.10E+02	1475	7755	1.77E-01	1.70E-01	0.95E+01	1.67E-01	1.

Excerpt from a data table describing with $p = 451$ variables the handwriting of $n = 174$ people, some of them suffering from Alzheimer's disease.

In the medical field in particular, it's common to have more descriptors than individuals.

From the study *Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking*, N. D. Cilia et al., 2022, and the associated dataset: UCI Machine Learning Repository. <https://doi.org/10.24432/C55D0K>.

Example: $p > n$

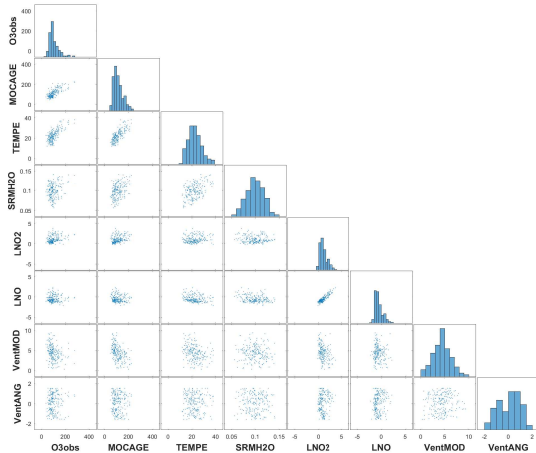
	A	B	C	D	E	F	G	H	I	J	K	
1	ID	air_time1	disp_index1	gmrt_in_air1	gmrt_on_paper1	max_x_extension1	max_y_extension1	mean_acc_in_air1	mean_acc_on_paper1	mean_gmrt1	mean_jerk_in_air1	mean
2	id_1	5160	1.25E-05	1.21E+02	8.69E+01	957	6601	3.62E-01	2.17E-01	1.04E+02	5.18E-02	2.
3	id_2	51980	1.60E-05	1.15E+02	8.34E+01	1694	6998	2.73E-01	1.45E-01	9.94E+01	3.98E-02	1.1
4	id_3	2600	1.03E-05	2.30E+02	1.73E+02	2333	5802	3.87E-01	1.81E-01	2.01E+02	6.42E-02	2.1
5	id_4	2130	1.03E-05	3.69E+02	1.83E+02	1756	8159	5.57E-01	1.65E-01	2.76E+02	9.04E-02	2.
6	id_5	2310	6.86E-06	2.58E+02	1.11E+02	987	4732	2.66E-01	1.45E-01	1.85E+02	3.75E-02	1.1
7	id_6	1920	1.14E-05	2.00E+02	1.10E+02	1548	6260	2.13E-01	1.43E-01	1.55E+02	2.84E-02	1.
8	id_7	6415	1.16E-05	2.77E+02	2.80E+02	1837	13414	6.78E-01	1.93E-01	2.78E+02	1.22E-01	2.
9	id_8	1510	6.94E-06	2.84E+02	1.55E+02	2883	4663	6.69E-01	1.68E-01	2.19E+02	1.23E-01	2.
10	id_9	4860	1.31E-05	2.37E+02	3.09E+02	3171	7348	2.77E-01	2.14E-01	2.73E+02	4.08E-02	2.
11	id_10	6265	1.26E-05	3.82E+02	3.54E+02	5568	12313	1.28E+00	1.93E-01	3.68E+02	2.34E-01	1.
12	id_11	2985	1.27E-05	2.21E+02	9.32E+01	1938	6711	3.67E-01	1.53E-01	1.57E+02	5.66E-02	1.1
13	id_12	1970	1.07E-05	2.31E+02	9.06E+01	1434	5643	2.10E-01	1.44E-01	1.61E+02	3.17E-02	1.1
14	id_13	3890	1.05E-05	1.84E+02	1.46E+02	1528	7011	2.50E-01	1.82E-01	1.65E+02	3.21E-02	2.
15	id_14	1190	8.49E-06	3.48E+02	1.98E+02	1739	7297	1.89E-01	1.59E-01	2.73E+02	2.50E-02	1.
16	id_15	2900	1.14E-05	3.05E+02	1.31E+02	1214	8202	9.80E-01	1.28E-01	2.18E+02	1.85E-01	1.1
17	id_16	4955	1.19E-05	3.07E+02	2.09E+02	1652	8863	7.13E-01	1.80E-01	2.58E+02	1.34E-01	2.1
18	id_17	5655	1.01E-05	1.25E+02	1.20E+02	1336	6170	4.82E-01	1.32E-01	1.22E+02	8.48E-02	1.1
19	id_18	12980	1.05E-05	1.65E+02	6.96E+01	11195	7222	4.13E-01	1.61E-01	1.17E+02	7.09E-02	1.1
20	id_19	6765	1.11E-05	7.99E+01	1.10E+02	1476	7765	1.77E-01	1.70E-01	0.96E+02	1.67E-02	1.

Excerpt from a data table describing with $p = 451$ variables the handwriting of $n = 174$ people, some of them suffering from Alzheimer's disease.

In the medical field in particular, it's common to have more descriptors than individuals.

From the study *Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking*, N. D. Cilia et al., 2022, and the associated dataset: UCI Machine Learning Repository. <https://doi.org/10.24432/C55D0K>.

Example: strong correlation between explanatory variables



“Ozone” example → correlation between variables NO and NO2

Example: strong correlation... (cont'd)

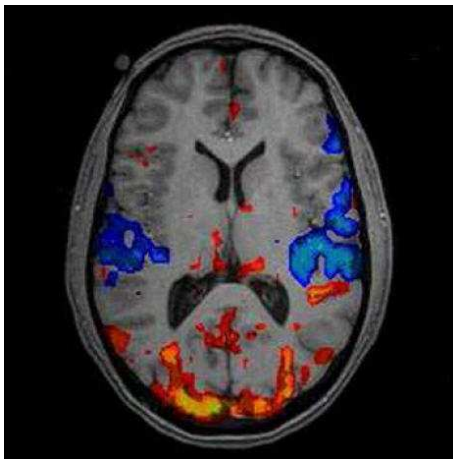
Vector $\hat{\beta}$ obtained by OLS regression:

β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
103.4	28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

Observations:

- ▶ The negative coefficient associated to NO2 is surprising
 ▮ hazardous interpretation of the coefficients
- ▶ The least influential variables (small coefficients) could perhaps be removed from the model?

Example: $p \gg n$ **and** strong correlation



Functional Magnetic Resonance Imaging (fMRI), with approximately, $p \approx 300000$ voxels

Typically, $n \approx 10$ or 100 !

One possible solution: penalized regression

A **penalty** term is added to the empirical risk:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \underbrace{\|\underline{Y} - \underline{X}\beta\|^2}_{\text{data "fidelity"}} + \underbrace{\lambda}_{\text{hyperparameter}} \underbrace{\Omega(\beta)}_{\text{penalty}}. \quad (\star)$$

NB: here and later on, $\|\cdot\|$ denotes the Euclidean norm.

Expected benefits of penalization:

- ▶ make the solution of (\star) unique,
- ▶ take prior information into account (this is related to the Bayesian approach),
- ▶ avoid over-fitting when the family of predictor functions is "large" (for linear models: $p \gg n$),
- ▶ make it easier to interpret the resulting model.

One possible solution: penalized regression

A **penalty** term is added to the empirical risk:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \underbrace{\|\underline{Y} - \underline{X}\beta\|^2}_{\text{data "fidelity"}} + \underbrace{\lambda}_{\text{hyperparameter}} \underbrace{\Omega(\beta)}_{\text{penalty}}. \quad (\star)$$

NB: here and later on, $\|\cdot\|$ denotes the Euclidean norm.

Expected benefits of penalization:

- ▶ make the solution of (\star) **unique**,
- ▶ take **prior information** into account
(this is related to the Bayesian approach),
- ▶ **avoid over-fitting** when the family of predictor functions is
"large" (for linear models: $p \gg n$),
- ▶ make it **easier to interpret** the resulting model.

Lecture outline

1 – Regularized regression (or classification): penalization

1.1 – Limitations of “ordinary least squares”

1.2 – Ridge regression

1.3 – LASSO regression

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

Ridge regression

Penalty

$$\Omega(\beta) = \|\beta\|^2 = \sum_{j=1}^p \beta_j^2$$

(usually, β_0 is not penalized)

$$\hat{\beta}^{\text{RIDGE}} = \operatorname{argmin}_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|^2$$

It can be proved that (\Leftrightarrow see PC):

$$\hat{\beta}^{\text{RIDGE}} = \left(\underline{X}^{\top} \underline{X} + \lambda I_{p+1} \right)^{-1} \underline{X}^{\top} \underline{Y}.$$

\Rightarrow When $\lambda \nearrow$, the conditioning of $(\underline{X}^{\top} \underline{X} + \lambda I_{p+1})$ improves.

Remark: $\hat{\beta}^{\text{RIDGE}}$ has a Bayesian interpretation

(\Leftrightarrow see PC too).

Ridge regression

Penalty

$$\Omega(\beta) = \|\beta\|^2 = \sum_{j=1}^p \beta_j^2$$

(usually, β_0 is not penalized)

$$\hat{\beta}^{\text{RIDGE}} = \operatorname{argmin}_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|^2$$

It can be proved that (\Rightarrow see PC):

$$\hat{\beta}^{\text{RIDGE}} = \left(\underline{X}^{\top} \underline{X} + \lambda I_{p+1} \right)^{-1} \underline{X}^{\top} \underline{Y}.$$

\Rightarrow When $\lambda \nearrow$, the **conditioning** of $(\underline{X}^{\top} \underline{X} + \lambda I_{p+1})$ **improves**.

Remark: $\hat{\beta}^{\text{RIDGE}}$ has a Bayesian interpretation

(\Rightarrow see PC too).

