CentraleSupélec

Statistique et apprentissage

Chargés de cours (ordre alphabétique) :

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

[†] Coordinateur du cours

Cours 8/9

Régularisation et sélection de modèles

Objectifs du cours 8

- Introduire la régression/classification pénalisée
- Savoir estimer l'erreur de généralisation
- Savoir déterminer les hyper-paramètres d'un modèle ou choisir un modèle

Plan du cours

Plan du cours

Plan du cours

Limites des « moindres carrés ordinaires »

Rappel : matrice $\underline{X}$ de taille #individus $\times$ #variables $(n \times (p+1))$.

Situations critiques pour la régression linéaire :

- lorsque la matrice $\underline{X}^\top \underline{X}$ n'est pas inversible
- ou mal conditionnée

Cas typiques

1. lorsque le nombre de variables est grand $(p+1 > n)$,
2. lorsque il y a de fortes corrélations entre les variables explicatives.

Limites des « moindres carrés ordinaires »

Rappel : matrice $\underline{X}$ de taille #individus $\times$ #variables $(n \times (p+1))$.

Situations critiques pour la régression linéaire :

- lorsque la matrice $\underline{X}^\top \underline{X}$ n'est pas inversible
- ou mal conditionnée

Cas typiques

1. lorsque le nombre de variables est grand $(p+1 > n)$,
2. lorsque il y a de fortes corrélations entre les variables explicatives.

Limites des « moindres carrés ordinaires »

Rappel : matrice $\underline{X}$ de taille #individus $\times$ #variables $(n \times (p+1))$.

Situations critiques pour la régression linéaire :

- lorsque la matrice $\underline{X}^\top \underline{X}$ n'est pas inversible
- ou mal conditionnée

Cas typiques

1. lorsque le nombre de variables est grand $(p+1 > n)$,
2. lorsque il y a de fortes corrélations entre les variables explicatives.

Exemple : $p > n$

	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 on_pa
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.
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.
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.
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.
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.
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.
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.
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.
19	id_18	12980	1.05E-05	1.65E+02	6.86E+01	11195	7222	4.13E-01	1.61E-01	1.17E+02	7.09E-02	1.
20	id_19	6365	1.11E-05	7.83E+01	1.19E+02	1425	7755	1.27E-01	1.20E-01	9.85E+01	1.62E-02	1

Extrait d'un tableau de données décrivant, au moyen de $p = 451$ variables, l'écriture manuscrite de $n = 174$ personnes dont certaines sont atteintes de la maladie d'Alzheimer.

Il est courant, dans le domaine médical notamment, d'avoir plus de descripteurs que d'individus.

D'après l'étude *Diagnosing Alzheimer's disease from on-line handwriting : A novel dataset and performance benchmarking*, N. D. Cilia et al., 2022, et son jeu de données associé : UCI Machine Learning Repository. https://doi.org/10.24432/C55D0K.

Exemple : $p > n$

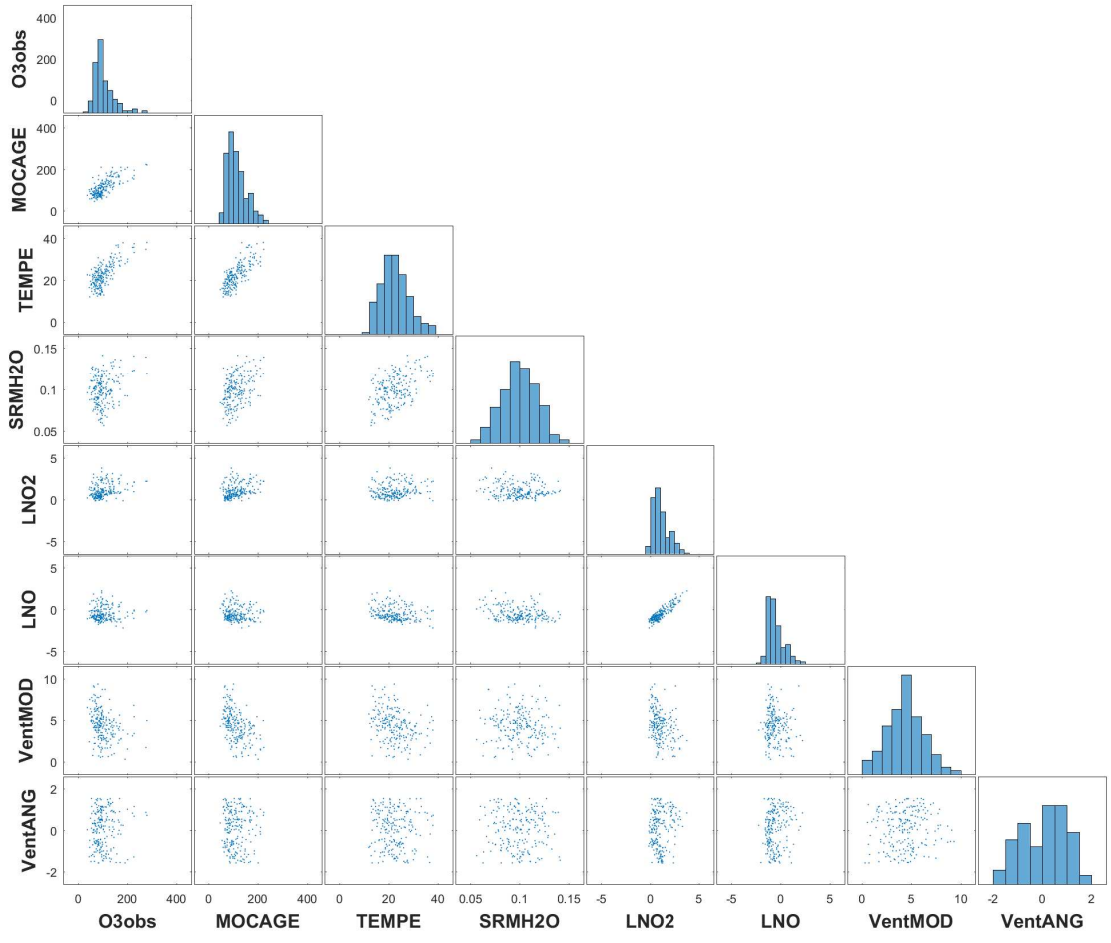
	A	B	C	D	E	F	G	H	I	J	K	
1	ID	air_time1	disp_index1	gmrt_in_air1	gmrt_on_paper1	max_x_extension1	max_y_extension1	mean_acc_in_air1	mean_acc_on_paper1	mean_gmrt1	mean_jerk_in_air1	mean on_pa
2	id_1	5160	1.25E-05	1.21E+02	8.69E+01	957	6601	3.62E-01	2.17E-01	1.04E+02	5.18E-02	2.
3	id_2	51980	1.60E-05	1.15E+02	8.34E+01	1694	6998	2.73E-01	1.45E-01	9.94E+01	3.98E-02	1.
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.
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.
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.
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.
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.
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.
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.
19	id_18	12980	1.05E-05	1.65E+02	6.86E+01	11195	7222	4.13E-01	1.61E-01	1.17E+02	7.09E-02	1.
20	id_19	6365	1.11E-05	7.83E+01	1.19E+02	1425	7755	1.27E-01	1.20E-01	9.85E+01	1.62E-02	1

Extrait d'un tableau de données décrivant, au moyen de $p = 451$ variables, l'écriture manuscrite de $n = 174$ personnes dont certaines sont atteintes de la maladie d'Alzheimer.

Il est courant, dans le domaine médical notamment, d'avoir plus de descripteurs que d'individus.

D'après l'étude *Diagnosing Alzheimer's disease from on-line handwriting : A novel dataset and performance benchmarking*, N. D. Cilia et al., 2022, et son jeu de données associé : UCI Machine Learning Repository. `https://doi.org/10.24432/C55D0K`.

Exemple : forte corrélation entre variables explicatives



Exemple « Ozone » $\longrightarrow$ corrélation entre les variables NO et NO2

Exemple : forte corrélation. . . (suite)

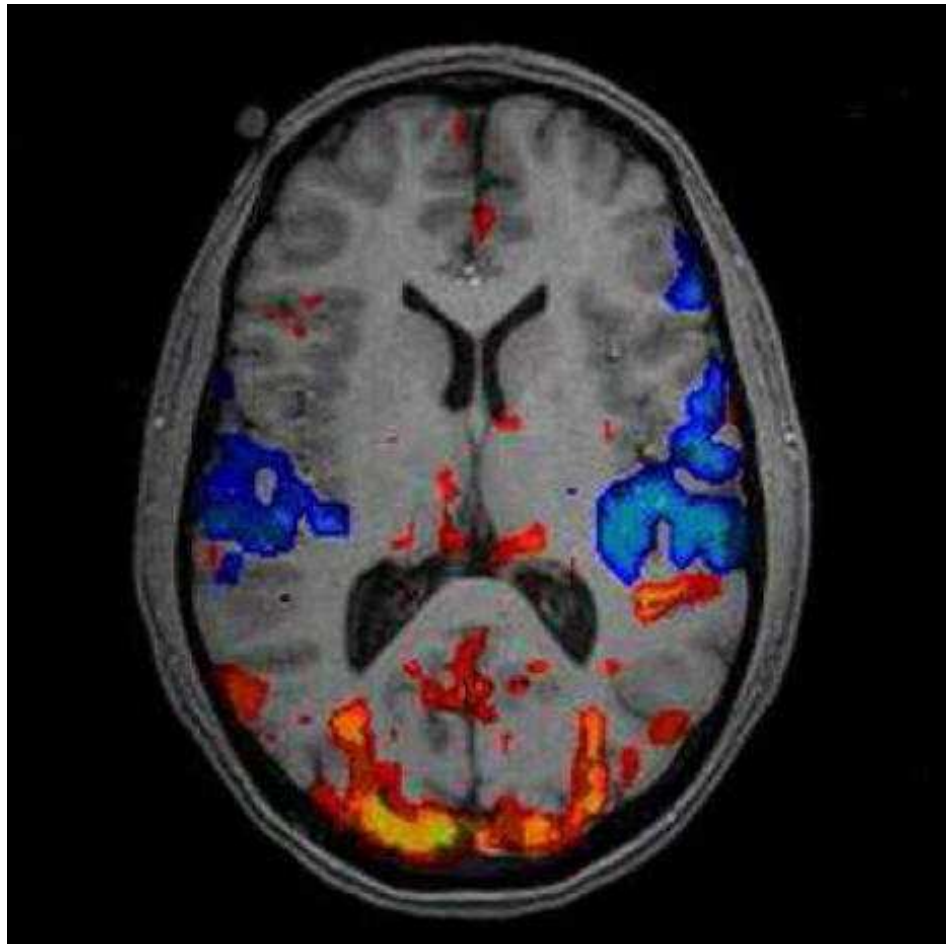
Vecteur $\hat{\beta}$ obtenu par régression linéaire :

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 :

- Le coefficient négatif associé à NO2 est surprenant
  ➠ interprétation hasardeuse des coefficients de régression
- Les variables les moins influentes (petits coefficients) pourraient être supprimées du modèle ?

Exemple : $p \gg n$ **et** forte corrélation



Imagerie par Résonance Magnétique Fonctionnelle (IRMf),
$p \approx 300000$ *voxels*

Typiquement, $n \approx 10$ ou 100 !

Solution possible : régression pénalisée

Au critère des moindres carrées (SCR), on ajoute une pénalité :

$$\hat{\beta} = \text{argmin}_\beta \quad \underbrace{\|\underline{Y} - \underline{X}\beta\|^2}_{\text{« attache » aux données}} + \underbrace{\lambda}_{\text{hyperparamètre}} \underbrace{\Omega(\beta)}_{\text{pénalité}} . \qquad (\star)$$

NB : ici et dans la suite, $\|\cdot\|$ désigne la norme euclidienne.

Intérêts de la pénalité.

- rendre la solution de $(\star)$ unique,
- apporter de l'a priori (lien avec l'approche bayésienne),
- pouvoir traiter les cas où $p \gg n$,
- faciliter l'interprétation des coefficients de régression.

Solution possible : régression pénalisée

Au critère des moindres carrées (SCR), on ajoute une pénalité :

$$\hat{\beta} = \text{argmin}_\beta \quad \underbrace{\|\underline{Y} - \underline{X}\beta\|^2}_{\text{« attache » aux données}} \quad + \quad \underbrace{\lambda}_{\text{hyperparamètre}} \quad \underbrace{\Omega(\beta)}_{\text{pénalité}} . \qquad (\star)$$

NB : ici et dans la suite, $\|\cdot\|$ désigne la norme euclidienne.

Intérêts de la pénalité.

- rendre la solution de $(\star)$ unique,
- apporter de l'a priori (lien avec l'approche bayésienne),
- pouvoir traiter les cas où $p \gg n$,
- faciliter l'interprétation des coefficients de régression.

Plan du cours

Régression ridge

Pénalité

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

(en général, on ne pénalise pas β_0)

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

On montre que (✏ voir TD) :

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

➠ Lorsque $\lambda \nearrow$, le conditionnement de $\left(\underline{X}^\top \underline{X} + \lambda I_{p+1}\right)$ s'améliore.

Remarque : $\hat{\beta}^{\text{RIDGE}}$ admet une interprétation bayésienne
(✏ voir TD également).

Régression ridge

Pénalité

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

(en général, on ne pénalise pas β_0)

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

On montre que (✏ voir TD) :

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

➠ Lorsque $\lambda \nearrow$, le conditionnement de $\left(\underline{X}^\top \underline{X} + \lambda I_{p+1}\right)$ s'améliore.

Remarque : $\hat{\beta}^{\text{RIDGE}}$ admet une interprétation bayésienne (✏ voir TD également).

Régression ridge

Pénalité

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

(en général, on ne pénalise pas β_0)

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

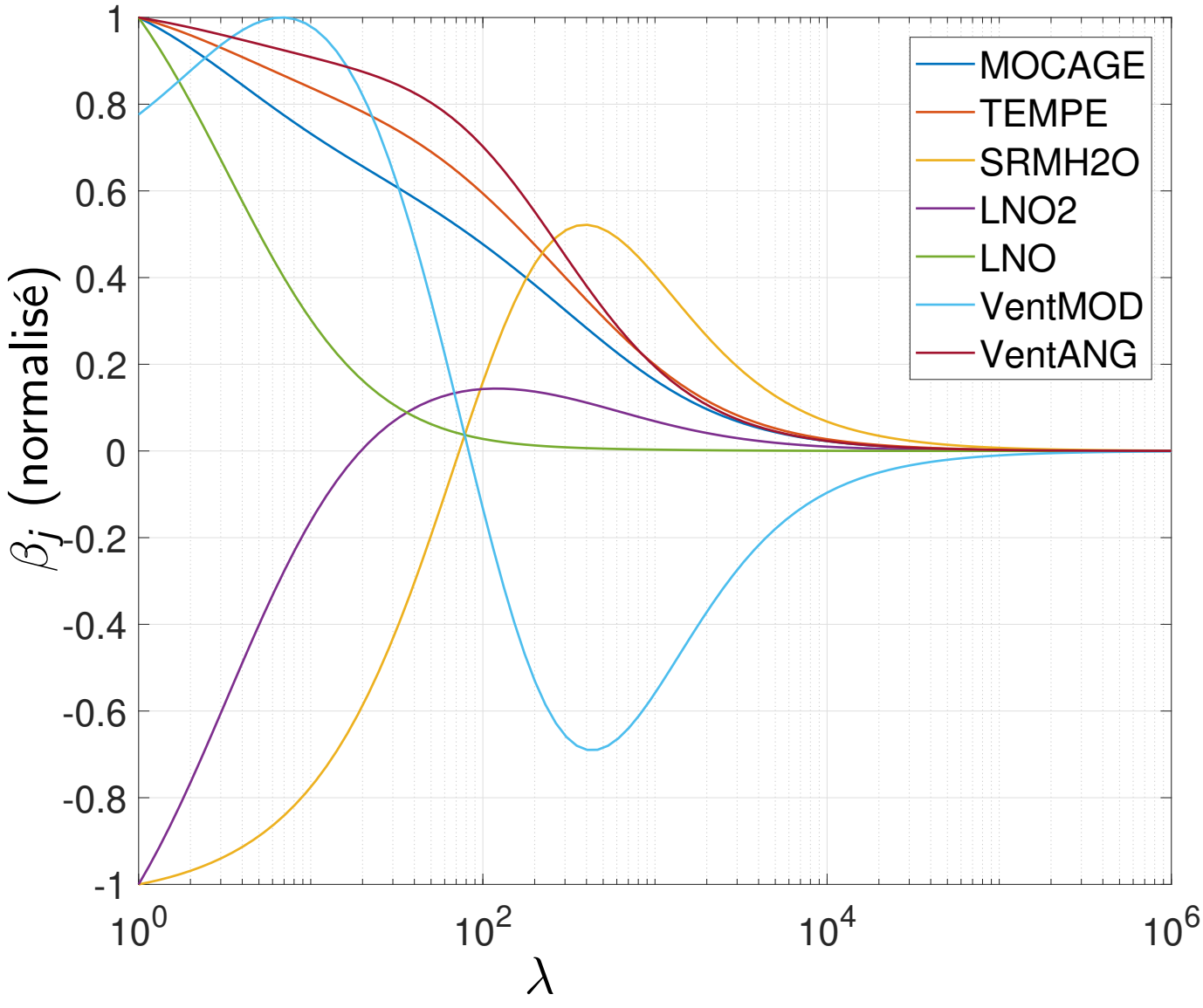
On montre que (✏ voir TD) :

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

➠ Lorsque $\lambda \nearrow$, le conditionnement de $\left(\underline{X}^\top \underline{X} + \lambda I_{p+1}\right)$ s'améliore.

