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

Lecture outline

Lecture outline

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.

Example: $p > n$

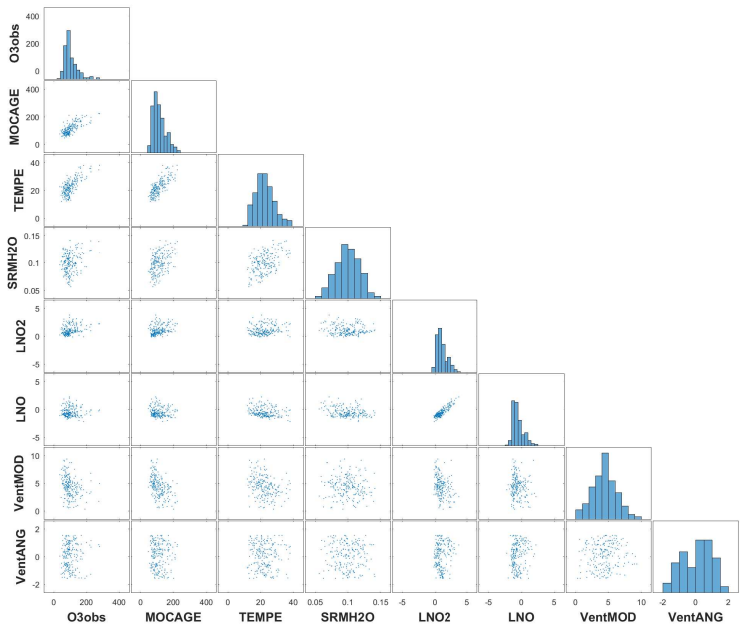
	A	B	C	D	E	F	G	H	I	J	K	
1	ID	air_time1	disp_index1	gmrt_in_air1	gmrt_on_paper1	max_x_extension1	max_y_extension1	mean_acc_in_air1	mean_acc_on_paper1	mean_gmrt1	mean_jerk_in_air1	mean on_pa
2	id_1	5160	1.25E-05	1.21E+02	8.69E+01	957	6601	3.62E-01	2.17E-01	1.04E+02	5.18E-02	2.
3	id_2	51980	1.60E-05	1.15E+02	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.

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, ant the associated dataset: UCI Machine Learning Repository. `https://doi.org/10.24432/C55D0K`.

Example: strong correlation between explanatory variables



"Ozone" example $\longrightarrow$ correlation between variables NO and NO2

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

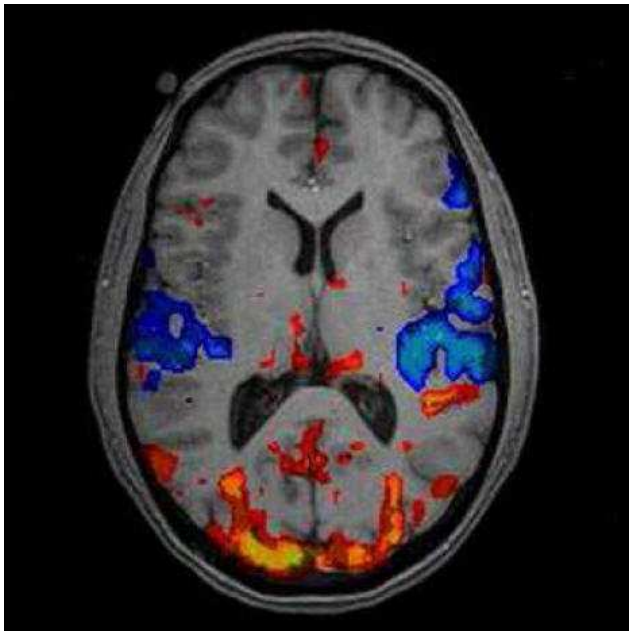
Vector $\hat{\beta}$ obtained by OLS regression:

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:

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

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

Ridge regression

Penalty

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

(usually, β_0 is not penalized)

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

It can be proved that (✏ see PC):