Ridge regression

Penalty

$$\Omega(\beta) = \|\beta\|^2 = \sum_{j=1}^p \beta_j^2$$

(usually, β_0 is not penalized)

$$\hat{\beta}^{\text{RIDGE}} = \operatorname{argmin}_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|^2$$

It can be proved that (\Rightarrow see PC):

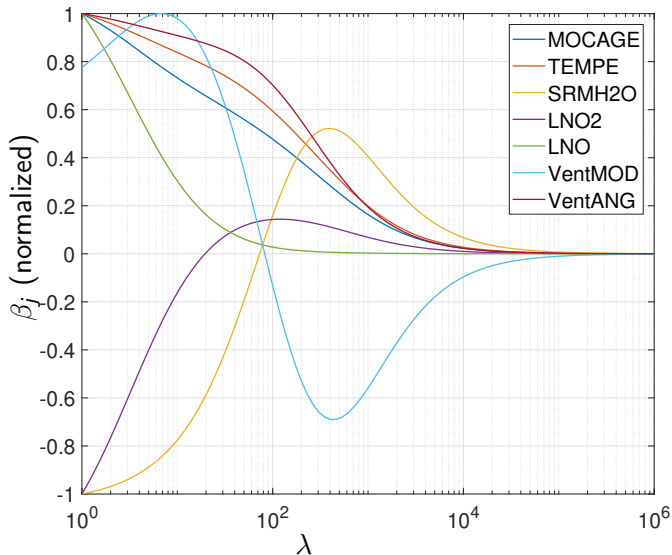
$$\hat{\beta}^{\text{RIDGE}} = \left(\underline{X}^{\top} \underline{X} + \lambda I_{p+1} \right)^{-1} \underline{X}^{\top} \underline{Y}.$$

\Rightarrow When $\lambda \nearrow$, the conditioning of $(\underline{X}^{\top} \underline{X} + \lambda I_{p+1})$ improves.

Remark: $\hat{\beta}^{\text{RIDGE}}$ has a Bayesian interpretation

(\Rightarrow see PC too).

“Ozone” example: Evolution of $\hat{\beta}^{RIDGE}$ as a function of λ



Lecture outline

1 – Regularized regression (or classification): penalization

1.1 – Limitations of “ordinary least squares”

1.2 – Ridge regression

1.3 – LASSO regression

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

LASSO regression

Penalty

$$\Omega(\beta) = \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

(usually, β_0 is not penalized)

$$\hat{\beta}^{\text{LASSO}} = \underset{\beta}{\operatorname{argmin}} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (\star)$$

Minimization of the criterion

- ▶ no explicit solution for $\hat{\beta}^{\text{LASSO}}$ (except in some cases,
 ⇒ dedicated algorithms

exercise 1

LASSO regression

Penalty

$$\Omega(\beta) = \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

(usually, β_0 is not penalized)

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (\star)$$

Minimization of the criterion

- ▶ **no explicit solution** for $\hat{\beta}^{\text{LASSO}}$ (except in some cases,
 ➡ dedicated algorithms

▶ exercise 1

LASSO regression: reformulation

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (\star)$$

- ▶ Let $\hat{\beta}$ denote the OLS estimator of β :

$$\hat{\beta}^{\text{LASSO}} = \hat{\beta} \quad \text{for } \lambda = 0$$

- ▶ Since $\|\underline{Y} - \underline{X}\beta\|^2 = \|\underline{X}(\beta - \hat{\beta})\|^2 + c$, we have:

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\beta} \|\underline{X}(\beta - \hat{\beta})\|^2 + \lambda \|\beta\|_1$$

- ▶ Reformulation with a constraint: it can be proved that there exists $c_{\lambda} \in \mathbb{R}^+$ such that

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\|\beta\|_1 \leq c_{\lambda}} \|\underline{X}(\beta - \hat{\beta})\|^2$$

LASSO regression: reformulation

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda\|\beta\|_1 \quad (\star)$$

- ▶ Let $\hat{\beta}$ denote the OLS estimator of β :

$$\hat{\beta}^{\text{LASSO}} = \hat{\beta} \quad \text{for } \lambda = 0$$

- ▶ Since $\|\underline{Y} - \underline{X}\beta\|^2 = \|\underline{X}(\beta - \hat{\beta})\|^2 + c$, we have:

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\beta} \|\underline{X}(\beta - \hat{\beta})\|^2 + \lambda\|\beta\|_1$$

- ▶ Reformulation with a constraint: it can be proved that there exists $c_{\lambda} \in \mathbb{R}^+$ such that

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\|\beta\|_1 \leq c_{\lambda}} \|\underline{X}(\beta - \hat{\beta})\|^2$$

LASSO regression: reformulation

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (\star)$$

- ▶ Let $\hat{\beta}$ denote the OLS estimator of β :

$$\hat{\beta}^{\text{LASSO}} = \hat{\beta} \quad \text{for } \lambda = 0$$

- ▶ Since $\|\underline{Y} - \underline{X}\beta\|^2 = \|\underline{X}(\beta - \hat{\beta})\|^2 + c$, we have:

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\beta} \|\underline{X}(\beta - \hat{\beta})\|^2 + \lambda \|\beta\|_1$$

- ▶ Reformulation with a **constraint**: it can be proved that there exists $c_{\lambda} \in \mathbb{R}^+$ such that

$$\hat{\beta}^{\text{LASSO}} = \operatorname{argmin}_{\|\beta\|_1 \leq c_{\lambda}} \|\underline{X}(\beta - \hat{\beta})\|^2$$

LASSO regression: intuitive interpretation

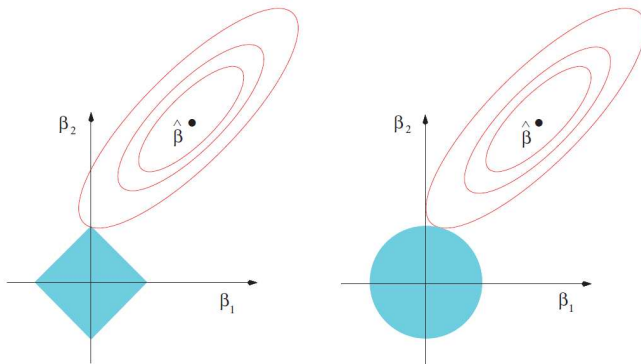
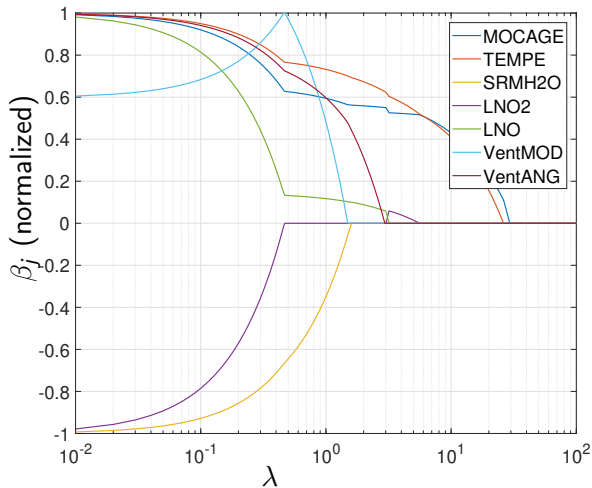


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

“Ozone” example: $\hat{\beta}^{\text{LASSO}}$ versus λ



When $\lambda \nearrow$, the number of coefficients equal to zero \nearrow

“Ozone” example: $\hat{\beta}^{\text{LASSO}}$ for several λ

With $\lambda = 0$ (OLS)

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

⇒ The coefficient for NO2 may seem surprising

With $\lambda = 0.5$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
18.1	17.2	-2.1	0	4.9	2.2	1.9

⇒ One of the two correlated variables is discarded,
makes it easier to interpret the coefficients

With $\lambda = 3$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
15.9	14.1	0	0	2.2	0	0

⇒ The remaining variables are progressively discarded

Choice of the hyper-parameter λ ?

“Ozone” example: $\hat{\beta}^{\text{LASSO}}$ for several λ

With $\lambda = 0$ (OLS)

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

⇒ The coefficient for NO2 may seem surprising

With $\lambda = 0.5$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
18.1	17.2	-2.1	0	4.9	2.2	1.9

⇒ One of the two correlated variables is discarded,
makes it easier to **interpret the coefficients**

With $\lambda = 3$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
15.9	14.1	0	0	2.2	0	0

⇒ The remaining variables are progressively discarded

Choice of the hyper-parameter λ ?

“Ozone” example: $\hat{\beta}^{\text{LASSO}}$ for several λ

With $\lambda = 0$ (OLS)

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

⇒ The coefficient for NO2 may seem surprising

With $\lambda = 0.5$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
18.1	17.2	-2.1	0	4.9	2.2	1.9

⇒ One of the two correlated variables is discarded,
makes it easier to interpret the coefficients

With $\lambda = 3$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
15.9	14.1	0	0	2.2	0	0

⇒ The remaining variables are progressively discarded

Choice of the hyper-parameter λ ?

“Ozone” example: $\hat{\beta}^{\text{LASSO}}$ for several λ

With $\lambda = 0$ (OLS)

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

⇒ The coefficient for NO2 may seem surprising

With $\lambda = 0.5$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
18.1	17.2	-2.1	0	4.9	2.2	1.9

⇒ One of the two correlated variables is discarded,
makes it easier to interpret the coefficients

With $\lambda = 3$

MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
15.9	14.1	0	0	2.2	0	0

⇒ The remaining variables are progressively discarded

Choice of the hyper-parameter λ ?

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

2.1 – Problem

2.2 – Zoom in on an illuminating special case

2.3 – Training set and test set

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

2.1 – Problem

2.2 – Zoom in on an illuminating special case

2.3 – Training set and test set

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

Problem

Back to the **general setting** (regression/classification).