Remarque : $\hat{\beta}^{\text{RIDGE}}$ admet une interprétation bayésienne (✏ voir TD également).

Exemple « Ozone » : Évolution de $\hat{\beta}^{RIDGE}$ en fonction de λ



Plan du cours

Régression LASSO

Pénalité

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

(en général, on ne pénalise pas β_0)

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

Minimisation du critère.

- pas d'expression explicite de $\hat{\beta}^{\text{LASSO}}$ (sauf dans certains cas, exercice 1)
- algorithmes spécifiques

Régression LASSO

Pénalité

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

(en général, on ne pénalise pas β_0)

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

Minimisation du critère.

- pas d'expression explicite de $\hat{\beta}^{\text{LASSO}}$ (sauf dans certains cas, exercice 1)
  - algorithmes spécifiques

Régression LASSO : reformulation

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

- Soit $\hat{\beta}$ l'estimateur de β au sens des moindres carrés (ordinaires) :

$$\hat{\beta}^{\text{LASSO}} = \hat{\beta} \qquad \text{pour } \lambda = 0$$

- Comme $\|\underline{Y} - \underline{X}\beta\|^2 = \|\underline{X}(\beta - \hat{\beta})\|^2 + c$, il vient :

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

- Reformulation à l'aide d'une contrainte : on peut montrer qu'il existe $c_\lambda \in \mathbb{R}^+$ telle que

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

Régression LASSO : reformulation

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

- Soit $\hat{\beta}$ l'estimateur de β au sens des moindres carrés (ordinaires) :

$$\hat{\beta}^{\text{LASSO}} = \hat{\beta} \qquad \text{pour } \lambda = 0$$

- Comme $\|\underline{Y} - \underline{X}\beta\|^2 = \|\underline{X}(\beta - \hat{\beta})\|^2 + c$, il vient :

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

- Reformulation à l'aide d'une contrainte : on peut montrer qu'il existe $c_\lambda \in \mathbb{R}^+$ telle que

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

Régression LASSO : reformulation

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

- Soit $\hat{\beta}$ l'estimateur de β au sens des moindres carrés (ordinaires) :

$$\hat{\beta}^{\text{LASSO}} = \hat{\beta} \qquad \text{pour } \lambda = 0$$

- Comme $\|\underline{Y} - \underline{X}\beta\|^2 = \|\underline{X}(\beta - \hat{\beta})\|^2 + c$, il vient :

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

- Reformulation à l'aide d'une contrainte : on peut montrer qu'il existe $c_\lambda \in \mathbb{R}^+$ telle que

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

Régression LASSO : explication intuitive

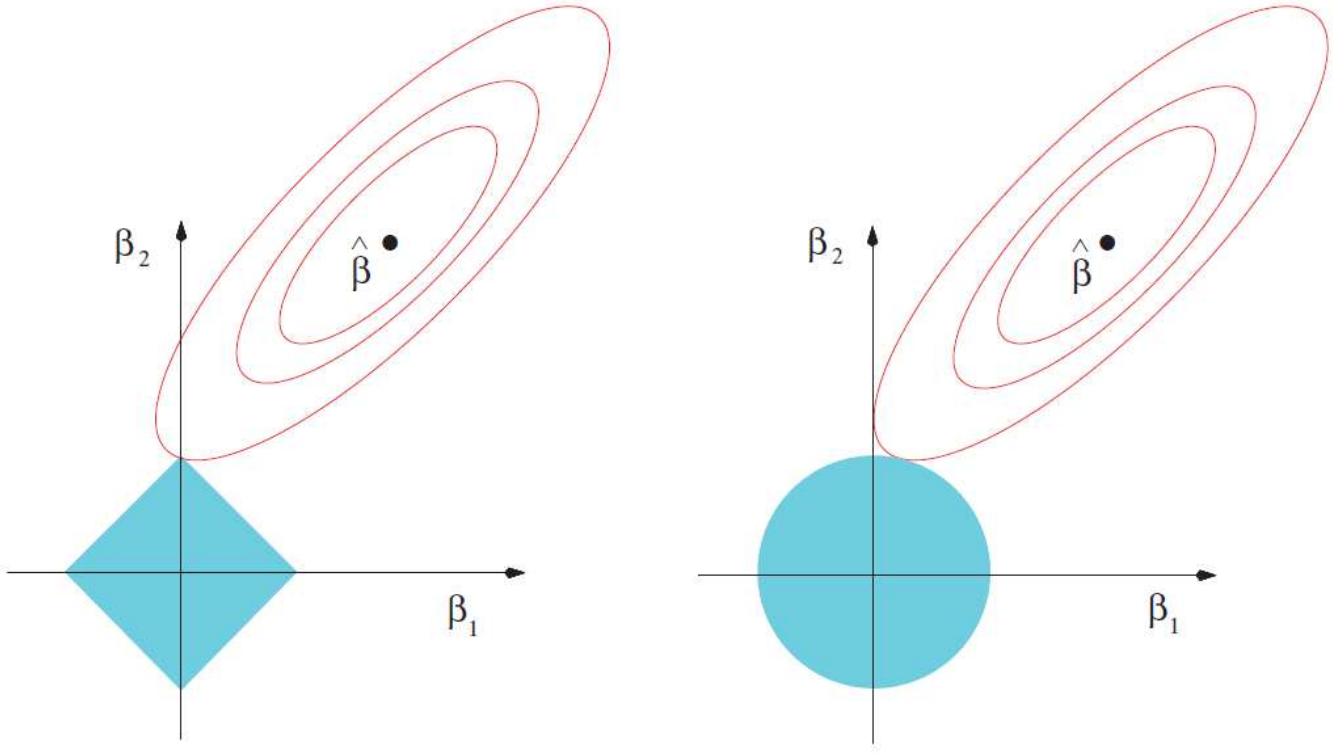
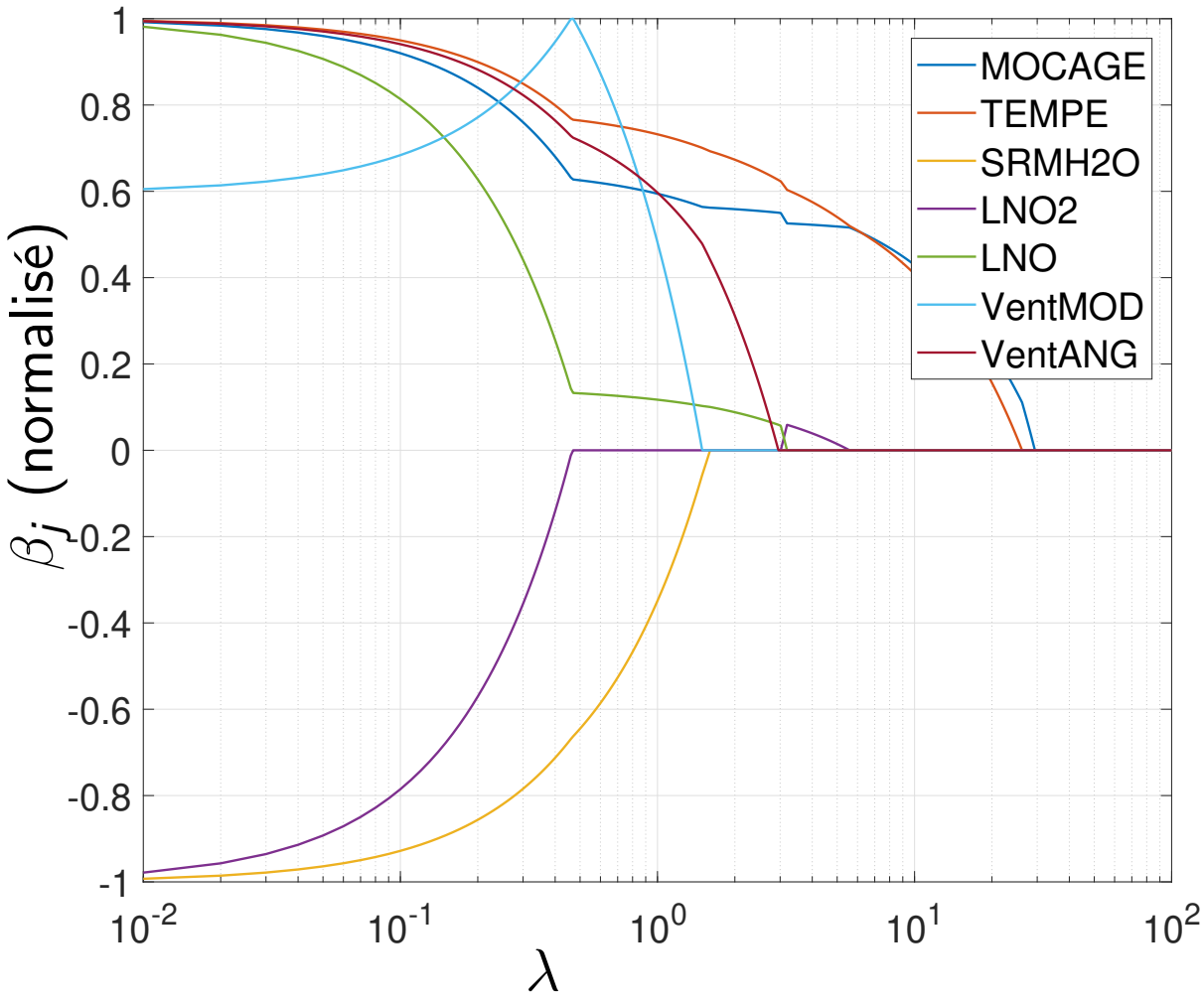


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are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

Exemple « Ozone » : $\hat{\beta}^{\text{LASSO}}$ en fonction de λ



Quand $\lambda \nearrow$, le nombre de coefficients égaux à 0 $\nearrow$

Exemple « Ozone » : $\hat{\beta}^{\text{LASSO}}$ pour différents λ

Avec $\lambda = 0$ (MCO)

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

➠ Le coefficient associé à NO2 peut sembler surprenant

Avec $\lambda = 0.5$

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

➠ Conservation d'une des deux variables les plus corrélées, facilite l'interprétation des coefficients

Avec $\lambda = 3$

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

➠ Mise à zéro progressive des coefficients restants

Choix de l'hyper-paramètre λ ?

Exemple « Ozone » : $\hat{\beta}^{\text{LASSO}}$ pour différents λ

Avec $\lambda = 0$ (MCO)

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

➠ Le coefficient associé à NO2 peut sembler surprenant

Avec $\lambda = 0.5$

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

➠ Conservation d'une des deux variables les plus corrélées,
facilite l'interprétation des coefficients

Avec $\lambda = 3$

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

➠ Mise à zéro progressive des coefficients restants

Choix de l'hyper-paramètre λ ?

Exemple « Ozone » : $\hat{\beta}^{\text{LASSO}}$ pour différents λ

Avec $\lambda = 0$ (MCO)

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

➠ Le coefficient associé à NO2 peut sembler surprenant

Avec $\lambda = 0.5$

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

➠ Conservation d'une des deux variables les plus corrélées, facilite l'interprétation des coefficients

Avec $\lambda = 3$

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

➠ Mise à zéro progressive des coefficients restants

Choix de l'hyper-paramètre λ ?

Exemple « Ozone » : $\hat{\beta}^{\text{LASSO}}$ pour différents λ

Avec $\lambda = 0$ (MCO)

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

➠ Le coefficient associé à NO2 peut sembler surprenant

Avec $\lambda = 0.5$

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

➠ Conservation d'une des deux variables les plus corrélées, facilite l'interprétation des coefficients

Avec $\lambda = 3$

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

➠ Mise à zéro progressive des coefficients restants

Choix de l'hyper-paramètre λ ?

Plan du cours

Plan du cours

Problème

Revenons à un cadre général (régression/classification).

Soit $\hat{h}$ un prédicteur $\mathcal{X} \to \mathcal{Y}$ appris à partir des données :

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

Rappel : étant donnée une fonction de perte L, on définit le risque, ou erreur de généralisation :

$$\begin{aligned} \mathscr{R}(\hat{h}) \;&=\; \mathbb{E}\left(L(Y, \hat{h}(X)) \;\middle|\; \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{h}(x)) \, \mathrm{P}^{X,Y}(\mathrm{d}x, \mathrm{d}y). \end{aligned}$$

Exemples : $L(y, \tilde{y}) = (y - \tilde{y})^2$, $L(y, \tilde{y}) = |y - \tilde{y}|$, $L(y, \tilde{y}) = \mathbb{1}_{y \neq \tilde{y}}$, $\ldots$

Problème

Comment estimer ce risque (qui dépend de la loi $\mathrm{P}^{X,Y}$ inconnue) ?

Problème

Revenons à un cadre général (régression/classification).

Soit $\hat{h}$ un prédicteur $\mathcal{X} \to \mathcal{Y}$ appris à partir des données :

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

Rappel : étant donnée une fonction de perte L, on définit le risque, ou erreur de généralisation :

$$\begin{aligned} \mathscr{R}(\hat{h}) \;&=\; \mathbb{E}\left(L(Y, \hat{h}(X)) \;\middle|\; \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{h}(x)) \, \mathrm{P}^{X,Y}(\mathrm{d}x, \mathrm{d}y). \end{aligned}$$

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

Problème

Comment estimer ce risque (qui dépend de la loi $\mathrm{P}^{X,Y}$ inconnue) ?

Problème

Revenons à un cadre général (régression/classification).

Soit $\hat{h}$ un prédicteur $\mathcal{X} \to \mathcal{Y}$ appris à partir des données :

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

Rappel : étant donnée une fonction de perte L, on définit le risque, ou erreur de généralisation :

$$\begin{aligned} \mathscr{R}(\hat{h}) \;&=\; \mathbb{E}\left(L(Y, \hat{h}(X)) \;\middle|\; \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{h}(x)) \, \mathrm{P}^{X,Y}(\mathrm{d}x, \mathrm{d}y). \end{aligned}$$

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

Problème

Comment estimer ce risque (qui dépend de la loi $\mathrm{P}^{X,Y}$ inconnue) ?