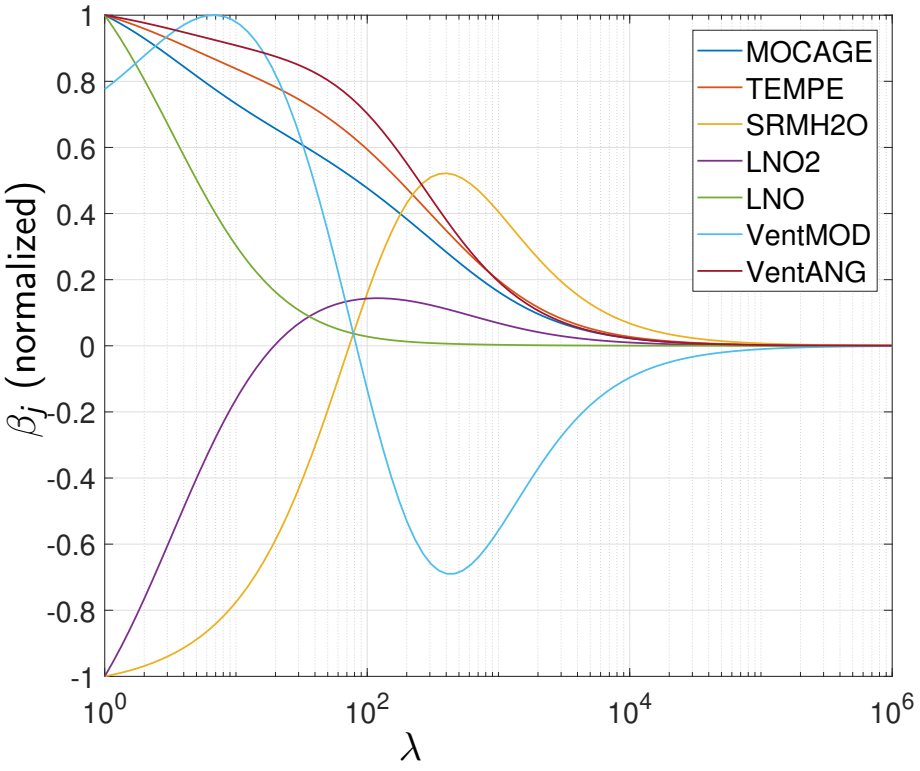
$$\hat{\beta}^{\text{RIDGE}} = \left(\underline{X}^\top \underline{X} + \lambda I_{p+1}\right)^{-1} \underline{X}^\top \underline{Y}.$$

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

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

(✏ see PC too).

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



Lecture outline

LASSO regression

Penalty

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

(usually, β_0 is not penalized)

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

Minimization of the criterion

- no explicit solution for $\hat{\beta}^{\text{LASSO}}$ (except in some cases, exercice 1)
  - dedicated algorithms