Let \hat{h} be a predictor $\mathcal{X} \rightarrow \mathcal{Y}$ learned from data:

$$\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n)) = \hat{h}(x; \underline{X}, \underline{Y}).$$

Recall that, given a loss function L , we define the risk, or generalization error :

$$\begin{aligned}\mathcal{R}(\hat{h}) &= \mathbb{E} \left(L(Y, \hat{h}(X)) \mid \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) P^{\underline{X}, \underline{Y}}(dx, dy).\end{aligned}$$

Examples. $L(y, \tilde{y}) = (y - \tilde{y})^2$, $L(y, \tilde{y}) = |y - \tilde{y}|$, $L(y, \tilde{y}) = \mathbb{1}_{y \neq \tilde{y}}$, ...

Problem

How can we estimate this risk (which depends on $P^{\underline{X}, \underline{Y}}$) ?

Problem

Back to the general setting (regression/classification).

Let \hat{h} be a predictor $\mathcal{X} \rightarrow \mathcal{Y}$ learned from data:

$$\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n)) = \hat{h}(x; \underline{X}, \underline{Y}).$$

Recall that, given a loss function L , we define the **risk**, or **generalization error** :

$$\begin{aligned}\mathcal{R}(\hat{h}) &= \mathbb{E} \left(L(Y, \hat{h}(X)) \mid \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) P^{\underline{X}, \underline{Y}}(dx, dy).\end{aligned}$$

Examples. $L(y, \tilde{y}) = (y - \tilde{y})^2$, $L(y, \tilde{y}) = |y - \tilde{y}|$, $L(y, \tilde{y}) = \mathbb{1}_{y \neq \tilde{y}}$, ...

Problem

How can we estimate this risk (which depends on $P^{\underline{X}, \underline{Y}}$) ?

Problem

Back to the general setting (regression/classification).

Let \hat{h} be a predictor $\mathcal{X} \rightarrow \mathcal{Y}$ learned from data:

$$\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n)) = \hat{h}(x; \underline{X}, \underline{Y}).$$

Recall that, given a loss function L , we define the risk, or generalization error :

$$\begin{aligned}\mathcal{R}(\hat{h}) &= \mathbb{E} \left(L(Y, \hat{h}(X)) \mid \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) P^{\underline{X}, \underline{Y}}(dx, dy).\end{aligned}$$

Examples. $L(y, \tilde{y}) = (y - \tilde{y})^2$, $L(y, \tilde{y}) = |y - \tilde{y}|$, $L(y, \tilde{y}) = \mathbb{1}_{y \neq \tilde{y}}$, ...

Problem

How can we **estimate this risk** (which depends on $P^{\underline{X}, \underline{Y}}$) ?

Refresher: empirical risk

We call **empirical risk** the risk

$$\hat{\mathcal{R}}_n = \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) \hat{P}_n(dx, dy) = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{h}(X_i))$$

computed with $P^{X,Y}$ equal to $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i, Y_i}$.

Question

Is this empirical risk $\hat{\mathcal{R}}_n$, in general, a “good” estimator of the true risk $\mathcal{R}(\hat{h})$?



the data is used twice !

Intuition: It is “risky” to estimate the risk from the error observed on the same data already used to construct \hat{h} ...

Refresher: empirical risk

We call empirical risk the risk

$$\hat{\mathcal{R}}_n = \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) \hat{P}_n(dx, dy) = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{h}(X_i))$$

computed with $P^{X,Y}$ equal to $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i, Y_i}$.

Question

Is this empirical risk $\hat{\mathcal{R}}_n$, in general, a “good” estimator of the true risk $\mathcal{R}(\hat{h})$?



the data is used twice !

Intuition: It is “risky” to estimate the risk from the error observed on the same data already used to construct \hat{h} ...

Refresher: empirical risk

We call empirical risk the risk

$$\hat{\mathcal{R}}_n = \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) \hat{P}_n(dx, dy) = \frac{1}{n} \sum_{i=1}^n L(\mathbf{Y}_i, \hat{h}(\mathbf{X}_i))$$

computed with $P^{\mathbf{X}, \mathbf{Y}}$ equal to $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{X}_i, \mathbf{Y}_i}$.

Question

Is this empirical risk $\hat{\mathcal{R}}_n$, in general, a “good” estimator of the true risk $\mathcal{R}(\hat{h})$?



the data is used twice !

Intuition: It is “risky” to estimate the risk from the error observed on the same data already used to construct \hat{h} ...

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

2.1 – Problem

2.2 – Zoom in on an illuminating special case

2.3 – Training set and test set

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

Zoom in on an illuminating special case

Consider the case of “ordinary” linear regression:

- ▶ $h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)},$
- ▶ quadratic loss: $L(y, \tilde{y}) = (y - \tilde{y})^2,$
- ▶ $p + 1 \leq n$ and $\underline{X}^\top \underline{X}$ an a.s. invertible $(p + 1) \times (p + 1)$ matrix.

Empirical risk minimization : $\hat{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}.$

Remark: link between $\hat{\mathcal{R}}_n$ and the coefficient R^2 of determination:

$$\begin{aligned} R^2 &= 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}^\top X_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{\hat{\mathcal{R}}_n}{\widehat{\text{var}}_n(Y)} \quad \text{with } \widehat{\text{var}}_n(Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

Zoom in on an illuminating special case

Consider the case of “ordinary” linear regression:

- ▶ $h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)},$
- ▶ quadratic loss: $L(y, \tilde{y}) = (y - \tilde{y})^2,$
- ▶ $p + 1 \leq n$ and $\underline{X}^\top \underline{X}$ an a.s. invertible $(p + 1) \times (p + 1)$ matrix.

Empirical risk minimization : $\hat{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}.$

Remark: link between $\hat{\mathcal{R}}_n$ and the coefficient R^2 of determination:

$$\begin{aligned} R^2 &= 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}^\top X_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{\hat{\mathcal{R}}_n}{\widehat{\text{var}}_n(Y)} \quad \text{with } \widehat{\text{var}}_n(Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

Zoom in on an illuminating special case

Consider the case of “ordinary” linear regression:

- ▶ $h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)},$
- ▶ quadratic loss: $L(y, \tilde{y}) = (y - \tilde{y})^2,$
- ▶ $p + 1 \leq n$ and $\underline{X}^\top \underline{X}$ an a.s. invertible $(p + 1) \times (p + 1)$ matrix.

Empirical risk minimization : $\hat{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}.$

Remark: link between $\hat{\mathcal{R}}_n$ and the coefficient R^2 of determination:

$$\begin{aligned} R^2 &= 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}^\top X_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{\hat{\mathcal{R}}_n}{\widehat{\text{var}}_n(Y)} \quad \text{with } \widehat{\text{var}}_n(Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

Zoom on an illuminating special case (cont'd)

Consider the generalization error wrt responses only:

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right),$$

with, for all i , \tilde{Y}_i and Y_i iid conditionally to \underline{X} .

Proposition

Assume that the unknown distribution $P^{X,Y}$ is such that $Y_i = \beta^\top X_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, independent of X_i . Then

$$\mathbb{E} \left(\tilde{\mathcal{R}}_n \right) = \sigma^2 \left(1 + \frac{p+1}{n} \right),$$

$$\mathbb{E} \left(\hat{\mathcal{R}}_n \right) = \sigma^2 \left(1 - \frac{p+1}{n} \right).$$

Zoom on an illuminating special case (cont'd)

Consider the generalization error wrt responses only:

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right),$$

with, for all i , \tilde{Y}_i and Y_i iid conditionally to \underline{X} .

Proposition

Assume that the unknown distribution $P^{X,Y}$ is such that $Y_i = \beta^\top X_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, independent of X_i . Then

$$\mathbb{E} \left(\tilde{\mathcal{R}}_n \right) = \sigma^2 \left(1 + \frac{p+1}{n} \right),$$

$$\mathbb{E} \left(\hat{\mathcal{R}}_n \right) = \sigma^2 \left(1 - \frac{p+1}{n} \right).$$

Zoom on an illuminating special case (cont'd)

Consider the generalization error wrt responses only:

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right),$$

with, for all i , \tilde{Y}_i and Y_i iid conditionally to \underline{X} .

Proposition

Assume that the unknown distribution $P^{X,Y}$ is such that $Y_i = \beta^\top X_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, independent of X_i . Then

$$\mathbb{E} \left(\tilde{\mathcal{R}}_n \right) = \sigma^2 \left(1 + \frac{p+1}{n} \right),$$

$$\mathbb{E} \left(\hat{\mathcal{R}}_n \right) = \sigma^2 \left(1 - \frac{p+1}{n} \right).$$

Zoom on an illuminating special case (cont'd)

Interpretation. On average, the empirical risk under-estimates the generalization error:

$$\mathbb{E} \left(\tilde{\mathcal{R}}_n - \hat{\mathcal{R}}_n \right) = 2 \frac{p+1}{n} \sigma^2 > 0.$$

Another way of looking at this result. Set

$$\eta = \frac{p+1}{n} = \frac{\text{number of coefficients}}{\text{sample size}}.$$

Then

$$\frac{\mathbb{E} \left(\tilde{\mathcal{R}}_n \right)}{\mathbb{E} \left(\hat{\mathcal{R}}_n \right)} = \frac{1+\eta}{1-\eta} \xrightarrow{\eta \rightarrow 1} +\infty.$$

Zoom on an illuminating special case (cont'd)

Interpretation. On average, the empirical risk under-estimates the generalization error:

$$\mathbb{E} \left(\tilde{\mathcal{R}}_n - \hat{\mathcal{R}}_n \right) = 2 \frac{p+1}{n} \sigma^2 > 0.$$

Another way of looking at this result. Set

$$\eta = \frac{p+1}{n} = \frac{\text{number of coefficients}}{\text{sample size}}.$$

Then

$$\frac{\mathbb{E} \left(\tilde{\mathcal{R}}_n \right)}{\mathbb{E} \left(\hat{\mathcal{R}}_n \right)} = \frac{1 + \eta}{1 - \eta} \xrightarrow[\eta \rightarrow 1]{} +\infty.$$

Zoom on an illuminating special case (cont'd)