Rappel : risque empirique

On appelle risque empirique le risque

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

calculé avec $\mathrm{P}^{X,Y}$ égal à la mesure empirique $\hat{\mathrm{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$.

Question

Le risque empirique $\widehat{\mathscr{R}}_n$ est-il, en général, un « bon » estimateur du risque réel $\mathscr{R}(\hat{h})$?

⚠ double utilisation des données !

Intuition : Il est « dangereux » d'estimer le risque à partir de l'erreur commise sur les données ayant servi à construire $\hat{h}$...

Rappel : risque empirique

On appelle risque empirique le risque

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

calculé avec $\mathrm{P}^{X,Y}$ égal à la mesure empirique $\hat{\mathrm{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$.

Question

Le risque empirique $\widehat{\mathscr{R}}_n$ est-il, en général, un « bon » estimateur du risque réel $\mathscr{R}(\hat{h})$?

⚠ double utilisation des données !

Intuition : Il est « dangereux » d'estimer le risque à partir de l'erreur commise sur les données ayant servi à construire $\hat{h}$...

Rappel : risque empirique

On appelle risque empirique le risque

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

calculé avec $\mathrm{P}^{X,Y}$ égal à la mesure empirique $\hat{\mathrm{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$.

Question

Le risque empirique $\widehat{\mathscr{R}}_n$ est-il, en général, un « bon » estimateur du risque réel $\mathscr{R}(\hat{h})$?

double utilisation des données !

Intuition : Il est « dangereux » d'estimer le risque à partir de l'erreur commise sur les données ayant servi à construire $\hat{h}$...

Plan du cours

Zoom sur un cas particulier instructif

Considérons le cas de la régression linéaire « ordinaire » :

- $h(x) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)}$,
- perte quadratique : $L(y, \tilde{y}) = (y - \tilde{y})^2$,
- $p + 1 \leq n$ et $\underline{X}^\top \underline{X}$ matrice $(p+1) \times (p+1)$ p.s. inversible.

Minimisation du risque empirique : $\hat{\beta} = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{Y}$.

Remarque : lien entre $\hat{\mathscr{R}}_n$ et le coefficient de détermination R^2 :

$$R^2 = 1 - \frac{\mathrm{SCR}(\hat{\beta})}{\mathrm{SCT}} = 1 - \frac{\sum_{i=1}^n \left(Y_i - \hat{\beta}^\top X_i\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}$$

$$= 1 - \frac{\widehat{\mathscr{R}}_n}{\widehat{\mathrm{var}}_n(Y)} \qquad \text{avec } \widehat{\mathrm{var}}_n(Y) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2.$$

Zoom sur un cas particulier instructif

Considérons le cas de la régression linéaire « ordinaire » :

- $h(x) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)}$,
- perte quadratique : $L(y, \tilde{y}) = (y - \tilde{y})^2$,
- $p + 1 \leq n$ et $\underline{X}^\top \underline{X}$ matrice $(p+1) \times (p+1)$ p.s. inversible.

Minimisation du risque empirique : $\hat{\beta} = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{Y}$.

Remarque : lien entre $\hat{\mathscr{R}}_n$ et le coefficient de détermination R^2 :

$$R^2 = 1 - \frac{\mathrm{SCR}(\hat{\beta})}{\mathrm{SCT}} = 1 - \frac{\sum_{i=1}^n \left(Y_i - \hat{\beta}^\top X_i\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}$$

$$= 1 - \frac{\hat{\mathscr{R}}_n}{\widehat{\mathrm{var}}_n(Y)} \qquad \text{avec } \widehat{\mathrm{var}}_n(Y) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2.$$

Zoom sur un cas particulier instructif

Considérons le cas de la régression linéaire « ordinaire » :

- $h(x) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)}$,
- perte quadratique : $L(y, \tilde{y}) = (y - \tilde{y})^2$,
- $p + 1 \leq n$ et $\underline{X}^\top \underline{X}$ matrice $(p+1) \times (p+1)$ p.s. inversible.

Minimisation du risque empirique : $\hat{\beta} = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{Y}$.

Remarque : lien entre $\hat{\mathscr{R}}_n$ et le coefficient de détermination R^2 :

$$R^2 = 1 - \frac{\mathrm{SCR}(\hat{\beta})}{\mathrm{SCT}} = 1 - \frac{\sum_{i=1}^n \left(Y_i - \hat{\beta}^\top X_i\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}$$

$$= 1 - \frac{\widehat{\mathscr{R}}_n}{\widehat{\mathrm{var}}_n(Y)} \qquad \text{avec } \widehat{\mathrm{var}}_n(Y) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2.$$

Zoom sur un cas particulier instructif (suite)

Considérons l'erreur liée à la généralisation sur les réponses :

$$\widetilde{\mathscr{R}}_n \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{Y}_i - \hat{\beta}^\top X_i\right)^2 \;\middle|\; \underline{X}, \underline{Y}\right),$$

avec, pour tout i, $\widetilde{Y}_i$ et Y_i iid conditionnellement à $\underline{X}$.

Proposition

Supposons la loi inconnue $\mathrm{P}^{X,Y}$ telle que $Y_i = \beta^\top X_i + \varepsilon_i$, avec $\varepsilon_i \sim \mathscr{N}(0, \sigma^2)$, indépendante de X_i. Alors

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

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

Zoom sur un cas particulier instructif (suite)

Considérons l'erreur liée à la généralisation sur les réponses :

$$\widetilde{\mathscr{R}}_n \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{Y}_i - \hat{\beta}^\top X_i\right)^2 \;\middle|\; \underline{X}, \underline{Y}\right),$$

avec, pour tout i, $\widetilde{Y}_i$ et Y_i iid conditionnellement à $\underline{X}$.

Proposition

Supposons la loi inconnue $\mathrm{P}^{X,Y}$ telle que $Y_i = \beta^\top X_i + \varepsilon_i$, avec $\varepsilon_i \sim \mathscr{N}(0, \sigma^2)$, indépendante de X_i. Alors

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

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

Zoom sur un cas particulier instructif (suite)

Considérons l'erreur liée à la généralisation sur les réponses :

$$\widetilde{\mathscr{R}}_n \;=\; \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \left(\widetilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \;\middle|\; \underline{X}, \underline{Y} \right),$$

avec, pour tout i, $\widetilde{Y}_i$ et Y_i iid conditionnellement à $\underline{X}$.

Proposition

Supposons la loi inconnue $\mathrm{P}^{X,Y}$ telle que $Y_i = \beta^\top X_i + \varepsilon_i$, avec $\varepsilon_i \sim \mathscr{N}(0, \sigma^2)$, indépendante de X_i. Alors

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

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

Zoom sur un cas particulier instructif (suite)

Interprétation. En moyenne, le risque empirique sous-estime l'erreur de généralisation :

$$\mathbb{E}\left(\widetilde{\mathscr{R}}_n - \widehat{\mathscr{R}}_n\right) \;=\; 2\,\frac{p+1}{n}\,\sigma^2 \;>\; 0.$$

Autre façon de voir les choses : posons

$$\eta \;=\; \frac{p+1}{n} \;=\; \frac{\text{nombre de coefficients}}{\text{taille de l'échantillon}}.$$

Alors

$$\frac{\mathbb{E}\left(\widetilde{\mathscr{R}}_n\right)}{\mathbb{E}\left(\widehat{\mathscr{R}}_n\right)} \;=\; \frac{1+\eta}{1-\eta} \;\xrightarrow[\eta\to 1]{}\; +\infty.$$

Zoom sur un cas particulier instructif (suite)

Interprétation. En moyenne, le risque empirique sous-estime l'erreur de généralisation :

$$\mathbb{E}\left(\widetilde{\mathscr{R}}_n - \widehat{\mathscr{R}}_n\right) \;=\; 2\,\frac{p+1}{n}\,\sigma^2 \;>\; 0.$$

Autre façon de voir les choses : posons

$$\eta \;=\; \frac{p+1}{n} \;=\; \frac{\text{nombre de coefficients}}{\text{taille de l'échantillon}}.$$

Alors

$$\frac{\mathbb{E}\left(\widetilde{\mathscr{R}}_n\right)}{\mathbb{E}\left(\widehat{\mathscr{R}}_n\right)} \;=\; \frac{1+\eta}{1-\eta} \;\xrightarrow[\eta\to 1]{}\; +\infty.$$

Zoom sur un cas particulier instructif (suite)

Démonstration. Calculons d'abord $\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right)$ avec (rappel)

$$\widetilde{\mathscr{R}}_n = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{Y}_i - \hat{\beta}^\top X_i\right)^2 \;\middle|\; \underline{X}, \underline{Y}\right).$$

On a $\mathbb{E}\left(\widetilde{Y}_i \mid \underline{X}\right) = \mathbb{E}\left(\hat{\beta}^\top X_i \mid \underline{X}\right) = \beta^\top X_i$, donc

$$\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}\left(\widetilde{Y}_i - \hat{\beta}^\top X_i \mid \underline{X}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\mathrm{var}\left(\widetilde{Y}_i \mid \underline{X}\right)}_{=\sigma^2} + \underbrace{\mathrm{var}\left(\hat{\beta}^\top X_i \mid \underline{X}\right)}_{=\circledast}\right).$$

Zoom sur un cas particulier instructif (suite)

Démonstration. Calculons d'abord $\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right)$ avec (rappel)

$$\widetilde{\mathscr{R}}_n \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{Y}_i - \hat{\beta}^\top X_i\right)^2 \;\middle|\; \underline{X}, \underline{Y}\right).$$

On a $\mathbb{E}\left(\widetilde{Y}_i \mid \underline{X}\right) = \mathbb{E}\left(\hat{\beta}^\top X_i \mid \underline{X}\right) = \beta^\top X_i$, donc

$$\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}\left(\widetilde{Y}_i - \hat{\beta}^\top X_i \mid \underline{X}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\mathrm{var}\left(\widetilde{Y}_i \mid \underline{X}\right)}_{=\sigma^2} + \underbrace{\mathrm{var}\left(\hat{\beta}^\top X_i \mid \underline{X}\right)}_{=\circledast}\right).$$

Zoom sur un cas particulier instructif (suite)

Démonstration. Calculons d'abord $\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right)$ avec (rappel)

$$\widetilde{\mathscr{R}}_n \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{Y}_i - \hat{\beta}^\top X_i\right)^2 \;\middle|\; \underline{X}, \underline{Y}\right).$$

On a $\mathbb{E}\left(\widetilde{Y}_i \mid \underline{X}\right) = \mathbb{E}\left(\hat{\beta}^\top X_i \mid \underline{X}\right) = \beta^\top X_i$, donc

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

Zoom sur un cas particulier instructif (suite)

On a vu que $\text{var}\left(\hat{\beta} \mid \underline{X}\right) = \sigma^2 \left(\underline{X}^\top \underline{X}\right)^{-1}$. Donc :

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

En remarquant que $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, il vient :

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

Zoom sur un cas particulier instructif (suite)

On a vu que $\text{var}\left(\hat{\beta} \mid \underline{X}\right) = \sigma^2 \left(\underline{X}^\top \underline{X}\right)^{-1}$. Donc :

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

En remarquant que $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, il vient :

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

Zoom sur un cas particulier instructif (suite)

On a vu que $\text{var}\left(\hat{\beta} \mid \underline{X}\right) = \sigma^2 \left(\underline{X}^\top \underline{X}\right)^{-1}$. Donc :

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

En remarquant que $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, il vient :

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

Zoom sur un cas particulier instructif (suite)

On a vu que $\text{var}\left(\hat{\beta} \mid \underline{X}\right) = \sigma^2 \left(\underline{X}^\top \underline{X}\right)^{-1}$. Donc :

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

En remarquant que $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, il vient :

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

Zoom sur un cas particulier instructif (suite)

Ainsi, on a :

$$\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\mathsf{var}\left(\widetilde{Y}_i \mid \underline{X}\right)}_{=\sigma^2} + \underbrace{\mathsf{var}\left(\hat{\beta}^\top X_i \mid \underline{X}\right)}_{=\circledast}\right)$$
$$= \sigma^2 + \sigma^2\,\frac{p+1}{n} = \sigma^2\left(1+\frac{p+1}{n}\right).$$

D'où le résultat : $\mathbb{E}\left(\widetilde{\mathscr{R}}_n\right) = \sigma^2\left(1+\frac{p+1}{n}\right)$.

Exercice (✏ voir TD) : prouver la deuxième égalité, à savoir

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

□

Zoom sur un cas particulier instructif (suite)

Ainsi, on a :

$$\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\mathsf{var}\left(\widetilde{Y}_i \mid \underline{X}\right)}_{=\sigma^2} + \underbrace{\mathsf{var}\left(\hat{\beta}^\top X_i \mid \underline{X}\right)}_{=\circledast}\right)$$
$$= \sigma^2 + \sigma^2\,\frac{p+1}{n} = \sigma^2\left(1+\frac{p+1}{n}\right).$$

D'où le résultat : $\mathbb{E}\left(\widetilde{\mathscr{R}}_n\right) = \sigma^2\left(1+\frac{p+1}{n}\right)$.

Exercice (✏ voir TD) : prouver la deuxième égalité, à savoir

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

□

Plan du cours

Ensembles d'apprentissage et de test

Conclusion/extrapolation. Le risque empirique est en général

- un estimateur négativement biaisé du risque,
- avec un biais qui augmente lorsque $p \nearrow$.

Solution : partager les données en deux ensembles

- données d'apprentissage : construction de $\hat{h}$,
- données de test : estimation de l'erreur de généralisation.

Exemple :

Ensembles d'apprentissage et de test

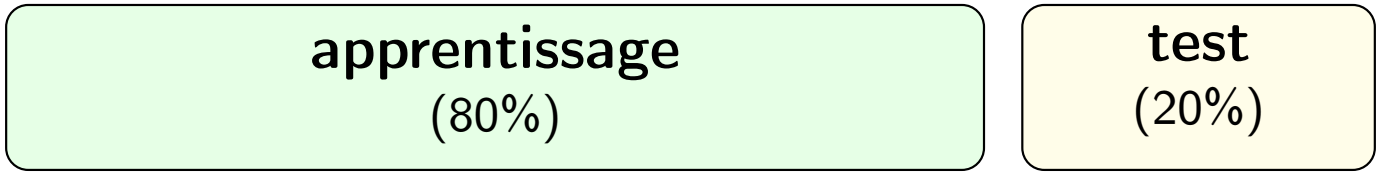
Conclusion/extrapolation. Le risque empirique est en général

- un estimateur négativement biaisé du risque,
- avec un biais qui augmente lorsque $p \nearrow$.