LASSO regression: reformulation

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

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

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

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

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

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

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

LASSO regression: intuitive interpretation

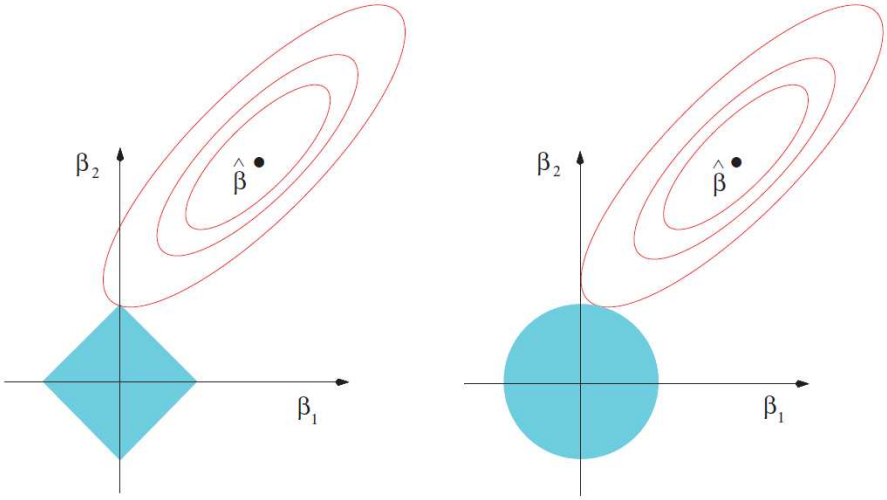
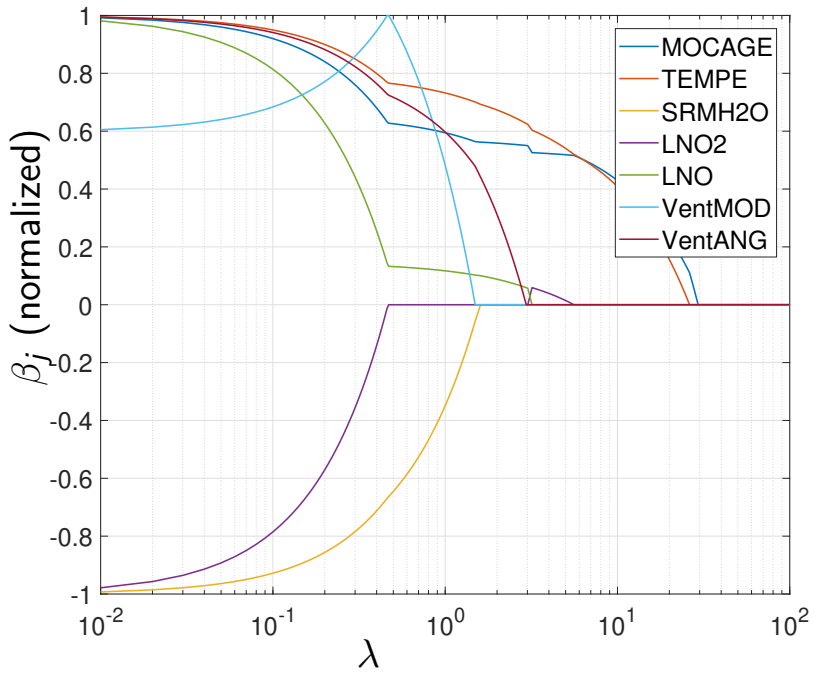


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are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

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 3

"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 ?

Lecture outline

Lecture outline

Problem

Back to the general setting (regression/classification).

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

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

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

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

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 $\mathrm{P}^{X,Y}$) ?

Refresher: empirical risk

We call empirical risk the risk

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

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

Question

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

⚠ 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

Zoom in on an illuminating special case

Consider the case of "ordinary" linear regression:

- $h(x) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)}$,
- 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 : $\hat{\beta} = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{Y}$.

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

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

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

Zoom on an illuminating special case (cont'd)

Consider the generalization error wrt responses only:

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

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

Proposition

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

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

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

Zoom on an illuminating special case (cont'd)

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

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

Another way of looking at this result. Set

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

Then

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

Zoom on an illuminating special case (cont'd)

Proof. Let us compute first $\mathbb{E}\left(\widetilde{\mathscr{R}}_n \mid \underline{X}\right)$ with (reminder)

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

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

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

Zoom on an illuminating special case (cont'd)

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

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

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

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

Zoom on an illuminating special case (cont'd)

Thus, we have:

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

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

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

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

□

Lecture outline

Training set and test set

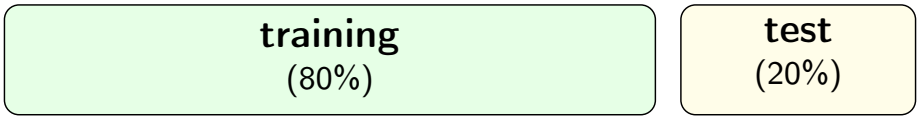
Conclusion/extrapolation. The empirical risk is in general