Proof. Let us compute first $\mathbb{E}(\tilde{\mathcal{R}}_n \mid \underline{X})$ with (reminder)

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right).$$

We have $\mathbb{E}(\tilde{Y}_i \mid \underline{X}) = \mathbb{E}(\hat{\beta}^\top X_i \mid \underline{X}) = \beta^\top X_i$, therefore

$$\begin{aligned} \mathbb{E}(\tilde{\mathcal{R}}_n \mid \underline{X}) &= \frac{1}{n} \sum_{i=1}^n \text{var}(\tilde{Y}_i - \hat{\beta}^\top X_i \mid \underline{X}) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\text{var}(\tilde{Y}_i \mid \underline{X})}_{=\sigma^2} + \underbrace{\text{var}(\hat{\beta}^\top X_i \mid \underline{X})}_{=0} \right). \end{aligned}$$

Zoom on an illuminating special case (cont'd)

Proof. Let us compute first $\mathbb{E}(\tilde{\mathcal{R}}_n \mid \underline{X})$ with (reminder)

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right).$$

We have $\mathbb{E}(\tilde{Y}_i \mid \underline{X}) = \mathbb{E}(\hat{\beta}^\top X_i \mid \underline{X}) = \beta^\top X_i$, therefore

$$\begin{aligned} \mathbb{E}(\tilde{\mathcal{R}}_n \mid \underline{X}) &= \frac{1}{n} \sum_{i=1}^n \text{var}(\tilde{Y}_i - \hat{\beta}^\top X_i \mid \underline{X}) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\text{var}(\tilde{Y}_i \mid \underline{X})}_{=\sigma^2} + \underbrace{\text{var}(\hat{\beta}^\top X_i \mid \underline{X})}_{=0} \right). \end{aligned}$$

Zoom on an illuminating special case (cont'd)

Proof. Let us compute first $\mathbb{E}(\tilde{\mathcal{R}}_n \mid \underline{X})$ with (reminder)

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right).$$

We have $\mathbb{E}(\tilde{Y}_i \mid \underline{X}) = \mathbb{E}(\hat{\beta}^\top X_i \mid \underline{X}) = \beta^\top X_i$, therefore

$$\begin{aligned} \mathbb{E}(\tilde{\mathcal{R}}_n \mid \underline{X}) &= \frac{1}{n} \sum_{i=1}^n \text{var}(\tilde{Y}_i - \hat{\beta}^\top X_i \mid \underline{X}) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\text{var}(\tilde{Y}_i \mid \underline{X})}_{=\sigma^2} + \underbrace{\text{var}(\hat{\beta}^\top X_i \mid \underline{X})}_{=\otimes} \right). \end{aligned}$$

Zoom on an illuminating special case (cont'd)

We already know that $\text{var}(\hat{\beta} \mid \underline{X}) = \sigma^2 (\underline{X}^\top \underline{X})^{-1}$. Therefore:

$$\begin{aligned} \circledast &= \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) \\ &= X_i^\top \text{var}(\hat{\beta} \mid \underline{X}) X_i \\ &= \sigma^2 X_i^\top (\underline{X}^\top \underline{X})^{-1} X_i \\ &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} X_i X_i^\top\right). \end{aligned}$$

By noting that $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, we get:

$$\begin{aligned} \sum_i \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} \sum_i X_i X_i^\top\right) \\ &= \sigma^2 \text{tr}(I_{p+1}) = \sigma^2 (p+1). \end{aligned}$$

Zoom on an illuminating special case (cont'd)

We already know that $\text{var}(\hat{\beta} \mid \underline{X}) = \sigma^2 (\underline{X}^\top \underline{X})^{-1}$. Therefore:

$$\begin{aligned} \circledast &= \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) \\ &= X_i^\top \text{var}(\hat{\beta} \mid \underline{X}) X_i \\ &= \sigma^2 X_i^\top (\underline{X}^\top \underline{X})^{-1} X_i \\ &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} X_i X_i^\top\right). \end{aligned}$$

By noting that $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, we get:

$$\begin{aligned} \sum_i \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} \sum_i X_i X_i^\top\right) \\ &= \sigma^2 \text{tr}(I_{p+1}) = \sigma^2 (p+1). \end{aligned}$$

Zoom on an illuminating special case (cont'd)

We already know that $\text{var}(\hat{\beta} \mid \underline{X}) = \sigma^2 (\underline{X}^\top \underline{X})^{-1}$. Therefore:

$$\begin{aligned} \circledast &= \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) \\ &= X_i^\top \text{var}(\hat{\beta} \mid \underline{X}) X_i \\ &= \sigma^2 X_i^\top (\underline{X}^\top \underline{X})^{-1} X_i \\ &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} X_i X_i^\top\right). \end{aligned}$$

By noting that $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, we get:

$$\begin{aligned} \sum_i \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} \sum_i X_i X_i^\top\right) \\ &= \sigma^2 \text{tr}(I_{p+1}) = \sigma^2 (p+1). \end{aligned}$$

Zoom on an illuminating special case (cont'd)

We already know that $\text{var}(\hat{\beta} \mid \underline{X}) = \sigma^2 (\underline{X}^\top \underline{X})^{-1}$. Therefore:

$$\begin{aligned} \circledast &= \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) \\ &= X_i^\top \text{var}(\hat{\beta} \mid \underline{X}) X_i \\ &= \sigma^2 X_i^\top (\underline{X}^\top \underline{X})^{-1} X_i \\ &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} X_i X_i^\top\right). \end{aligned}$$

By noting that $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, we get:

$$\begin{aligned} \sum_i \text{var}(\hat{\beta}^\top X_i \mid \underline{X}) &= \sigma^2 \text{tr}\left((\underline{X}^\top \underline{X})^{-1} \sum_i X_i X_i^\top\right) \\ &= \sigma^2 \text{tr}(I_{p+1}) = \sigma^2 (p+1). \end{aligned}$$

Zoom on an illuminating special case (cont'd)

Thus, we have:

$$\begin{aligned}\mathbb{E}\left(\tilde{\mathcal{R}}_n \mid \underline{X}\right) &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\text{var}\left(\tilde{Y}_i \mid \underline{X}\right)}_{=\sigma^2} + \underbrace{\text{var}\left(\hat{\beta}^\top X_i \mid \underline{X}\right)}_{=0} \right) \\ &= \sigma^2 + \sigma^2 \frac{p+1}{n} = \sigma^2 \left(1 + \frac{p+1}{n}\right).\end{aligned}$$

Hence the result: $\mathbb{E}\left(\tilde{\mathcal{R}}_n\right) = \sigma^2 \left(1 + \frac{p+1}{n}\right)$.

Exercise (\Rightarrow see PC): prove the second inequality, i.e.,

$$\mathbb{E}\left(\hat{\mathcal{R}}_n\right) = \sigma^2 \left(1 - \frac{p+1}{n}\right).$$



Zoom on an illuminating special case (cont'd)

Thus, we have:

$$\begin{aligned}\mathbb{E}\left(\tilde{\mathcal{R}}_n \mid \underline{X}\right) &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\text{var}\left(\tilde{Y}_i \mid \underline{X}\right)}_{=\sigma^2} + \underbrace{\text{var}\left(\hat{\beta}^\top X_i \mid \underline{X}\right)}_{=0} \right) \\ &= \sigma^2 + \sigma^2 \frac{p+1}{n} = \sigma^2 \left(1 + \frac{p+1}{n}\right).\end{aligned}$$

Hence the result: $\mathbb{E}\left(\tilde{\mathcal{R}}_n\right) = \sigma^2 \left(1 + \frac{p+1}{n}\right)$.

Exercise (\Rightarrow see PC): prove the second inequality, i.e.,

$$\mathbb{E}\left(\hat{\mathcal{R}}_n\right) = \sigma^2 \left(1 - \frac{p+1}{n}\right).$$



Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

2.1 – Problem

2.2 – Zoom in on an illuminating special case

2.3 – Training set and test set

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

Training set and test set

Conclusion/extrapolation. The empirical risk is in general

- ▶ a **negatively biased estimator** of the risk,
- ▶ with a **bias that is increasing when $p \nearrow$** .

Solution: split the data in two sets

- ▶ training data: used to construct \hat{h} ,
- ▶ test data: used to estimate the generalization error.

Example:

training
(80%)

test
(20%)

Training set and test set

Conclusion/extrapolation. The empirical risk is in general

- ▶ a negatively biased estimator of the risk,
- ▶ with a bias that is increasing when $p \nearrow$.

Solution: split the data in two sets

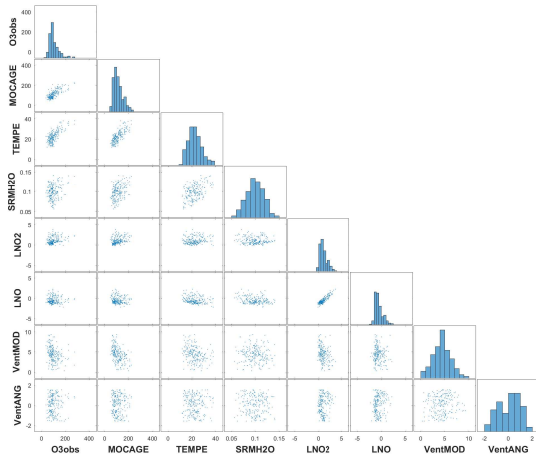
- ▶ **training** data: used to construct \hat{h} ,
- ▶ **test** data: used to estimate the generalization error.

Example:

training
(80%)

test
(20%)

Exemple “Ozone” (cont'd from lecture #6)



Goal: predict the ozone concentration on day $t + 1$
from data available on day t

“Ozone” example: 70/30

Here we use the 7 explanatory variables + 21 interactions $X_j X_k$ ($j \neq k$).