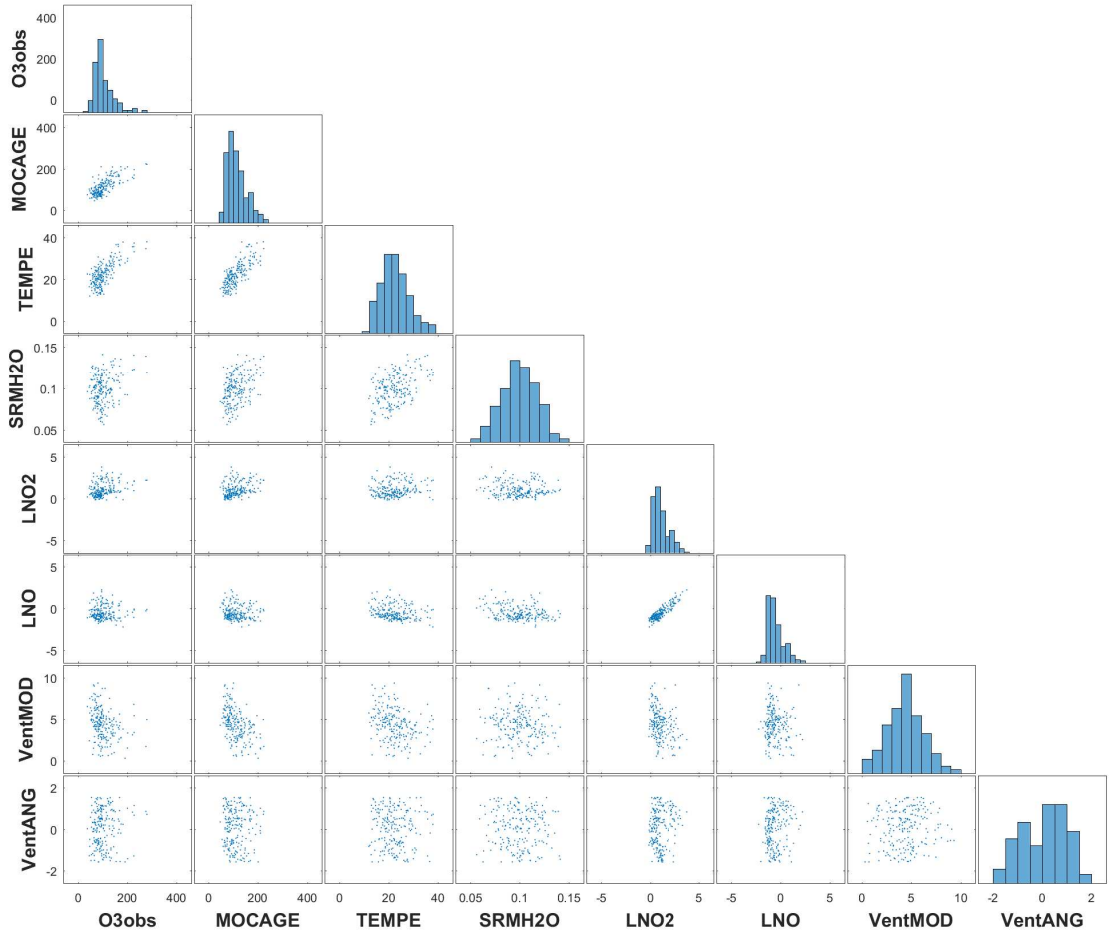
Solution : partager les données en deux ensembles

- données d'apprentissage : construction de $\hat{h}$,
- données de test : estimation de l'erreur de généralisation.

Exemple :



Exemple « Ozone » (rappel)



Objectif : savoir prédire la concentration d'ozone du jour $t+1$ à partir des données disponible au jour t.

Exemple « Ozone » : 70/30

On considère les 7 variables explicatives + 21 interactions $X_j X_k$ $(j \neq k)$.

Résultat de 10 partitions aléatoires, 70% / 30% :

R^2	$\hat{\mathscr{R}}_n$	$\hat{\mathscr{R}}_n^{\text{test}}$
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

Plan du cours

Plan du cours

Problème 1 : choisir une « bonne » famille $\mathscr{H}$

Exemple. Sélection de k variables parmi p. Soit $J \subset \{1, \dots p\}$:

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

➠ Définit une famille $\mathscr{H}_J$ avec $k_J = \text{card}(J) + 1$ paramètres.

Exemple. Développement dans une base, tronqué à l'ordre J :

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

➠ Définit une famille $\mathscr{H}_J$ avec $k_J = J + 1$ paramètres. ➠ complément

Problème : choix de modèle

Comment choisir la famille $\mathscr{H}_J$ (et, en particulier, sa « taille » k_J) ?

Remarque : remplacer $h(x)$ par $\ln \frac{h(x)}{1-h(x)}$ pour la régression logistique.

Problème 1 : choisir une « bonne » famille $\mathscr{H}$

Exemple. Sélection de k variables parmi p. Soit $J \subset \{1, \dots p\}$:

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

➠ Définit une famille $\mathscr{H}_J$ avec $k_J = \text{card}(J) + 1$ paramètres.

Exemple. Développement dans une base, tronqué à l'ordre J :

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

➠ Définit une famille $\mathscr{H}_J$ avec $k_J = J + 1$ paramètres. ➠ complément

Problème : choix de modèle

Comment choisir la famille $\mathscr{H}_J$ (et, en particulier, sa « taille » k_J) ?

Remarque : remplacer $h(x)$ par $\ln \frac{h(x)}{1-h(x)}$ pour la régression logistique.

Problème 1 : choisir une « bonne » famille $\mathscr{H}$

Exemple. Sélection de k variables parmi p. Soit $J \subset \{1, \dots p\}$:

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

➠ Définit une famille $\mathscr{H}_J$ avec $k_J = \text{card}(J) + 1$ paramètres.

Exemple. Développement dans une base, tronqué à l'ordre J :

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

➠ Définit une famille $\mathscr{H}_J$ avec $k_J = J + 1$ paramètres. ➠ complément

Problème : choix de modèle

Comment choisir la famille $\mathscr{H}_J$ (et, en particulier, sa « taille » k_J) ?

Remarque : remplacer $h(x)$ par $\ln \frac{h(x)}{1-h(x)}$ pour la régression logistique.

Problème 2 : choisir un hyper-paramètre

La plupart des méthodes nécessitent un « réglage » ...

- Régression ridge/LASSO : $\hat{\beta} = \text{argmin}\, \hat{\mathscr{R}}^{\text{pen}}_{n,\lambda}$, avec

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

- Arbres de décision, réseaux de neurones : structure
  (par ex. nombre de niveaux de l'arbre / de couches du réseau)

- Méthode des k plus proches voisins : $h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i$,
  avec $\mathcal{V}_{n,k}(x)$ les indices des k plus proches voisins de x.

Problème

Comment choisir la valeur de ces hyper-paramètres ?

Problème 2 : choisir un hyper-paramètre

La plupart des méthodes nécessitent un « réglage » ...

- Régression ridge/LASSO : $\hat{\beta} = \text{argmin}\, \hat{\mathscr{R}}^{\text{pen}}_{n,\lambda}$, avec

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

- Arbres de décision, réseaux de neurones : structure

  (par ex. nombre de niveaux de l'arbre / de couches du réseau)

- Méthode des k plus proches voisins : $h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i$,

  avec $\mathcal{V}_{n,k}(x)$ les indices des k plus proches voisins de x.

Problème

Comment choisir la valeur de ces hyper-paramètres ?

Problème 2 : choisir un hyper-paramètre

La plupart des méthodes nécessitent un « réglage » ...

- Régression ridge/LASSO : $\hat{\beta} = \text{argmin}\, \hat{\mathscr{R}}^{\text{pen}}_{n,\lambda}$, avec

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

- Arbres de décision, réseaux de neurones : structure
  (par ex. nombre de niveaux de l'arbre / de couches du réseau)

- Méthode des k plus proches voisins : $h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i$,
  avec $\mathcal{V}_{n,k}(x)$ les indices des k plus proches voisins de x.

Problème

Comment choisir la valeur de ces hyper-paramètres ?

Attention au sur-apprentissage

Idée

Choisir la famille $\mathscr{H}_J$, ou l'hyperparamètre λ, de façon à minimiser (une estimation de) l'erreur de généralisation.

⚠ à nouveau, le risque empirique $\hat{\mathscr{R}}_n$ estimé sur les données d'apprentissage ne convient pas !

Exemple. Régression polynomiale, $x \in \mathbb{R}$, degré $\leq J$:

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

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

Rappel : en régression linéaire, on a vu que le risque empirique souffre d'un biais proportionnel au nombre de paramètres dans le modèle.

Attention au sur-apprentissage

Idée

Choisir la famille $\mathscr{H}_J$, ou l'hyperparamètre λ, de façon à minimiser (une estimation de) l'erreur de généralisation.

à nouveau, le risque empirique $\hat{\mathscr{R}}_n$ estimé sur les données d'apprentissage ne convient pas !

Exemple. Régression polynomiale, $x \in \mathbb{R}$, degré $\leq J$:

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

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

Rappel : en régression linéaire, on a vu que le risque empirique souffre d'un biais proportionnel au nombre de paramètres dans le modèle.

Attention au sur-apprentissage

Idée

Choisir la famille $\mathscr{H}_J$, ou l'hyperparamètre λ, de façon à minimiser (une estimation de) l'erreur de généralisation.

à nouveau, le risque empirique $\hat{\mathscr{R}}_n$ estimé sur les données d'apprentissage ne convient pas !

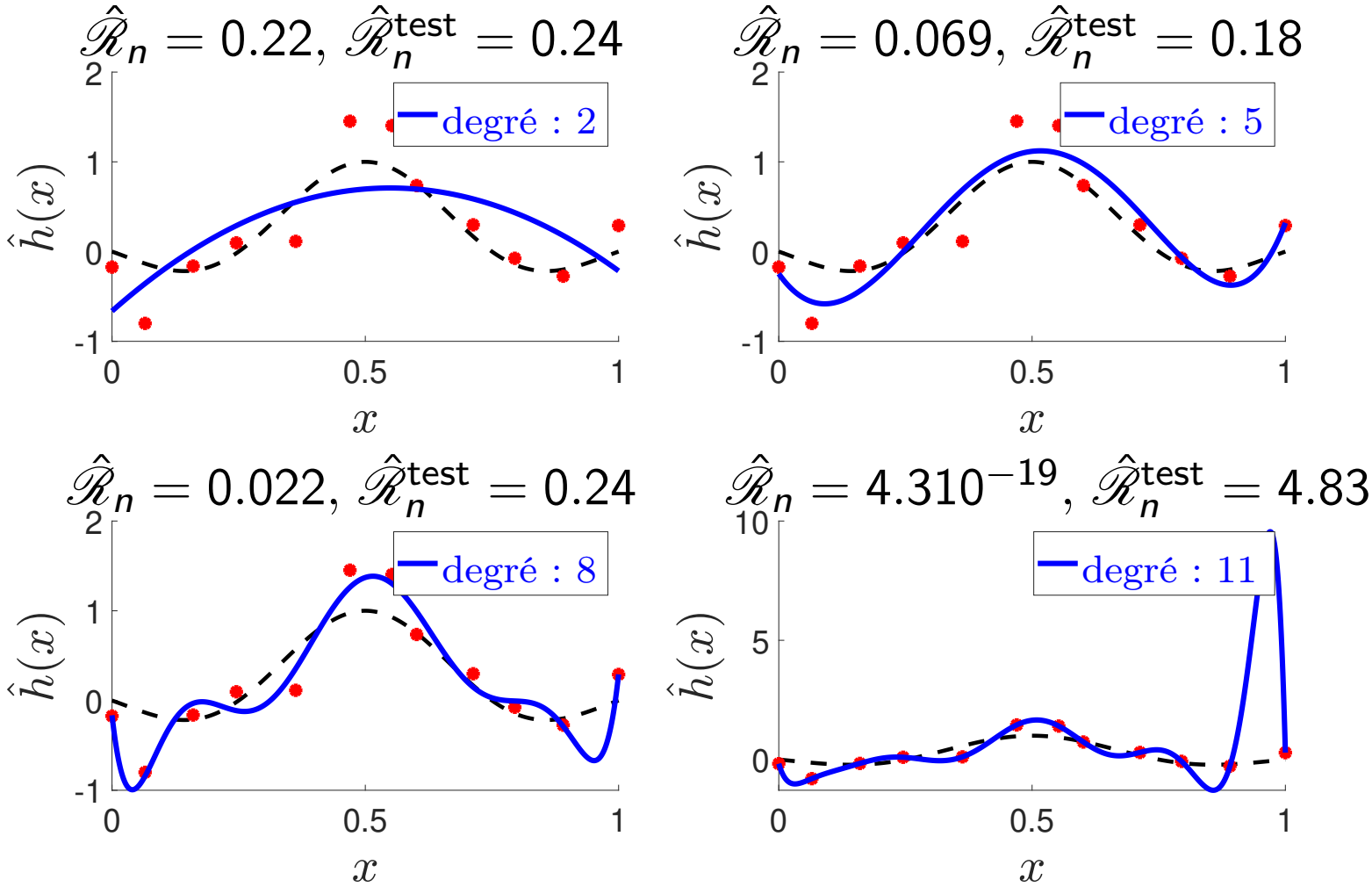
Exemple. Régression polynomiale, $x \in \mathbb{R}$, degré $\leq J$:

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

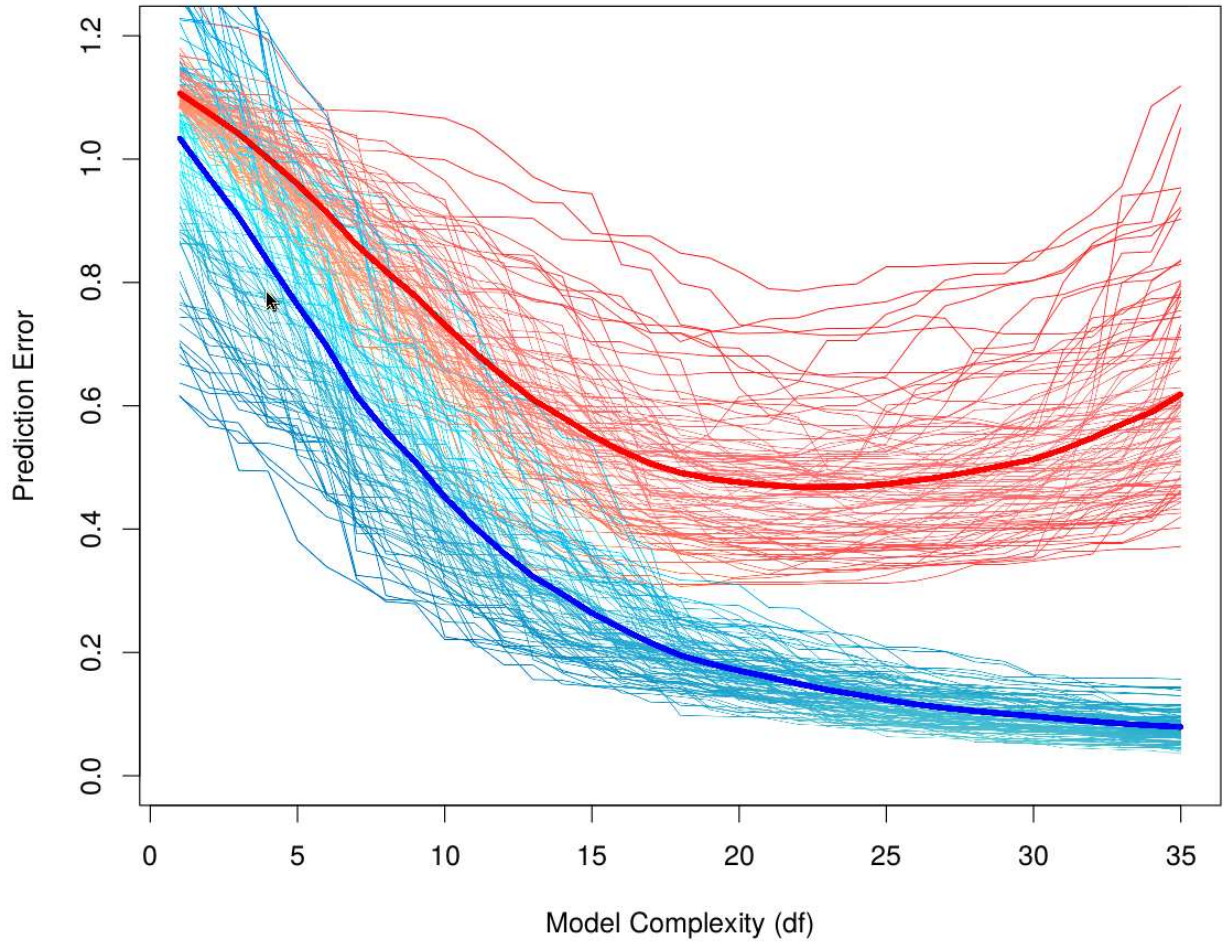
avec $J = 2, 5, 8, 11$.

Rappel : en régression linéaire, on a vu que le risque empirique souffre d'un biais proportionnel au nombre de paramètres dans le modèle.