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

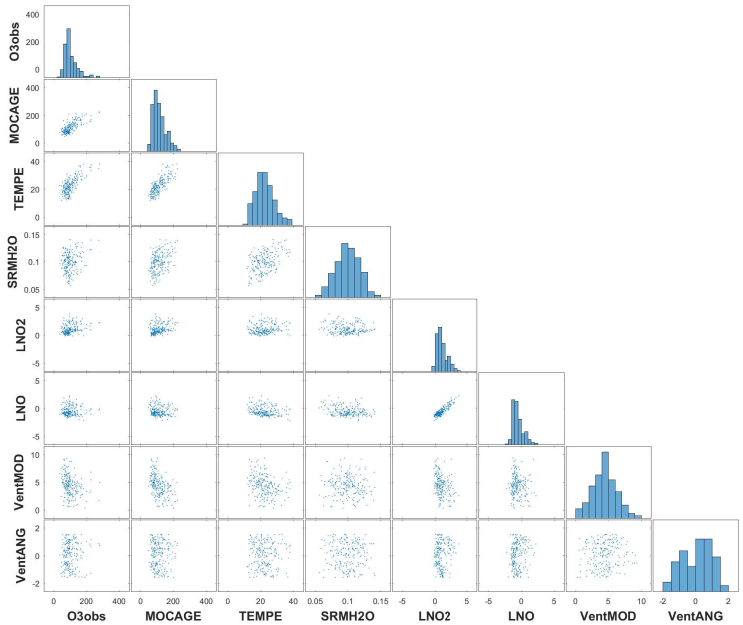
Solution: split the data in two sets

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

Example:



Exemple "Ozone" (cont'd from lecture #6)



Goal: predict the ozone concentation 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	$\hat{\mathscr{R}}_n$	$\hat{\mathscr{R}}_n^{\text{test}}$
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

Lecture outline

Lecture outline

Problem #1: choosing a "good" family $\mathscr{H}$

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 $\mathscr{H}_J$ 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 $\mathscr{H}_J$ with $k_J = J + 1$ parameters. ➠ complement

Problem: model selection

How to choose the family $\mathscr{H}_J$ (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: $\hat{\beta} = \text{argmin}\, \hat{\mathscr{R}}^{\text{pen}}_{n,\lambda}$, with

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

- 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 $\mathscr{H}_J$, or the hyperparameter λ, in order to minimize (an estimation of) the generalization error.

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

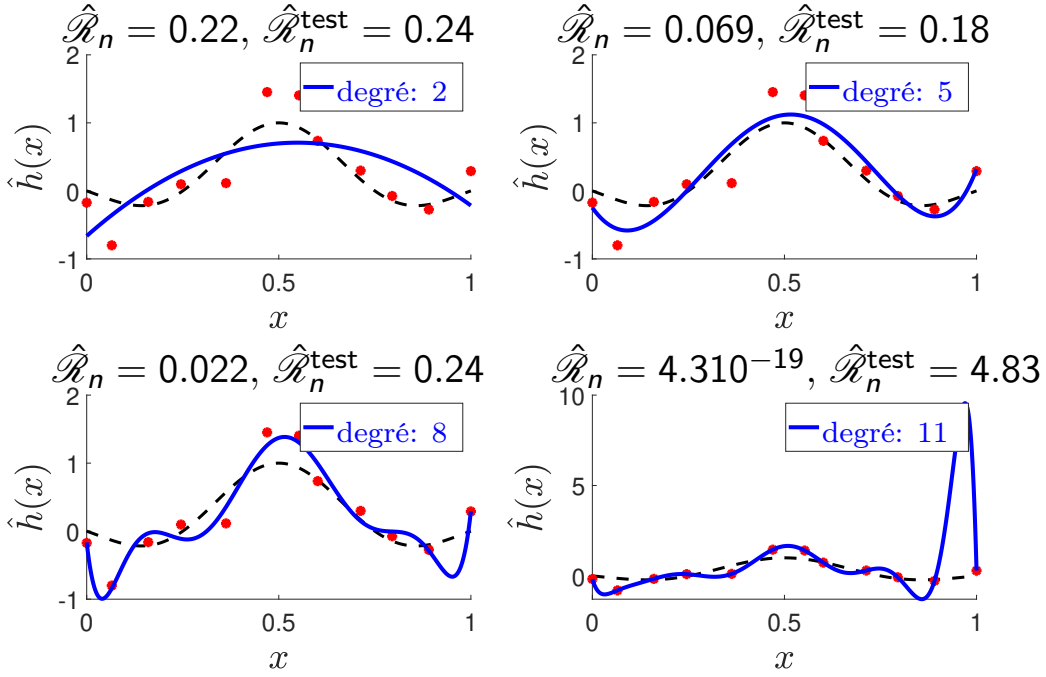
Example. Polynomial regression with $x \in \mathbb{R}$, degree $\leq J$:

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

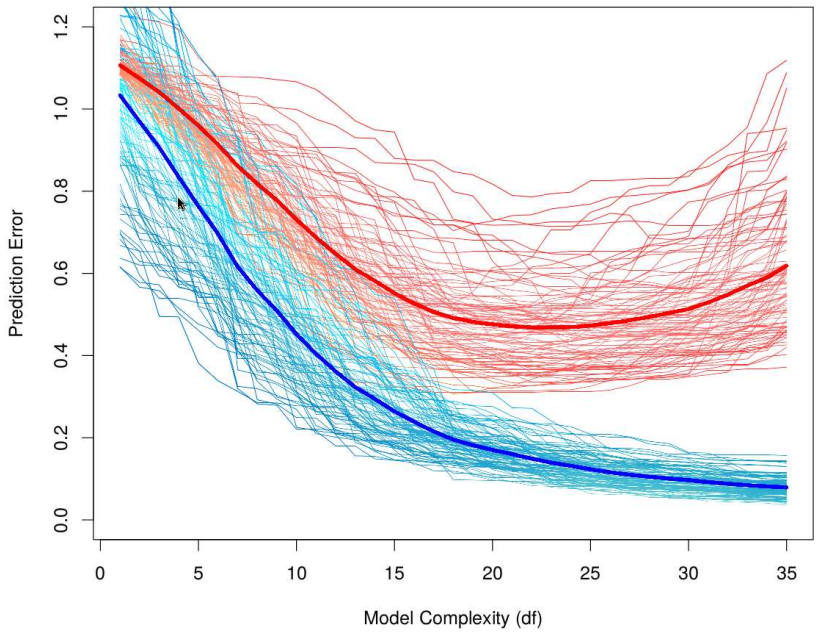
with $J = 2, 5, 8, 11$.

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 $\hat{\mathscr{R}}_n$ / 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 $\mathscr{H}$ or λ but...

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

Lecture outline

Solution: validation set

Idea: split the data in three sets

- training data: construct $\hat{h}$ with given $\mathscr{H}/\lambda$,
- validation set: choose $\mathscr{H}$, 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 $k = 4$:

train. 1 (20%)	**train. 1** (20%)	**train. 1** (20%)	**valid. 1** (20%)	
train. 2 (20%)	**train. 2** (20%)	**valid. 2** (20%)	**train. 2** (20%)	**test** (e.g., 20%)
train. 3 (20%)	**valid. 3** (20%)	**train. 3** (20%)	**train. 3** (20%)	
valid. 4 (20%)	**train. 4** (20%)	**train. 4** (20%)	**train. 4** (20%)	