Results from 10 random splits, 70% / 30%:

R^2	$\hat{\mathcal{R}}_n$	$\hat{\mathcal{R}}_n^{\text{test}}$
77.2%	345.1	573.3
76.8%	371.4	496.0
77.3%	344.0	608.6
76.1%	350.5	606.1
78.6%	345.5	669.7
75.5%	399.9	476.6
71.4%	343.7	643.7
77.7%	377.3	524.7
81.8%	317.8	695.9
79.8%	373.2	554.3

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

3.1 – Problem

3.2 – Cross validation

3.3 – AIC criterion

4 – Exercises and solutions

5 – Appendices

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

3.1 – Problem

3.2 – Cross validation

3.3 – AIC criterion

4 – Exercises and solutions

5 – Appendices

Problem #1: choosing a “good” family \mathcal{H}

Example. Selection of k variables among p . Let $J \subset \{1, \dots, p\}$:

$$h(x) = \beta_0 + \sum_{j \in J} \beta_j x^{(j)}.$$

⇒ Defines a family \mathcal{H}_J with $k_J = \text{card}(J) + 1$ parameters.

Example. Expansion in a basis, truncated at order J :

$$h(x) = \sum_{k=0}^J \beta_j \psi_j(x).$$

⇒ Defines a family \mathcal{H}_J with $k_J = J + 1$ parameters.

⇒ complement

Problem: model selection

How to choose the family \mathcal{H}_J (and, in particular, its “size” k_J) ?

Remark: replace $h(x)$ with $\ln \frac{h(x)}{1-h(x)}$ for logistic regression.

Problem #1: choosing a “good” family \mathcal{H}

Example. Selection of k variables among p . Let $J \subset \{1, \dots, p\}$:

$$h(x) = \beta_0 + \sum_{j \in J} \beta_j x^{(j)}.$$

⇒ Defines a family \mathcal{H}_J with $k_J = \text{card}(J) + 1$ parameters.

Example. Expansion in a basis, truncated at order J :

$$h(x) = \sum_{k=0}^J \beta_k \psi_k(x).$$

⇒ Defines a family \mathcal{H}_J with $k_J = J + 1$ parameters.

⇒ complement

Problem: model selection

How to choose the family \mathcal{H}_J (and, in particular, its “size” k_J) ?

Remark: replace $h(x)$ with $\ln \frac{h(x)}{1-h(x)}$ for logistic regression.

Problem #1: choosing a “good” family \mathcal{H}

Example. Selection of k variables among p . Let $J \subset \{1, \dots, p\}$:

$$h(x) = \beta_0 + \sum_{j \in J} \beta_j x^{(j)}.$$

⇒ Defines a family \mathcal{H}_J with $k_J = \text{card}(J) + 1$ parameters.

Example. Expansion in a basis, truncated at order J :

$$h(x) = \sum_{k=0}^J \beta_k \psi_k(x).$$

⇒ Defines a family \mathcal{H}_J with $k_J = J + 1$ parameters.

⇒ complement

Problem: model selection

How to choose the family \mathcal{H}_J (and, in particular, its “size” k_J) ?

Remark: replace $h(x)$ with $\ln \frac{h(x)}{1-h(x)}$ for logistic regression.

Problem #2: choosing an hyper-parameter

Most methods require some “tuning” . . .

- ▶ Ridge/LASSO regression: $\hat{\beta} = \operatorname{argmin} \hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}$, with

$$\hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}(\beta) = \hat{\mathcal{R}}_n(\beta) + \lambda \sum_j |\beta_j|^q, \quad q \in \{1, 2\},$$

- ▶ Decision trees, neural networks: structure
(e.g., number of levels of the tree / layers in the network)
- ▶ The k -nearest neighbors method: $h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i$,
with $\mathcal{V}_{n,k}(x)$ the indices of the k nearest neighbors of x .

Problem

How to choose the value of such hyperparameters ?

Problem #2: choosing an hyper-parameter

Most methods require some “tuning” . . .

- ▶ Ridge/LASSO regression: $\hat{\beta} = \operatorname{argmin} \hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}$, with

$$\hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}(\beta) = \hat{\mathcal{R}}_n(\beta) + \lambda \sum_j |\beta_j|^q, \quad q \in \{1, 2\},$$

- ▶ Decision trees, neural networks: **structure**
(e.g., number of levels of the tree / layers in the network)
- ▶ The k -nearest neighbors method: $h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i$,
with $\mathcal{V}_{n,k}(x)$ the indices of the k nearest neighbors of x .

Problem

How to choose the value of such hyperparameters ?

Problem #2: choosing an hyper-parameter

Most methods require some “tuning”...

- ▶ Ridge/LASSO regression: $\hat{\beta} = \operatorname{argmin} \hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}$, with

$$\hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}(\beta) = \hat{\mathcal{R}}_n(\beta) + \lambda \sum_j |\beta_j|^q, \quad q \in \{1, 2\},$$

- ▶ Decision trees, neural networks: structure
(e.g., number of levels of the tree / layers in the network)
- ▶ The k -nearest neighbors method: $h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i$,
with $\mathcal{V}_{n,k}(x)$ the indices of the k nearest neighbors of x .

Problem

How to choose the value of such hyperparameters ?

Over-fitting: beware!

Idea

Choose the family \mathcal{H}_J , or the hyperparameter λ , in order to **minimize** (an estimation of) the **generalization error**.

 again, the empirical risk $\hat{\mathcal{R}}_n$, estimated on the training data, is not appropriate !

Example. Polynomial regression with $x \in \mathbb{R}$, degree $\leq J$:

$$h(x) = \beta_0 + \beta_1 x + \dots + \beta_J x^J,$$


with $J = 2, 5, 8, 11$.

Recall that, in linear regression, the empirical risk has a downward bias proportional to the number of parameters in the model.

Over-fitting: beware!

Idea

Choose the family \mathcal{H}_J , or the hyperparameter λ , in order to minimize (an estimation of) the generalization error.

 again, the empirical risk $\hat{\mathcal{R}}_n$, estimated on the training data, is not appropriate !

Example. Polynomial regression with $x \in \mathbb{R}$, degree $\leq J$:

$$h(x) = \beta_0 + \beta_1 x + \dots + \beta_J x^J,$$


with $J = 2, 5, 8, 11$.

Recall that, in linear regression, the empirical risk has a downward bias proportional to the number of parameters in the model.

Over-fitting: beware!

Idea

Choose the family \mathcal{H}_J , or the hyperparameter λ , in order to minimize (an estimation of) the generalization error.

 again, the empirical risk $\hat{\mathcal{R}}_n$, estimated on the training data, is not appropriate !

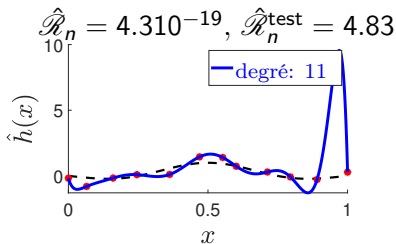
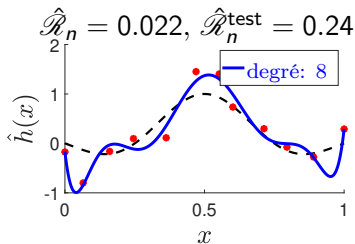
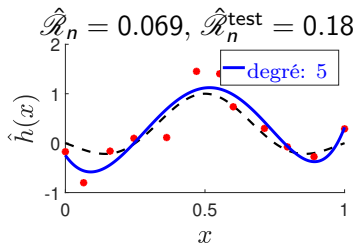
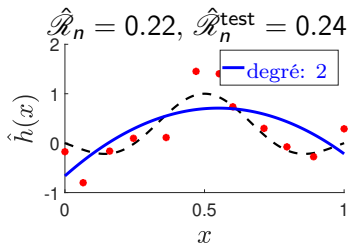
Example. Polynomial regression with $x \in \mathbb{R}$, degree $\leq J$:

$$h(x) = \beta_0 + \beta_1 x + \dots + \beta_J x^J,$$

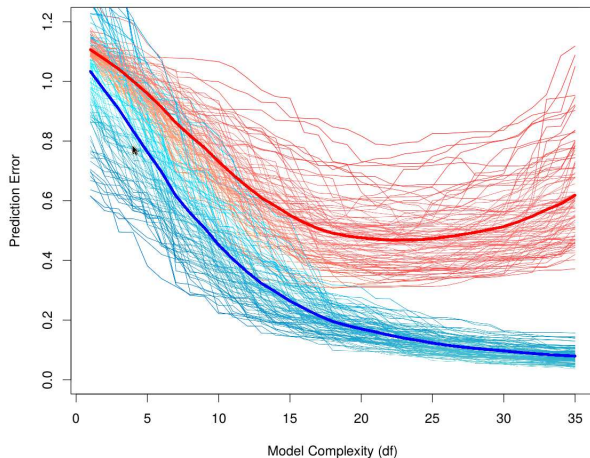
with $J = 2, 5, 8, 11$.

Recall that, in linear regression, the empirical risk has a downward bias proportional to the number of parameters in the model.

Example: polynomial regression



Understanding over-fitting: simulations



Blue: empirical risk $\hat{\mathcal{R}}_n$ / Red: error on the test set

Figure from Hastie, Tibshirani & Friedman (2017).
The Elements of Statistical Learning (12th edition), Springer.

Let's recapitulate. . .

Problem. We want to estimate the error to choose \mathcal{H} or λ but. . .

- ▶ it should be done neither on the **training data**
(\Rightarrow **over-fitting** problem),
- ▶ nor on the test data
(\Rightarrow bias in the final estimation of the generalization error).

Let's recapitulate...

Problem. We want to estimate the error to choose \mathcal{H} or λ but...

- ▶ it should be done neither on the training data
(\Rightarrow over-fitting problem),
- ▶ nor on the **test data**
(\Rightarrow **bias** in the final estimation of the generalization error).



Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

3.1 – Problem

3.2 – Cross validation

3.3 – AIC criterion

4 – Exercises and solutions

5 – Appendices

Solution: validation set

Idea: split the data in three sets

- ▶ **training** data: construct \hat{h} with given \mathcal{H}/λ ,
- ▶ **validation** set: choose \mathcal{H} , λ , etc.
- ▶ **test** data: estimate the generalization error.