Exemple : régression polynomiale



Comprendre le sur-apprentissage : simulations



Bleu : risque empirique $\hat{\mathscr{R}}_n$ / Rouge : erreur sur l'ensemble de test

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

Résumons. . .

Problème. On veut estimer l'erreur pour choisir $\mathscr{H}$ ou λ mais. . .

- il ne faut pas le faire sur les données d'apprentissage
  (➠ problème du sur-apprentissage),
- il ne faut pas non plus le faire avec les données de test
  (➠ biais dans l'évaluation finale de l'erreur).

Résumons. . .

Problème. On veut estimer l'erreur pour choisir $\mathscr{H}$ ou λ mais. . .

- il ne faut pas le faire sur les données d'apprentissage
  (➠ problème du sur-apprentissage),
- il ne faut pas non plus le faire avec les données de test
  (➠ biais dans l'évaluation finale de l'erreur).

Plan du cours

Solution : ensemble de validation

Idée : partager les données en trois ensembles

- données d'apprentissage : construction des $\hat{h}$ à $\mathscr{H}$ / λ fixés,
- données de validation : choix de $\mathscr{H}$, λ, etc.
- données de test : estimation de l'erreur de généralisation.

Validation simple (hold-out)

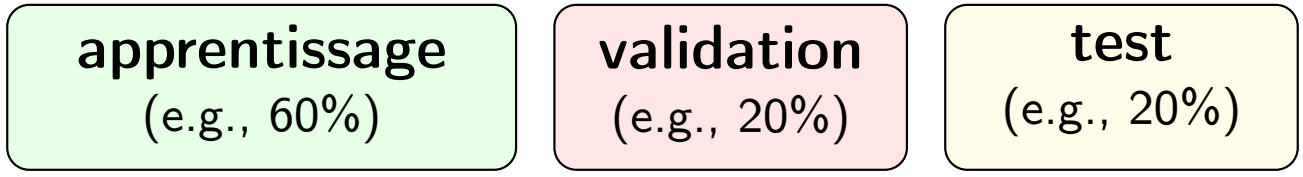
apprentissage (e.g., 60%)	validation (e.g., 20%)	test (e.g., 20%)

Solution : ensemble de validation

Idée : partager les données en trois ensembles

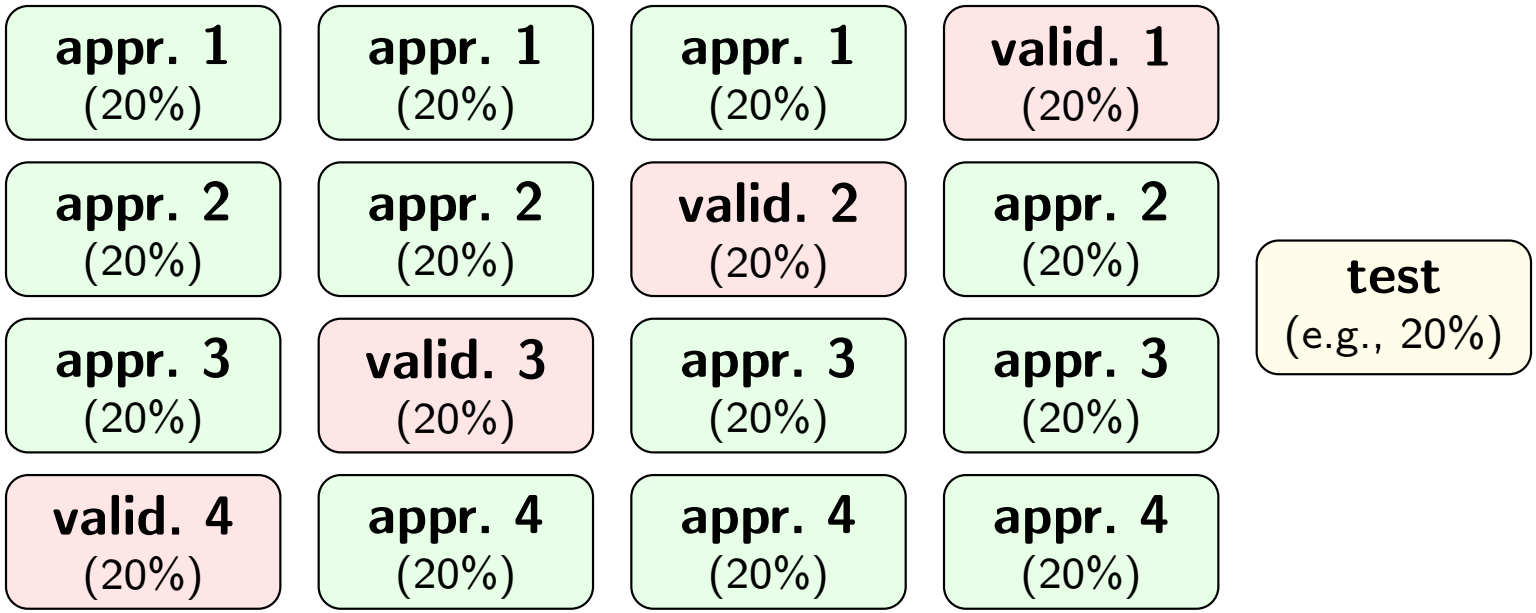
- données d'apprentissage : construction des $\hat{h}$ à $\mathscr{H}$ / λ fixés,
- données de validation : choix de $\mathscr{H}$, λ, etc.
- données de test : estimation de l'erreur de généralisation.

Validation simple (hold-out)



Amélioration : validation croisée

Validation croisée à k blocs (k-*fold*). Ici avec $k = 4$:



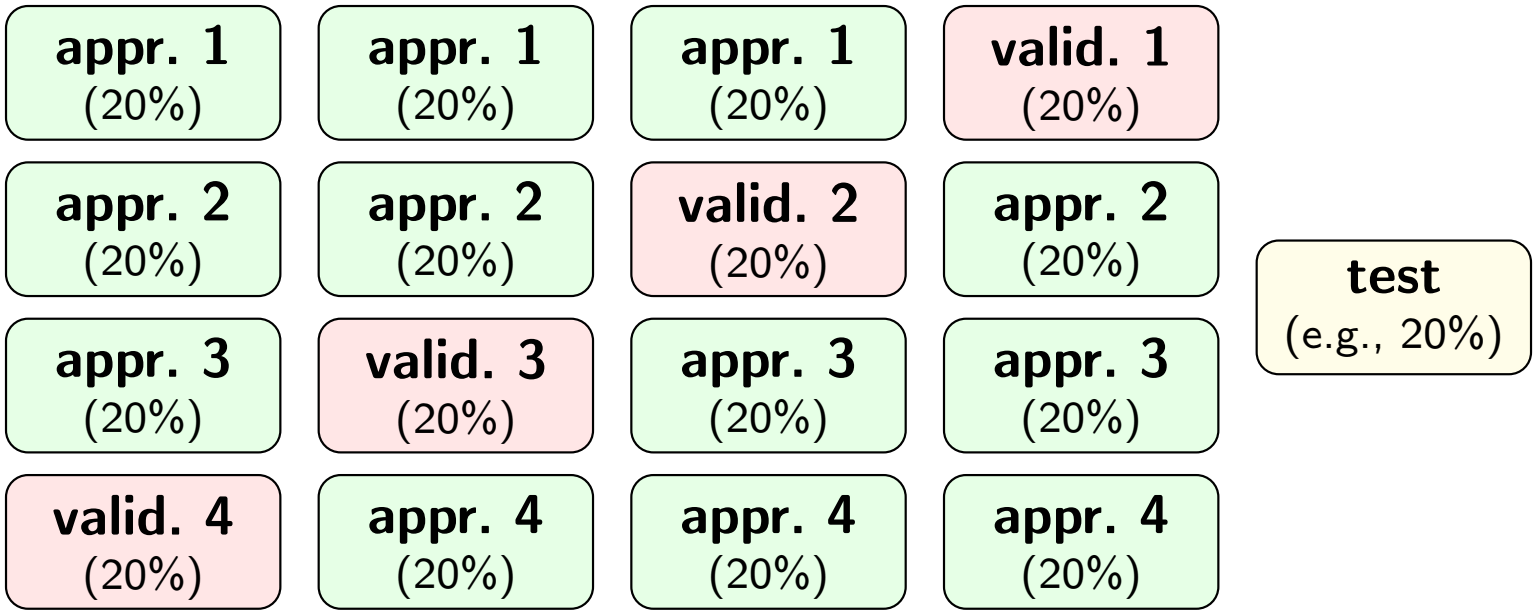
➠ on moyenne les erreurs sur les k ensembles de validation.

Cas particulier : validation croisée « leave one out »

- $k = n$ blocs (de taille $n/k = 1$).

Amélioration : validation croisée

Validation croisée à k blocs (k-*fold*). Ici avec $k = 4$:



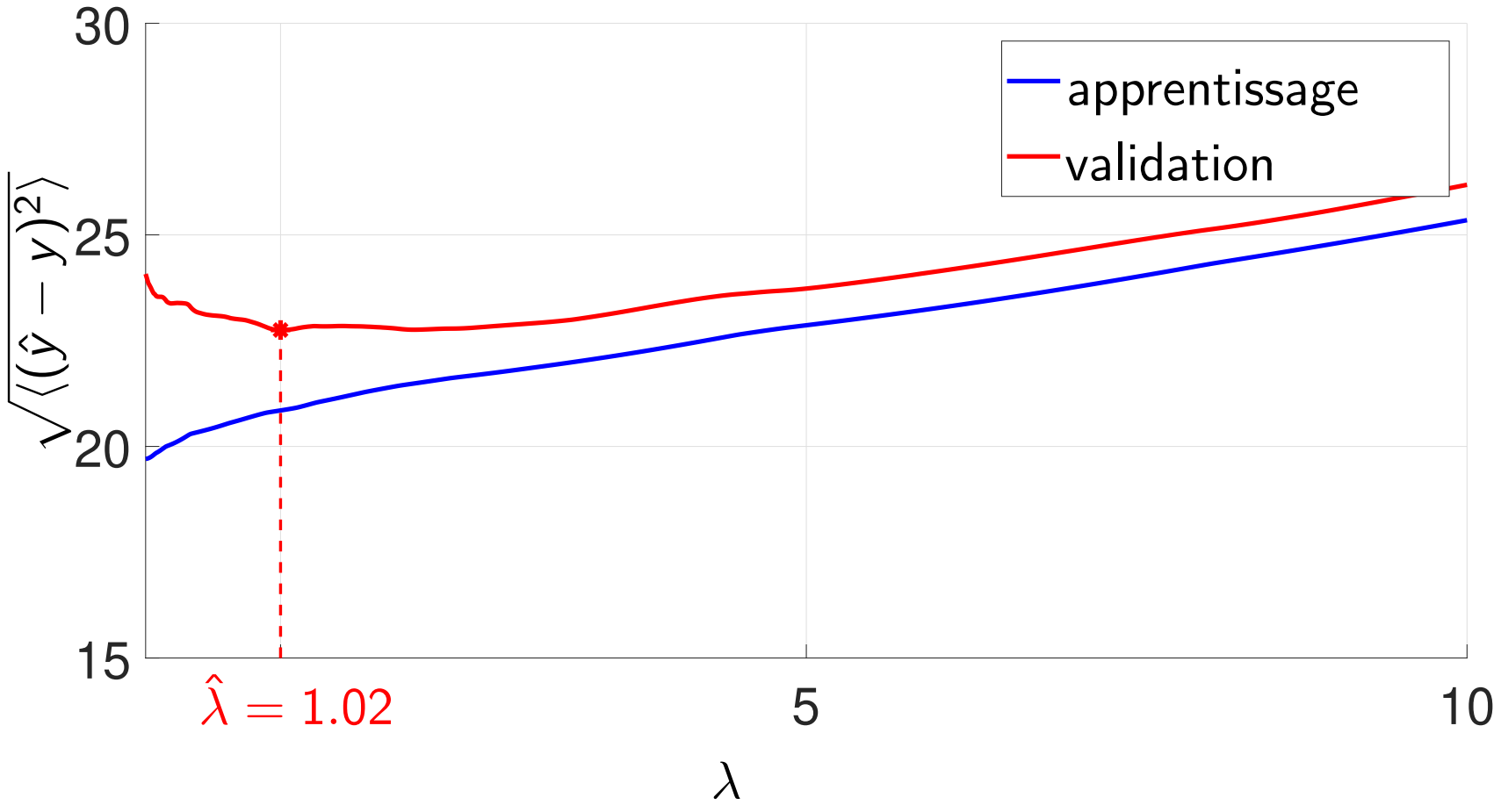
➠ on moyenne les erreurs sur les k ensembles de validation.

Cas particulier : validation croisée « leave one out »

- $k = n$ blocs (de taille $n/k = 1$).

Exemple « Ozone » : estimation du λ

- Prédicteur obtenu par une régression LASSO réalisée sur toutes les variables + leurs interactions
- $\hat{\lambda}$ obtenu par VC (LOO)



Exemple « Ozone » : interactions

- Ajout des variables de type $X^{(j)}X^{(j')}$ et $X^{(j)}X^{(j')}X^{(j'')}$.
- Régression LASSO (pénalisation L^1).
- Hyper-paramètre λ estimé par VC (10-fold).

modèle	$X^{(j)}$	$X^{(j)}\,X^{(j')}$	$X^{(j)}\,X^{(j')}\,X^{(j'')}$
nombre total de variables	7	35	119
nombre de variables sélectionnées ($\beta_j \neq 0$)	4	9	8
$\sqrt{MSE}$ VC (10-fold)	49.1	41.5	33.0
variables sélectionnées	MOCAGE TEMPE NO VentANG	MOCAGE TEMPE NO2 MOCAGE · TEMPE TEMPE^2 TEMPE · MH2O TEMPE · NO2 NO2 · VentANG VentANG · VentANG	MOCAGE TEMPE NO2 MOCAGE · TEMPE TEMPE^2 TEMPE · RMH2O TEMPE^2 · MOCAGE VentANG^2 · TEMPE

Plan du cours

Autre approche pour le choix de modèle : le critère AIC

Hypothèse : modèles statistiques paramétriques $\mathscr{M}_j$ pour $\mathrm{P}^{Y|X}$.

On note $\hat{\theta}_j^{\mathrm{EMV}}$ l'EMV de θ dans le modèle $\mathscr{M}_j$.

Alors le critère AIC peut être aussi utilisé pour le choix de modèle :

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

avec k_j le nombre de paramètres dans le modèle $\mathscr{M}_j$.

➠ voir TD pour une justification partielle (régression linéaire MCO)

Autre approche pour le choix de modèle : le critère AIC

Hypothèse : modèles statistiques paramétriques $\mathscr{M}_j$ pour $\mathrm{P}^{Y|X}$.

On note $\hat{\theta}_j^{\mathrm{EMV}}$ l'EMV de θ dans le modèle $\mathscr{M}_j$.

Alors le critère AIC peut être aussi utilisé pour le choix de modèle :

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

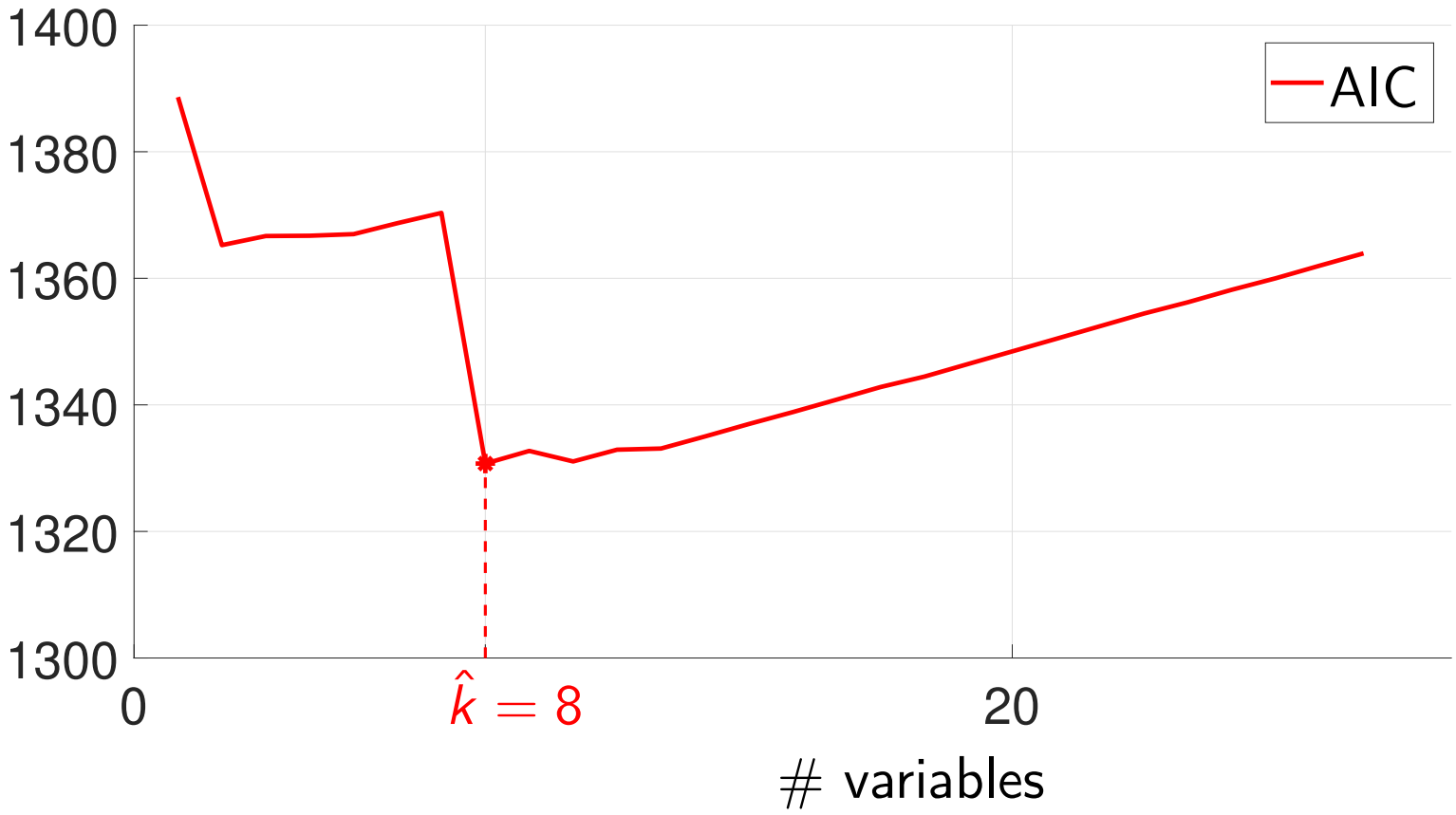
avec k_j le nombre de paramètres dans le modèle $\mathscr{M}_j$.

➠ voir TD pour une justification partielle (régression linéaire MCO)

Exemple « Ozone » : AIC

- Prédicteur obtenu par une régression OLS réalisée sur un nombre croissant de variables

  (d'abord les termes linéaires, puis les interactions)



Conclusion du cours et transition vers la séquence suivante

Nous avons vu et développerons en TD :

- les méthodes ridge et LASSO pour la régression linéaire pénalisée ;
- la problématique de l'estimation de l'erreur de généralisation (risque) ;
- la méthode de validation croisée pour le réglage des hyper-paramètres et le choix de modèle.

Nous aborderons dans la dernière séquence :

- les problématiques de l'apprentissage non supervisé ;
- l'analyse en composantes principales (ACP) pour la réduction de dimension ;
- l'algorithme des K-means pour le partitionnement (clustering).

Conclusion du cours et transition vers la séquence suivante

Nous avons vu et développerons en TD :

- les méthodes ridge et LASSO pour la régression linéaire pénalisée ;
- la problématique de l'estimation de l'erreur de généralisation (risque) ;
- la méthode de validation croisée pour le réglage des hyper-paramètres et le choix de modèle.

Nous aborderons dans la dernière séquence :

- les problématiques de l'apprentissage non supervisé ;
- l'analyse en composantes principales (ACP) pour la réduction de dimension ;
- l'algorithme des K-means pour le partitionnement (clustering).

Plan du cours

Plan du cours

Exercice 1 (Régression pénalisée)

corrigé

Soient $X_1, \ldots, X_n$ les exemples, à valeurs dans $\mathbb{R}^p$, et $Y_1, \ldots, Y_n$ les étiquettes, à valeurs dans $\mathbb{R}$. La relation liant Y_i à X_i est donnée par :

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

où β est le vecteur de paramètres à estimer, et ε_i une variable aléatoire de loi $\mathcal{N}(0, \sigma^2)$, indépendante de X_i.
On se propose d'estimer β en minimisant un critère de la forme

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

où $\mathcal{P}$ est un terme de pénalisation, et $\lambda \geq 0$ un hyper-paramètre.

Exercice 1 (Régression pénalisée)

corrigé

slide 12

On note $X = [X_1 \dots X_n]^\top$, la matrice $n \times p$ contenant les observations. **On se place dans le cas où $X^\top X = I_p$.**

Question

1. Donner l'expression de l'estimateur quand $\lambda = 0$. On notera cet estimateur $\hat{\beta}$.
2. On considère une pénalisation de la forme $\mathcal{P}(\beta) = \|\beta\|_2^2$. Donner l'expression de cet estimateur que l'on notera $\hat{\beta}^R$,et en déduire qu'il existe une constante $c_{1,\lambda}$ (que l'on explicitera) telle que $\hat{\beta}_j^R = c_{1,\lambda}\hat{\beta}_j$, $j = 1, \dots, p$.

Exercice 1 (Régression pénalisée)

corrigé

slide 12

Question

3. On considère une pénalisation de la forme $\mathcal{P}(\beta) = \|\beta\|_1$.
   En préambule, montrer que le minimum sur $\mathbb{R}$ de la fonction

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

   est atteint pour$\alpha^* = \text{sign}(x)\ \max(0, |x| - \lambda)$.

4. En déduire la solution du problème d'optimisation (1) pour $\mathcal{P}(\beta) = \|\beta\|_1$ que l'on exprimera en fonction de $\hat{\beta}$. On notera $\hat{\beta}^L$ cet estimateur.

Plan du cours

Corrigé de l'exercice 1

retour au sujet

1. On reconnait le critère des moindres carrés et l'on a :

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

2. Il s'agit de la régression ridge.

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

Il vient donc que $\hat{\beta}_j^R = (1 + 2\lambda)^{-1} \hat{\beta}_j$.

3. La fonction f n'est pas dérivable, mais elle est dérivable en tout point $\alpha \neq 0$ *et continue en* $\alpha = 0$. On peut donc déterminer son minimum en faisant une analyse de ses variations au moyen du signe de la dérivée, comme si elle était dérivable partout.
La dérivée en tout $\alpha \neq 0$ s'écrit

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

d'où

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

3. Supposons par exemple $x > 0$. Alors le deuxième cas dans le membre de droite de (2) est impossible, et il reste :

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

De même, toujours en supposant $x > 0$,

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

Ainsi f est strictement décroissante à gauche de $\max(0, x-\lambda)$, strictement croissante à droite, ce qui conclut le cas $x > 0$.
Le cas $x < 0$ s'en déduit comme précédemment.

retour au sujet

4. On va ici manipuler le problème d'optimisation initial de sorte à se ramener au problème d'optimisation de la question précédente :

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

Le produit croisé s'annule puisque le résidu $(Y - X\hat{\beta})$ est par construction, orthogonal à toute combinaison linéaire de colonnes de X et donc $(Y - X\hat{\beta})^\top(X\hat{\beta} - X\beta) = 0$.

retour au sujet

4. Le premier terme étant indépendant de β, il vient donc :

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

Le problème est séparable et, de la question précédente, il vient :

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

Plan du cours

Plan du cours

Non-linéarité dans les modèles linéaires...

Si le risque empirique $\hat{\mathscr{R}}(\hat{h})$ est élevé, plusieurs causes possibles :

- bruit : difficulté intrinsèque à prédire Y
  - ➠ erreur statistique irréductible.
- non-linéarité du prédicteur optimal par rapport aux $X^{(j)}$
  - ➠ erreur d'approximation, réductible.

Solution possible : $x^{(1)}, \ldots, x^{(p)} \mapsto \tilde{x}^{(1)}, \ldots, \tilde{x}^{(q)}$

- avec $\tilde{x}^{(j)}$ fonction de $x^{(1)}, \ldots, x^{(p)}$.
- Le modèle reste linéaire par rapport à β.

Non-linéarité dans les modèles linéaires...

Si le risque empirique $\hat{\mathscr{R}}(\hat{h})$ est élevé, plusieurs causes possibles :

- bruit : difficulté intrinsèque à prédire Y
  - ➠ erreur statistique irréductible.
- non-linéarité du prédicteur optimal par rapport aux $X^{(j)}$
  - ➠ erreur d'approximation, réductible.

Solution possible : $x^{(1)}, \ldots, x^{(p)} \mapsto \tilde{x}^{(1)}, \ldots, \tilde{x}^{(q)}$

- avec $\tilde{x}^{(j)}$ fonction de $x^{(1)}, \ldots, x^{(p)}$.
- Le modèle reste linéaire par rapport à β.

Non-linéarité dans les modèles linéaires...

Si le risque empirique $\hat{\mathscr{R}}(\hat{h})$ est élevé, plusieurs causes possibles :

- bruit : difficulté intrinsèque à prédire Y
  - ➠ erreur statistique irréductible.
- non-linéarité du prédicteur optimal par rapport aux $X^{(j)}$
  - ➠ erreur d'approximation, réductible.

Solution possible : $x^{(1)}, \ldots, x^{(p)} \mapsto \tilde{x}^{(1)}, \ldots, \tilde{x}^{(q)}$

- avec $\tilde{x}^{(j)}$ fonction de $x^{(1)}, \ldots, x^{(p)}$.
- Le modèle reste linéaire par rapport à β.

Exemples

Quelques exemples :

- transformations scalaires : $\ln(x^{(j)})$, $\sqrt{x^{(j)}}$, $(x^{(j)})^k \ldots$
- interactions (ici d'ordre deux) : $x^{(j)}x^{(k)}$, $j \neq k$,
- interactions d'ordre supérieur,
- développement (tronqué) dans une base...

si $q \gg p$, risque de sur-apprentissage.

Remarques : *feature engineering*

- Proposer de nouvelles variables pertinentes
  - ➠ expertise métier (ou choix de modèle...?)
- On peut utiliser ce principe pour *réduire* la dimension
  - ➠ extraction de caractéristiques (*features extraction*).

Exemples

Quelques exemples :

- transformations scalaires : $\ln(x^{(j)})$, $\sqrt{x^{(j)}}$, $(x^{(j)})^k \ldots$
- interactions (ici d'ordre deux) : $x^{(j)}x^{(k)}$, $j \neq k$,
- interactions d'ordre supérieur,
- développement (tronqué) dans une base...

 si $q \gg p$, risque de sur-apprentissage.

Remarques : *feature engineering*

- Proposer de nouvelles variables pertinentes
  - ➠ expertise métier (ou choix de modèle...?)
- On peut utiliser ce principe pour *réduire* la dimension
  - ➠ extraction de caractéristiques (*features extraction*).

Développement dans une base

Principe

Soit $\{\psi_m\}_{m>0}$ une base de fonctions de $L^2(\mathcal{X})^\dagger$.

On considère $\tilde{X}^{(m)} = \psi_m(X)$, $m = 1, \ldots, M$

➠ décomposition tronquée dans la base $\{\psi_m\}$.

Exemples de bases (de préférence orthogonales) :

- bases de polynômes,
- bases d'ondelettes,
- base de Fourier...

† ou autre espace de fonctions sur $\mathcal{X}$ supposé contenir le h^* optimal.

➠ retour au slide 28

Développement dans une base

Principe

Soit $\{\psi_m\}_{m>0}$ une base de fonctions de $L^2(\mathcal{X})^\dagger$.

On considère $\tilde{X}^{(m)} = \psi_m(X)$, $m = 1, \ldots, M$

➠ décomposition tronquée dans la base $\{\psi_m\}$.

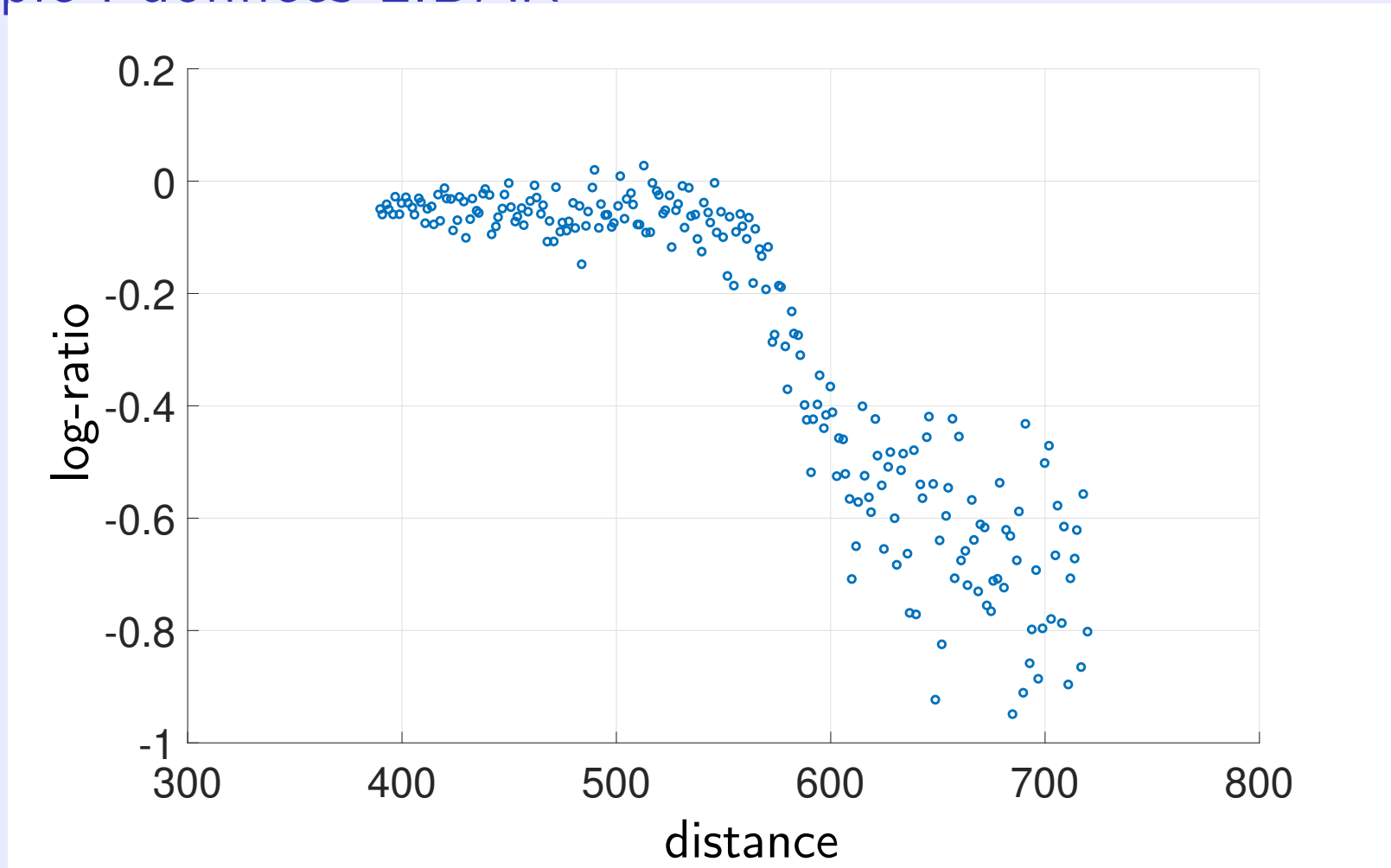
Exemples de bases (de préférence orthogonales) :

- bases de polynômes,
- bases d'ondelettes,
- base de Fourier...