Simple validation (hold-out)

training
(e.g., 60%)

validation
(e.g., 20%)

test
(e.g., 20%)

Solution: validation set

Idea: split the data in three sets

- ▶ training data: construct \hat{h} with given \mathcal{H}/λ ,
- ▶ validation set: choose \mathcal{H} , λ , etc.
- ▶ test data: estimate the generalization error.

Simple validation (hold-out)

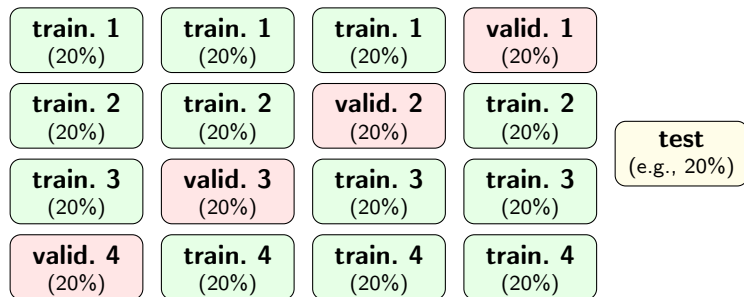
training
(e.g., 60%)

validation
(e.g., 20%)

test
(e.g., 20%)

Better validation: the cross validation method

k-fold cross-validation, here with $k = 4$:



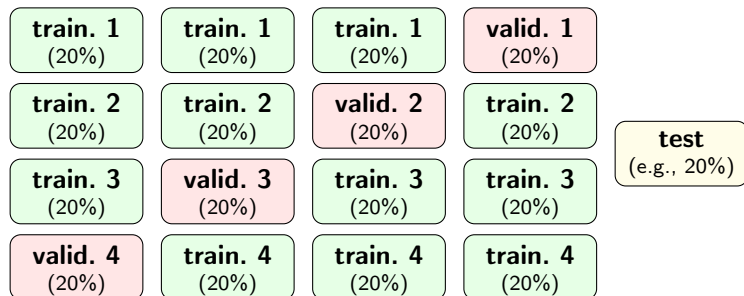
➡ the error is averaged over the k validation sets.

Special case: leave-one-out cross validation

▶ $k = n$ blocks (of size $n/k = 1$).

Better validation: the cross validation method

k -fold cross-validation, here with $k = 4$:



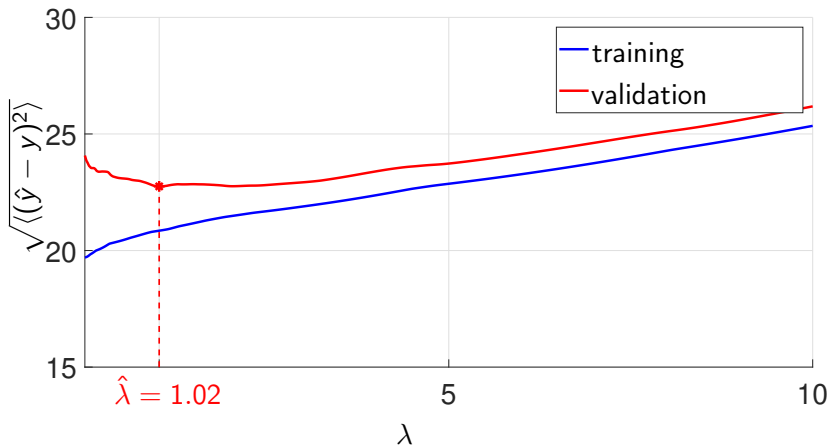
➡ the error is averaged over the k validation sets.

Special case: **leave-one-out** cross validation

▶ $k = n$ blocks (of size $n/k = 1$).

“Ozone” example: LASSO / choice of λ

- ▶ Predictor: LASSO regression using all variables and their interactions
- ▶ $\hat{\lambda}$ obtained by CV (LOO)



“Ozone” example: interactions

- ▶ We add variables of the form $X^{(j)}X^{(j')}$ and $X^{(j)}X^{(j')}X^{(j'')}$.
- ▶ LASSO regression (L^1 penalty).
- ▶ Hyper-parameter λ estimated through 10-fold CV.

model	$X^{(j)}$	$X^{(j)} X^{(j')}$	$X^{(j)} X^{(j')} X^{(j'')}$
total number of variables	7	35	119
number of selected variables ($\beta_j \neq 0$)	4	9	8
\sqrt{MSE} CV (10-fold)	49.1	41.5	33.0
selected variables	MOCAGE TEMPE NO VentANG	MOCAGE TEMPE NO2 MOCAGE · TEMPE TEMPE² TEMPE · MH2O TEMPE · NO2 NO2 · VentANG VentANG · VentANG	MOCAGE TEMPE NO2 MOCAGE · TEMPE TEMPE ² TEMPE · RMH2O TEMPE² · MOCAGE VentANG ² · TEMPE

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

3.1 – Problem

3.2 – Cross validation

3.3 – AIC criterion

4 – Exercises and solutions

5 – Appendices

Another approach to model selection: the AIC criterion

Assumption: parametric statistical models \mathcal{M}_j for $P^{Y|X}$.

Denote by $\hat{\theta}_j^{\text{MLE}}$ the MLE of θ in model \mathcal{M}_j .

Then the AIC criterion can also be used for model selection:

$$\hat{j} = \operatorname{argmin} \operatorname{AIC}(j), \quad \operatorname{AIC}(j) = -2 \ln \mathcal{L} \left(\hat{\theta}_j^{\text{MLE}}; \underline{X}, \underline{Y} \right) + 2k_j,$$

with k_j the number of parameters in model \mathcal{M}_j .

⇒ see PC for a partial justification (OLS linear regression)

Another approach to model selection: the AIC criterion

Assumption: parametric statistical models \mathcal{M}_j for $P^{Y|X}$.

Denote by $\hat{\theta}_j^{\text{MLE}}$ the MLE of θ in model \mathcal{M}_j .

Then the AIC criterion can also be used for model selection:

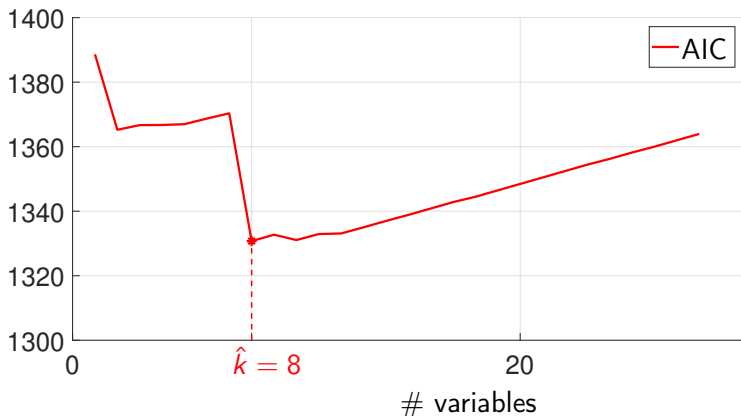
$$\hat{j} = \operatorname{argmin} \operatorname{AIC}(j), \quad \operatorname{AIC}(j) = -2 \ln \mathcal{L} \left(\hat{\theta}_j^{\text{MLE}}; \underline{X}, \underline{Y} \right) + 2k_j,$$

with k_j the number of parameters in model \mathcal{M}_j .

⇒ see PC for a partial justification (OLS linear regression)

“Ozone” example: AIC

- Predictor obtained by the ordinary least squares method, on an increasing number of variables
(linear terms first, then interactions)



Summary and preview

We have seen and will practice in PC 8:

- ▶ Ridge and LASSO regularization for penalized linear regression;
- ▶ the problem of estimating the generalization error (risk);
- ▶ the cross-validation method for hyper-parameter tuning and model selection.

We will cover in the last lecture:

- ▶ the challenges of unsupervised learning;
- ▶ principal component analysis (PCA) for dimension reduction;
- ▶ K -means algorithm for clustering.

Summary and preview

We have seen and will practice in PC 8:

- ▶ Ridge and LASSO regularization for penalized linear regression;
- ▶ the problem of estimating the generalization error (risk);
- ▶ the cross-validation method for hyper-parameter tuning and model selection.

We will cover in the last lecture:

- ▶ the challenges of unsupervised learning;
- ▶ principal component analysis (PCA) for dimension reduction;
- ▶ K -means algorithm for clustering.

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

4.1 – Questions

4.2 – Solutions

5 – Appendices

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

4.1 – Questions

4.2 – Solutions

5 – Appendices

Exercise 1 (Penalized regression)

Let X_1, \dots, X_n represent the examples, taking values in \mathbb{R}^p , and Y_1, \dots, Y_n be the labels, taking values in \mathbb{R} . The relationship between Y_i and X_i is given by:

$$Y_i = \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)} + \varepsilon_i,$$

where β is the parameter vector to be estimated, and ε_i is a random variable following $\mathcal{N}(0, \sigma^2)$, independent of X_i .

We aim to estimate β by minimizing a criterion of the form

$$\frac{1}{2} \sum_{i=1}^n \left(Y_i - \beta^\top X_i \right)^2 + \lambda \mathcal{P}(\beta) \quad (1)$$

where \mathcal{P} is a penalty term, and $\lambda \geq 0$ is a hyper-parameter.

We denote $X = [X_1 \dots X_n]^\top$, the $n \times p$ matrix containing the observations. **We are considering the case where $X^\top X = I_p$.**

Question

- 1 Give the expression of the estimator when $\lambda = 0$. Denote this estimator $\hat{\beta}$.
- 2 We consider a penalty of the form $\mathcal{P}(\beta) = \|\beta\|_2^2$. Give the expression of this estimator, denoted $\hat{\beta}^R$, and deduce that there exists a constant $c_{1,\lambda}$ (to be specified) such that $\hat{\beta}_j^R = c_{1,\lambda} \hat{\beta}_j$, $j = 1, \dots, p$.