† ou autre espace de fonctions sur $\mathcal{X}$ supposé contenir le h^* optimal.

➠ retour au slide 28

Exemple : données LIDAR



abscisse : distance entre le point de réflexion et la source laser

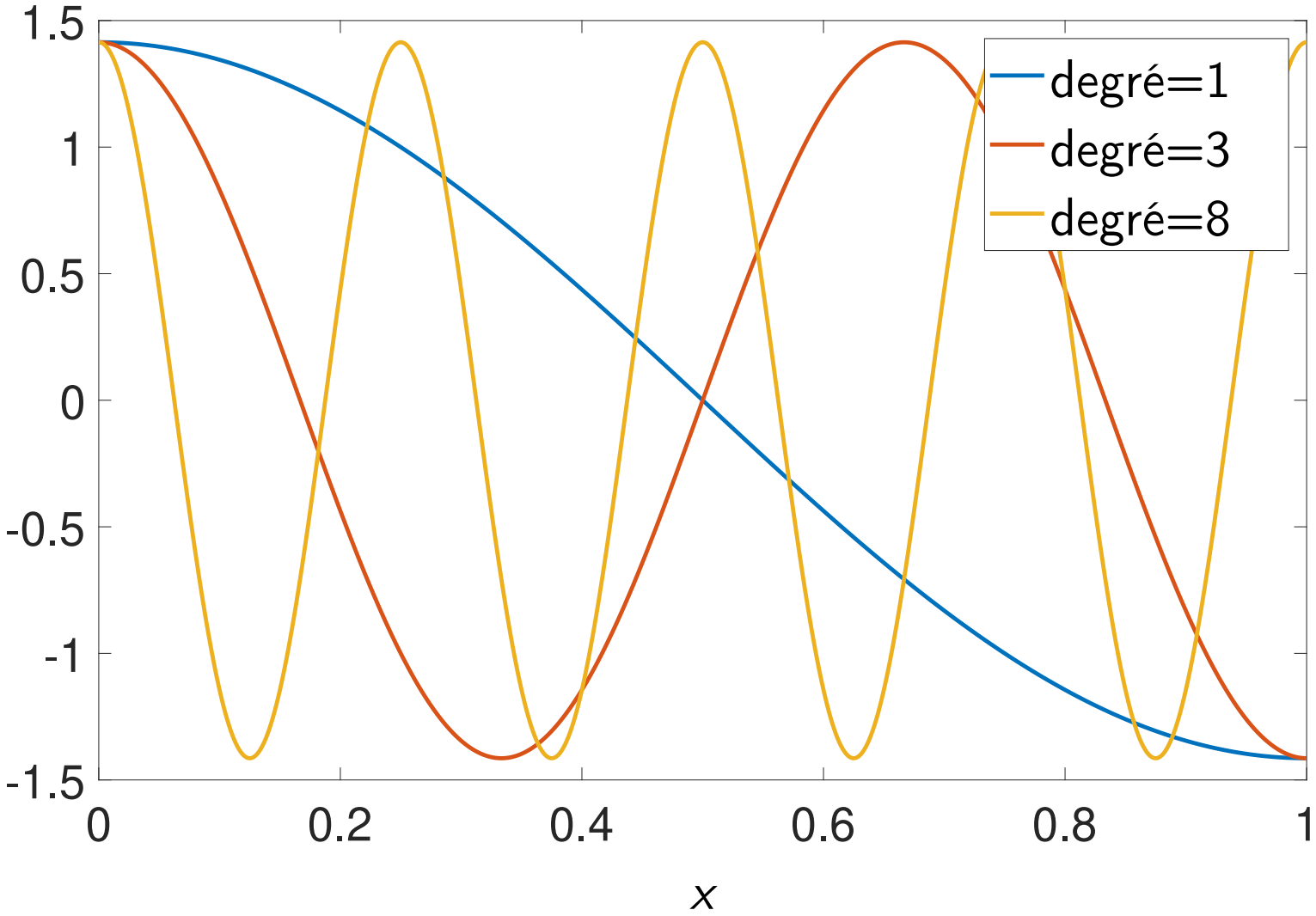
ordonnée : log-ratio de lumière reçue pour 2 sources de fréquences différentes

Données issues de `http://matt-wand.utsacademics.info/webspr/lidar.html`

LIDAR : LIght Detection And Ranging

retour au slide 28

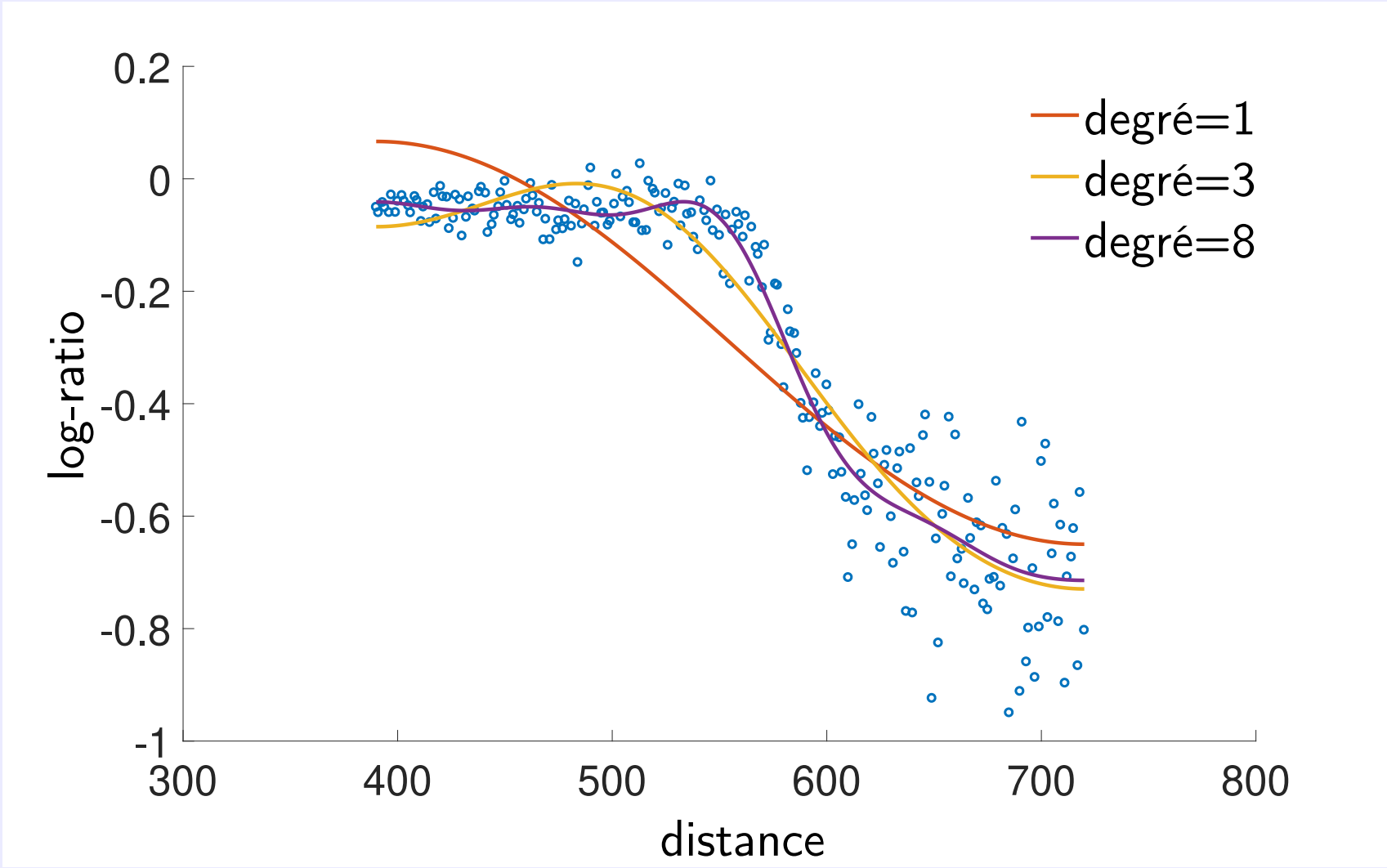
Base de cosinus orthogonaux (base de $L^2([0,1])$



retour au slide 28

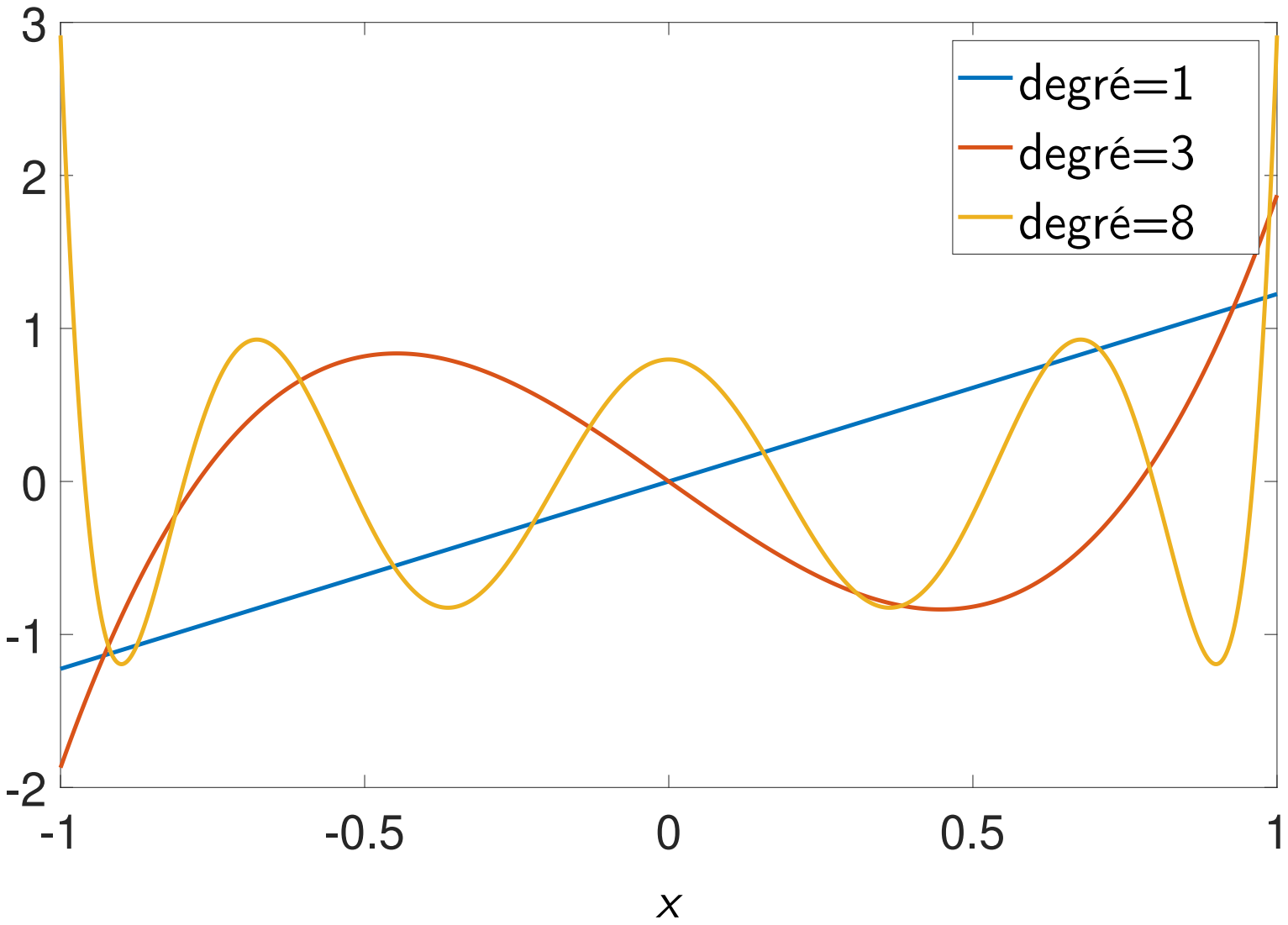
Exemple : données LIDAR (suite)

Fonction de perte quadratique, base de cosinus orthogonaux



retour au slide 28

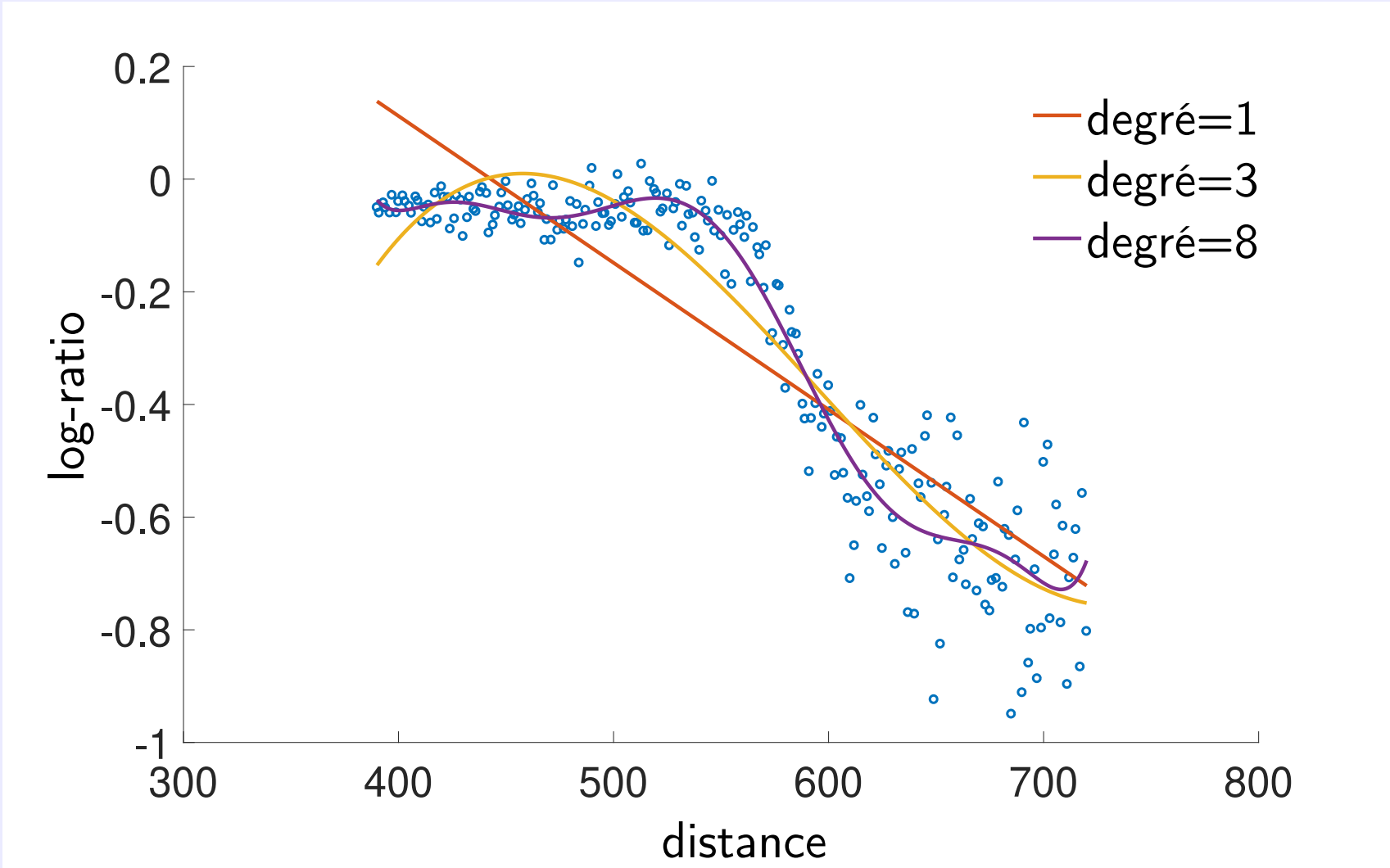
Polynômes de Legendre (base de $L^2([-1,1])$)



retour au slide 28

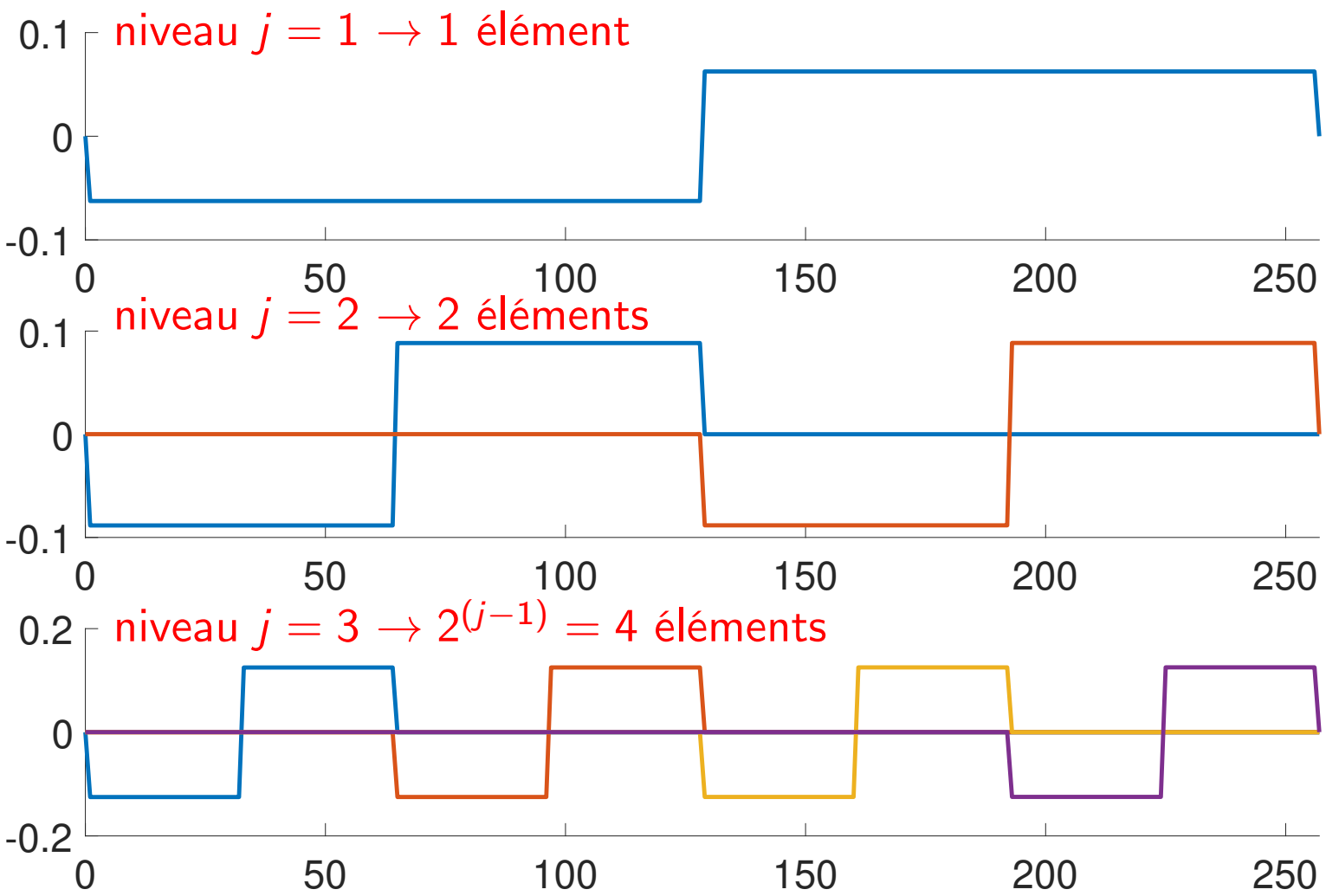
Exemple : données LIDAR (suite)

Fonction de perte quadratique + polynômes de Legendre



retour au slide 28

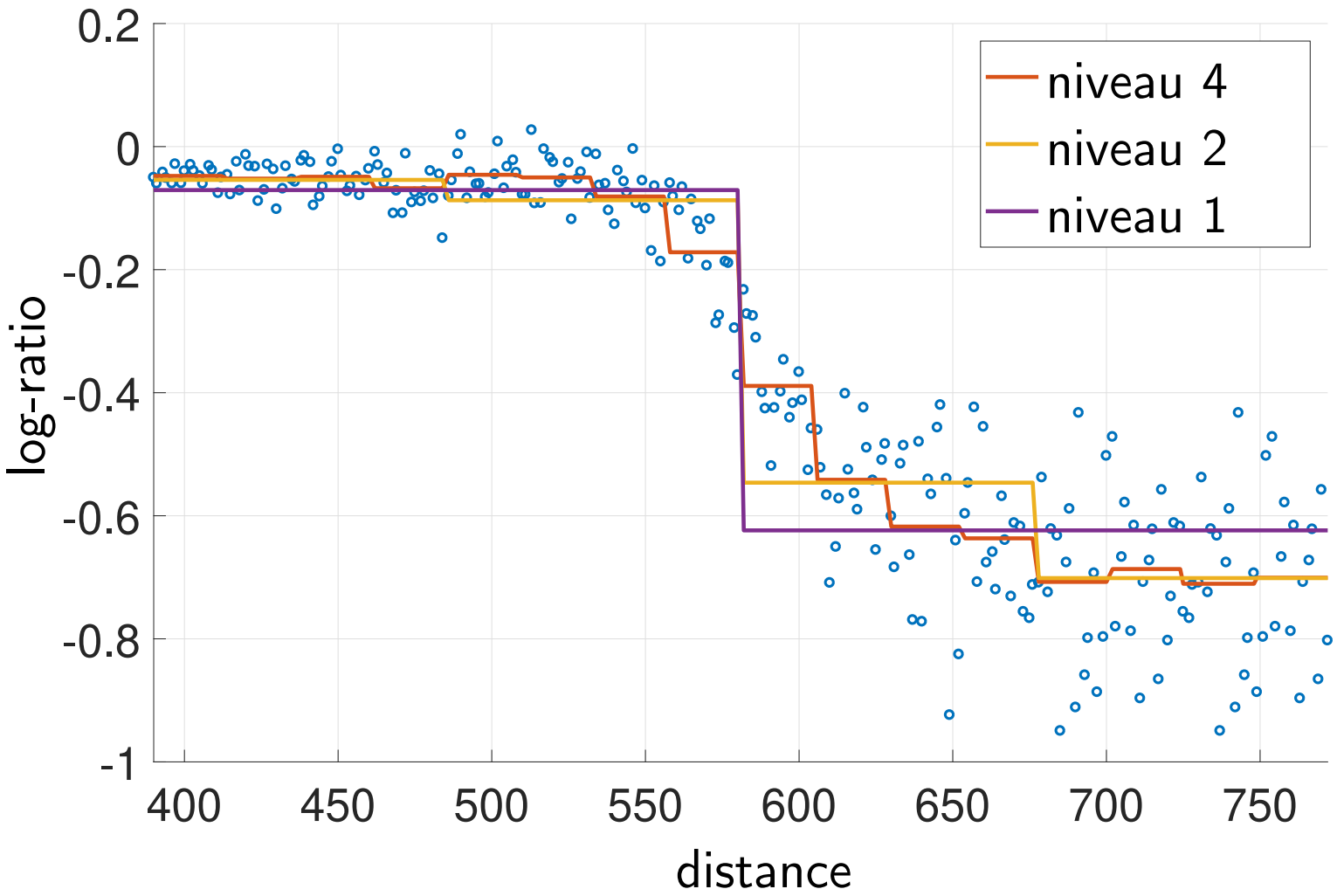
Base d'ondelettes de Haar



retour au slide 28

Exemple : données LIDAR (suite)

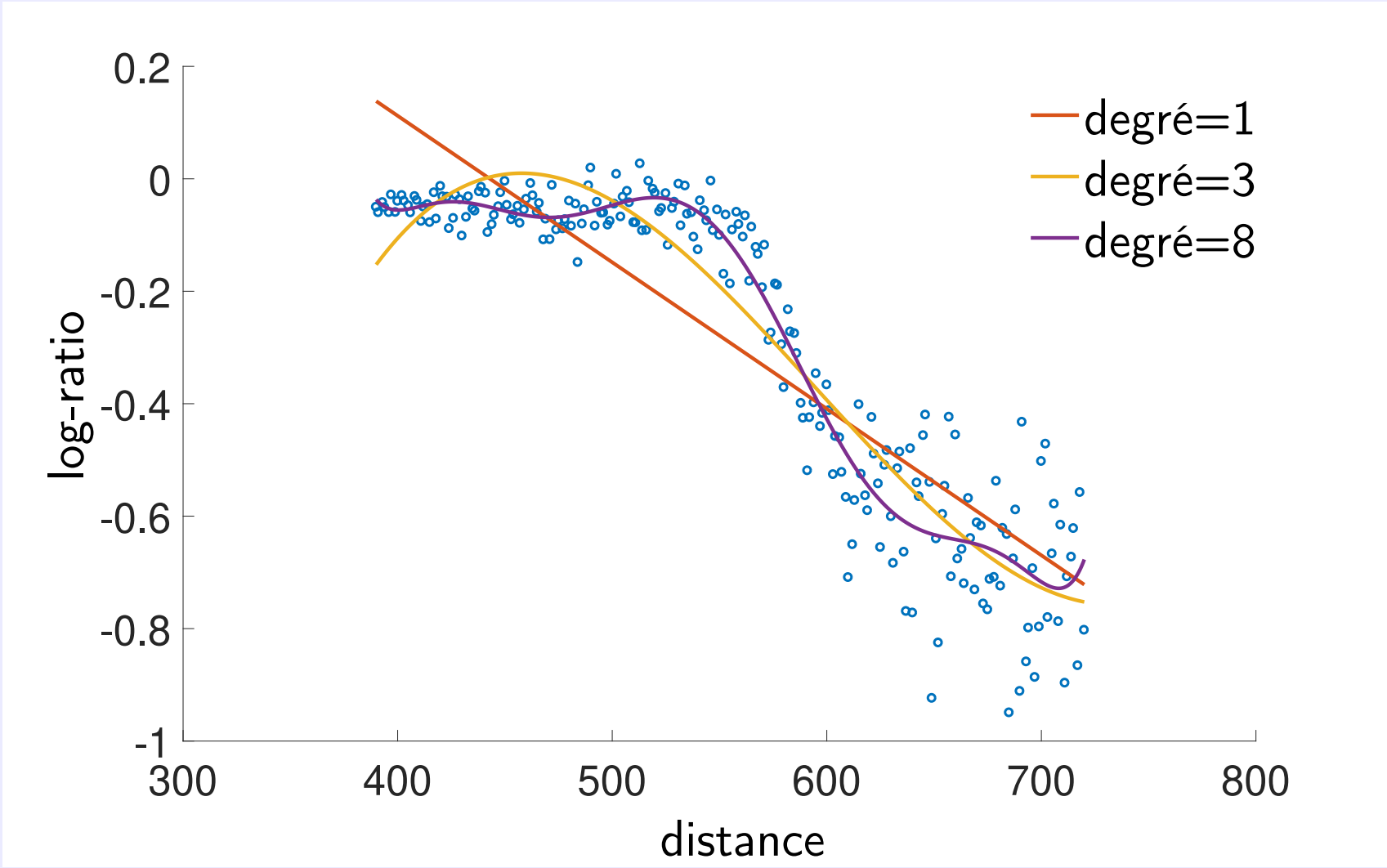
Fonction de perte quadratique + ondelettes de Haar



retour au slide 28

Exemple : données LIDAR (suite)

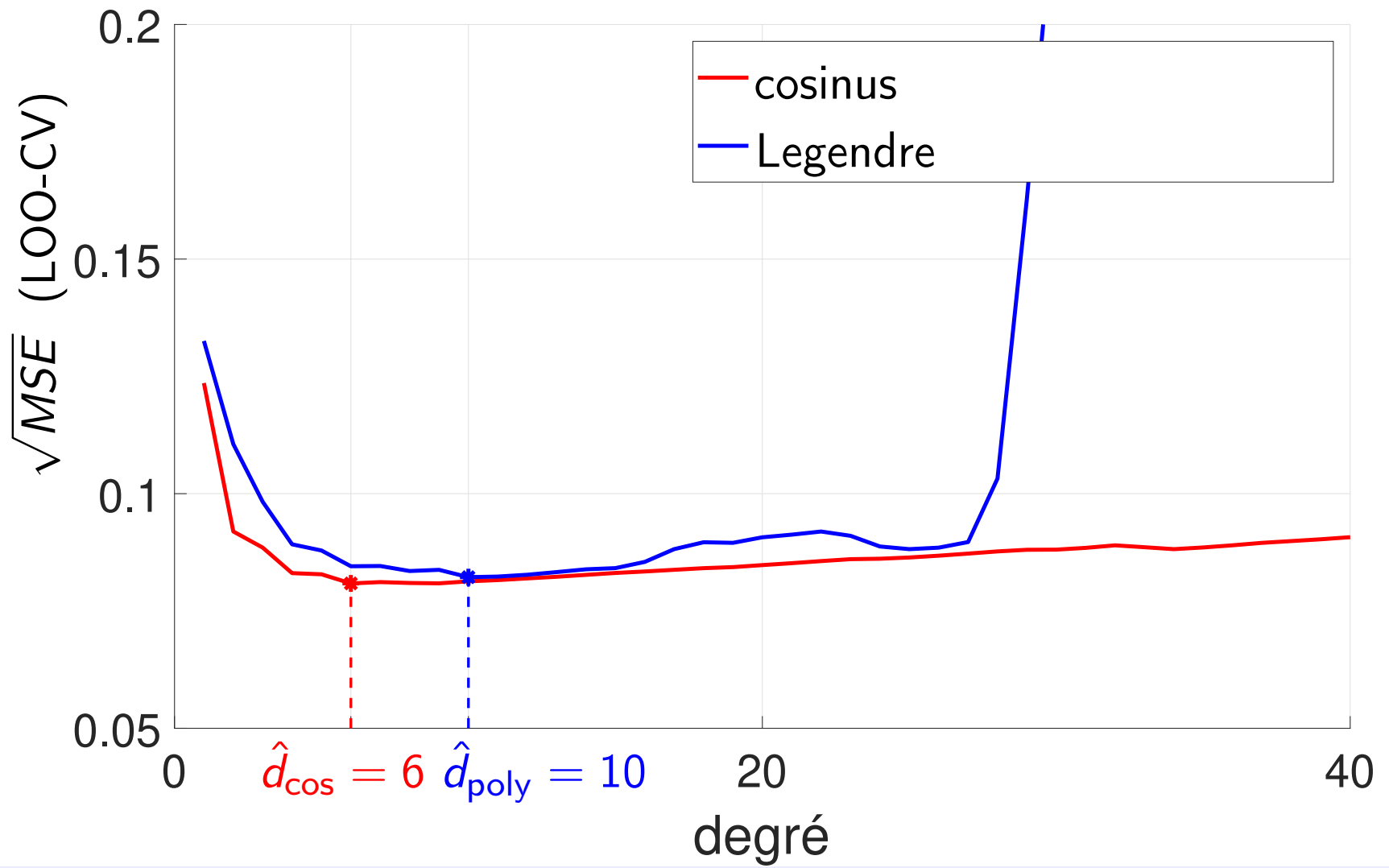
Fonction de perte quadratique + polynômes de Legendre



retour au slide 28

Exemple : données LIDAR (suite)

Choix de modèle



retour au slide 28