Question

- ③ We consider a penalty of the form $\mathcal{P}(\beta) = \|\beta\|_1$.
To begin with, demonstrate that the minimum on \mathbb{R} of the function

$$f : \alpha \mapsto \frac{1}{2}(x - \alpha)^2 + \lambda |\alpha|$$

is achieved at $\alpha^* = \text{sign}(x) \max(0, |x| - \lambda)$.

- ④ Deduce the solution of the optimization problem (1) for $\mathcal{P}(\beta) = \|\beta\|_1$, which will be expressed in terms of $\hat{\beta}$. Denote this estimator $\hat{\beta}^L$.

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

4.1 – Questions

4.2 – Solutions

5 – Appendices

- ① We recognize the least squares criterion, and we have:

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

- ② This corresponds to ridge regression..

$$\begin{aligned}\hat{\beta}^R &= (X^T X + 2\lambda I)^{-1} X^T Y \\ &= (1 + 2\lambda)^{-1} \hat{\beta}\end{aligned}$$

Therefore $\hat{\beta}_j^R = (1 + 2\lambda)^{-1} \hat{\beta}_j$.

- ③ The function f is not differentiable, but it is differentiable at every point $\alpha \neq 0$ and continuous at $\alpha = 0$. Thus, we can determine its minimum by analyzing its variations using the sign of the derivative, as if it were differentiable everywhere. The derivative at every $\alpha \neq 0$ is given by

$$f'(\alpha) = \begin{cases} \alpha - x + \lambda & \text{si } \alpha > 0, \\ \alpha - x - \lambda & \text{si } \alpha < 0, \end{cases}$$

hence

$$f'(\alpha) > 0 \quad \Leftrightarrow \quad (\alpha > x - \lambda \text{ et } \alpha > 0) \text{ ou } (\alpha > x + \lambda \text{ et } \alpha < 0). \quad (2)$$

- ③ Let's consider, for example, $x > 0$. Then, the second case in the right-hand side of (2) is impossible, and we're left with:

$$f'(\alpha) > 0 \Leftrightarrow \alpha > x - \lambda \text{ et } \alpha > 0 \Leftrightarrow \alpha > \max(0, x - \lambda). \quad (3)$$

Similarly, still assuming $x > 0$,

$$\begin{aligned} f'(\alpha) < 0 &\Leftrightarrow (\alpha < x - \lambda \text{ et } \alpha > 0) \text{ ou } (\alpha < x + \lambda \text{ et } \alpha < 0) \\ &\Leftrightarrow (0 < \alpha < \max(0, x - \lambda)) \text{ ou } (\alpha < 0) \\ &\Leftrightarrow (\alpha < \max(0, x - \lambda)) \text{ et } (\alpha \neq 0). \end{aligned}$$

Thus, f strictly decreases to the left of $\max(0, x - \lambda)$, and strictly increases to the right, which concludes the case $x > 0$. The case $x < 0$ follows similarly.

- 4 Here, we'll manipulate the initial optimization problem to reduce it to the optimization problem from the previous question.:

$$\begin{aligned}\hat{\beta}^L &= \operatorname{argmin}_{\beta} \frac{1}{2} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \\ &= \operatorname{argmin}_{\beta} \frac{1}{2} \|Y - X\hat{\beta} + X\hat{\beta} - X\beta\|^2 + \lambda \|\beta\|_1 \\ &= \operatorname{argmin}_{\beta} \frac{1}{2} \left\{ \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\beta\|^2 \right\} + \lambda \|\beta\|_1\end{aligned}$$

The cross product vanishes because the residual $(Y - X\hat{\beta})$ is, by construction, orthogonal to any linear combination of columns of X , thus $(Y - X\hat{\beta})^\top (X\hat{\beta} - X\beta) = 0$.

- 4 Since the first term is independent of β , we have:

$$\begin{aligned}\hat{\beta}^L &= \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|X\hat{\beta} - X\beta\|^2 + \lambda \|\beta\|_1 \\ &= \underset{\beta}{\operatorname{argmin}} \frac{1}{2} (\hat{\beta} - \beta)^\top X^\top X (\hat{\beta} - \beta) + \lambda \|\beta\|_1 \\ &= \underset{\beta}{\operatorname{argmin}} \frac{1}{2} (\hat{\beta} - \beta)^\top (\hat{\beta} - \beta) + \lambda \|\beta\|_1 \\ &= \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \sum_{j=1}^p (\hat{\beta}_j - \beta_j)^2 + \lambda |\beta_j|\end{aligned}$$

The problem is separable and, from the previous question, we have:

$$\hat{\beta}_j^L = \operatorname{sign}(\hat{\beta}_j) \max(0, |\hat{\beta}_j| - \lambda)$$

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

5.1 – Model building: feature engineering

Lecture outline

1 – Regularized regression (or classification): penalization

2 – Estimation of the risk (generalization error)

3 – Hyper-parameters, model selection

4 – Exercises and solutions

5 – Appendices

5.1 – Model building: feature engineering

Non-linearities in linear models...

If the empirical risk $\hat{\mathcal{R}}(\hat{h})$ is high, several possible causes:

- ▶ **noise**: intrinsic difficulty in predicting Y
 - irreducible **statistical error**.
- ▶ non-linearity of the optimal predictor wrt the $X^{(j)}$'s
 - reducible approximation error.

Possible workaround: $x^{(1)}, \dots, x^{(p)} \mapsto \tilde{x}^{(1)}, \dots, \tilde{x}^{(q)}$

- ▶ with $\tilde{x}^{(j)}$ function of $x^{(1)}, \dots, x^{(p)}$.
- ▶ The model is still linear with respect to β .

Non-linearities in linear models...

If the empirical risk $\hat{\mathcal{R}}(\hat{h})$ is high, several possible causes:

- ▶ noise: intrinsic difficulty in predicting Y
 - ▮ irreducible statistical error.
- ▶ **non-linearity** of the optimal predictor wrt the $X^{(j)}$'s
 - ▮ reducible **approximation error**.

Possible workaround: $x^{(1)}, \dots, x^{(p)} \mapsto \tilde{x}^{(1)}, \dots, \tilde{x}^{(q)}$

- ▶ with $\tilde{x}^{(j)}$ function of $x^{(1)}, \dots, x^{(p)}$.
- ▶ The model is still linear with respect to β .

Non-linearities in linear models...

If the empirical risk $\hat{\mathcal{R}}(\hat{h})$ is high, several possible causes:

- ▶ noise: intrinsic difficulty in predicting Y
 - ▢ irreducible statistical error.
- ▶ non-linearity of the optimal predictor wrt the $X^{(j)}$'s
 - ▢ reducible approximation error.

Possible workaround: $x^{(1)}, \dots, x^{(p)} \mapsto \tilde{x}^{(1)}, \dots, \tilde{x}^{(q)}$

- ▶ with $\tilde{x}^{(j)}$ function of $x^{(1)}, \dots, x^{(p)}$.
- ▶ The model is still **linear with respect to β** .

Examples

A few examples:

- ▶ **scalar transformations**: $\ln(x^{(j)})$, $\sqrt{x^{(j)}}$, $(x^{(j)})^k \dots$
- ▶ **interactions** (here, of order two): $x^{(j)}x^{(k)}$, $j \neq k$,
- ▶ higher-order interactions,
- ▶ (truncated) expansion in a basis...



if $q \gg p$, **risk of over-fitting**.

Remarks: *feature engineering*

- ▶ Proposing new relevant variables
 - ▢ domain expertise (or model selection...?)
- ▶ The same principle can be used to *reduce* dimension
 - ▢ features extraction.

Examples

A few examples:

- ▶ scalar transformations: $\ln(x^{(j)})$, $\sqrt{x^{(j)}}$, $(x^{(j)})^k \dots$
- ▶ interactions (here, of order two): $x^{(j)}x^{(k)}$, $j \neq k$,
- ▶ higher-order interactions,
- ▶ (truncated) expansion in a basis...



if $q \gg p$, risk of over-fitting.

Remarks: *feature engineering*

- ▶ Proposing new relevant variables
 - ▢ *domain expertise* (or model selection...?)
- ▶ The same principle can be used to *reduce* dimension
 - ▢ *features extraction*.

Expansion in a basis

Principle

Let $\{\psi_m\}_{m>0}$ be a function basis of $L^2(\mathcal{X})^\dagger$.

Consider $\tilde{X}^{(m)} = \psi_m(X)$, $m = 1, \dots, M$

⇒ truncated expansion in the basis $\{\psi_m\}$.

Examples of bases (preferably orthogonal):

- ▶ polynomial bases,
- ▶ wavelet bases,
- ▶ Fourier bases. . .

[†] or any other function space assumed to contain the optimal predictor h^* .

Expansion in a basis

Principle

Let $\{\psi_m\}_{m>0}$ be a function basis of $L^2(\mathcal{X})^\dagger$.

Consider $\tilde{X}^{(m)} = \psi_m(X)$, $m = 1, \dots, M$

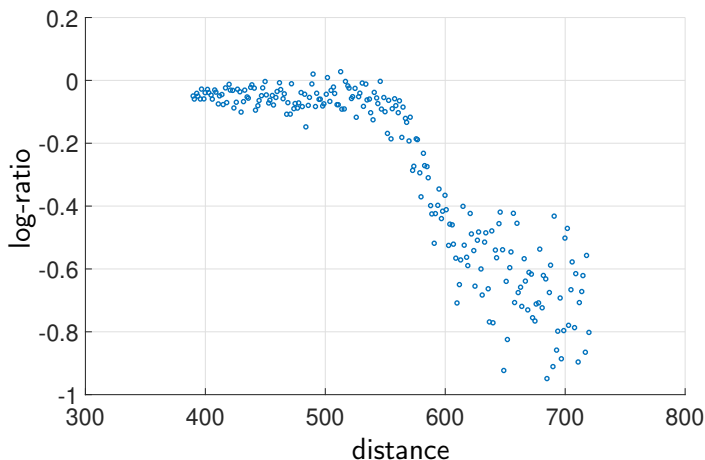
▮ truncated expansion in the basis $\{\psi_m\}$.

Examples of bases (preferably orthogonal):

- ▶ polynomial bases,
- ▶ wavelet bases,
- ▶ Fourier bases. . .

[†] or any other function space assumed to contain the optimal predictor h^* .

Example: LIDAR data



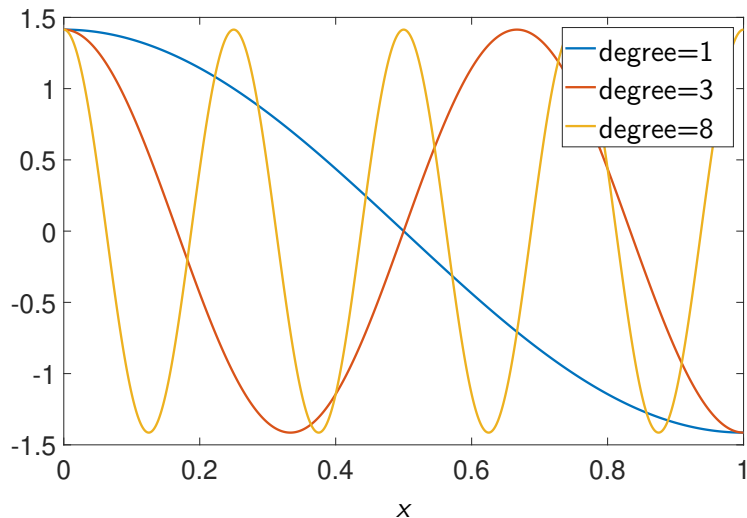
x-axis: distance travelled before the light is reflected back to its source

y-axis: logarithm of the ratio of received light from two laser sources

Data obtained from <http://matt-wand.utsacademics.info/webspr/lidar.html>

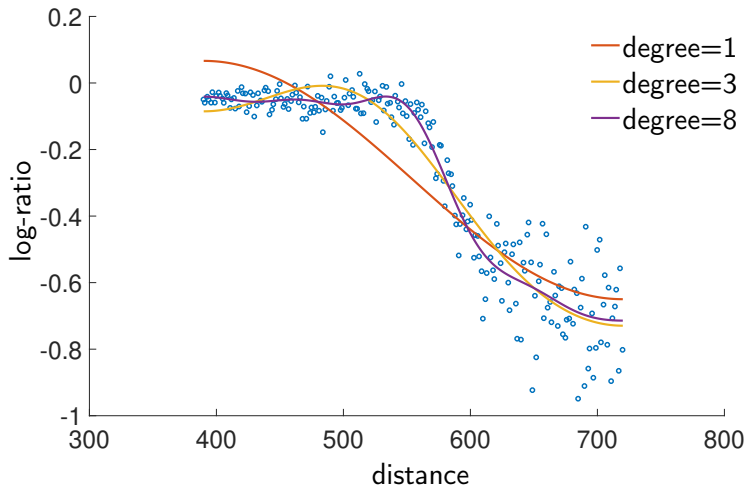
LIDAR: Light Detection And Ranging

Basis of orthogonal cosines (basis of $L^2([0, 1])$)

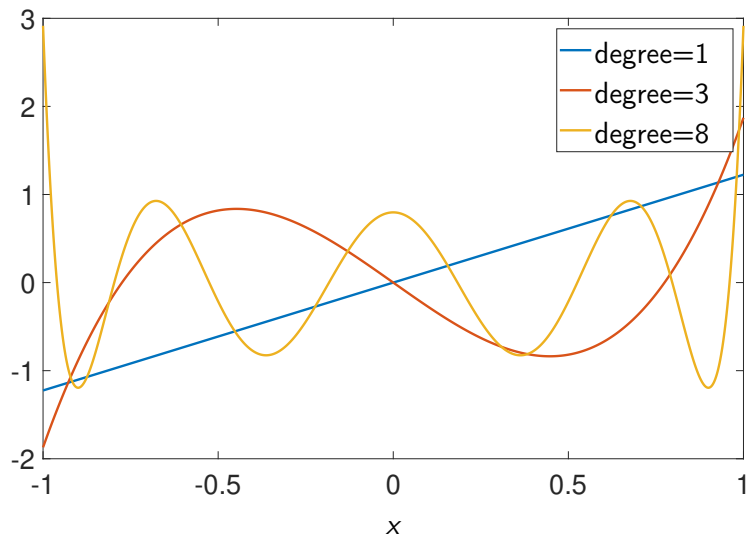


Example: LIDAR data (cont'd)

Quadratic loss + basis of cosines

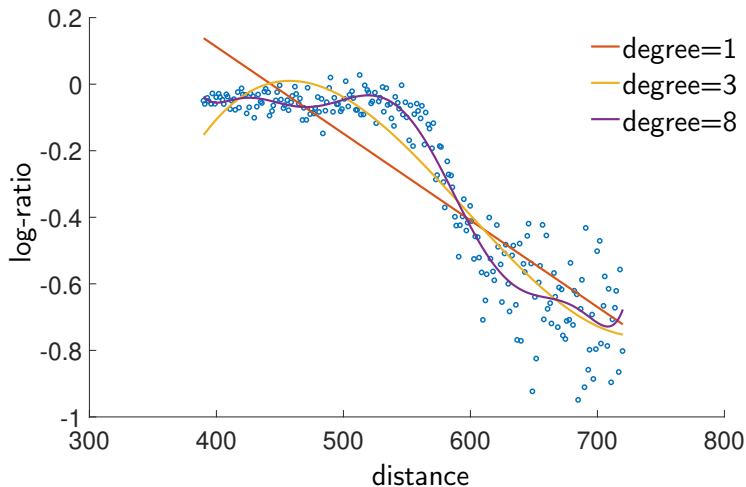


Legendre polynomials (orthonormal basis of $L^2([-1, 1])$)

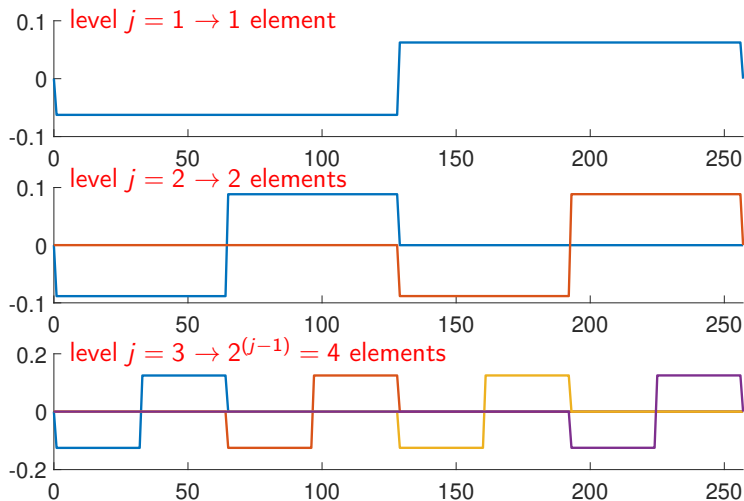


Example: LIDAR data (cont'd)

Quadratic loss + Legendre polynomials

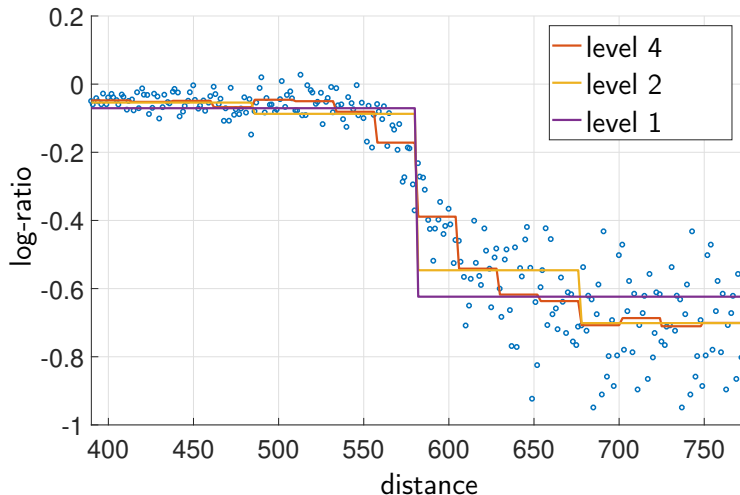


Haar wavelet basis



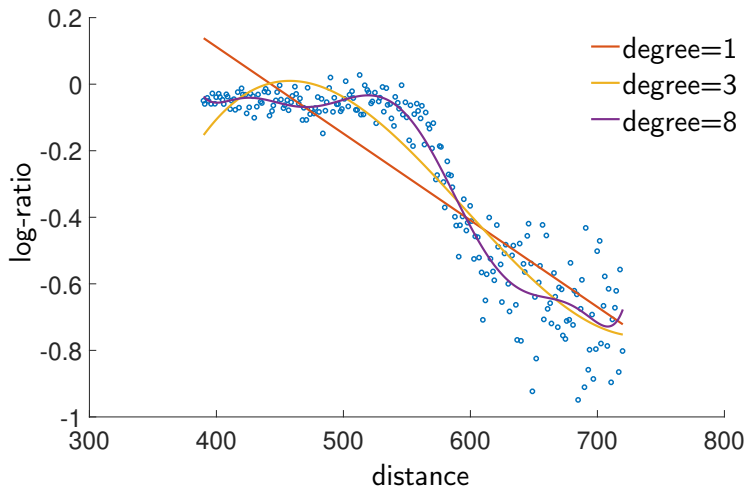
Example: LIDAR data (cont'd)

Quadratic loss + Haar wavelets



Example: LIDAR data (cont'd)

Quadratic loss + Legendre polynomials



Example: LIDAR data (cont'd)

Model selection