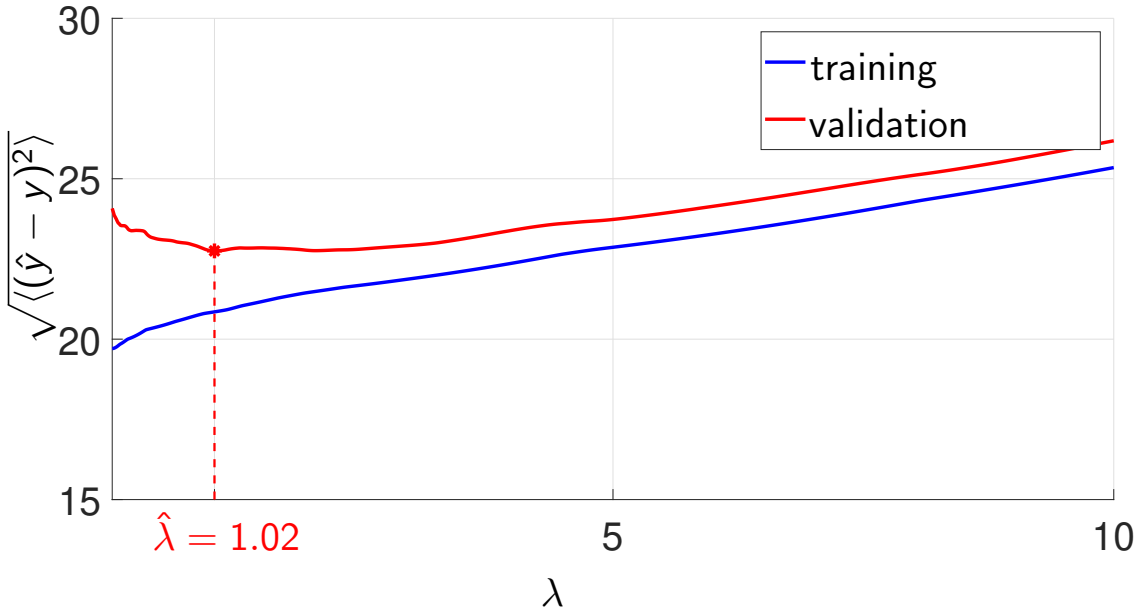
➠ the error is averaged over the k validation sets.

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 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

Lecture outline

Another approach to model selection: the AIC criterion

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

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

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

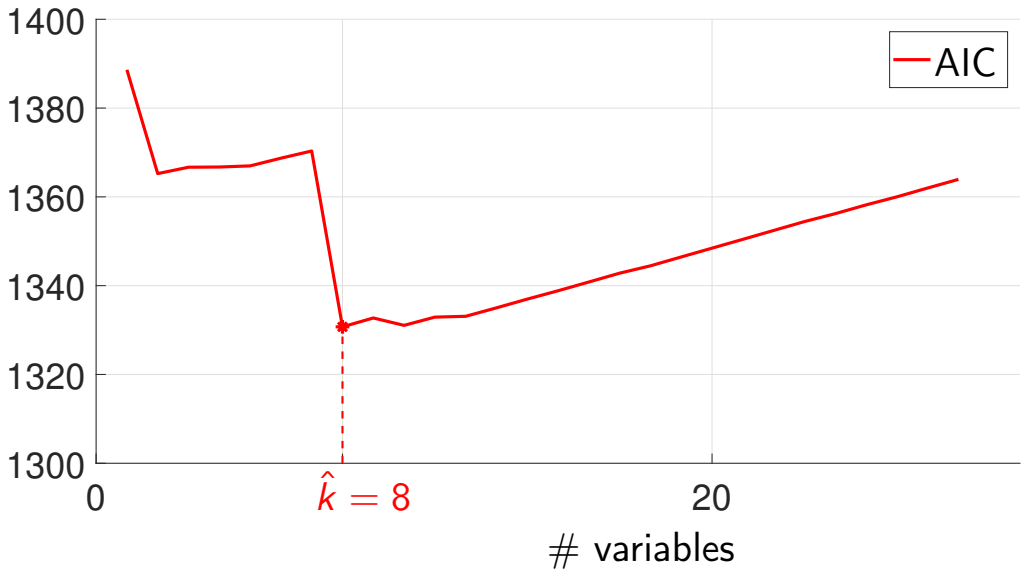
$$\hat{j} = \operatorname{argmin} \mathrm{AIC}(j), \qquad \mathrm{AIC}(j) \;=\; -2 \ln \mathcal{L}\left(\hat{\theta}_j^{\mathrm{MLE}}; \underline{X}, \underline{Y}\right) + 2k_j,$$

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

➠ 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.

Lecture outline

Lecture outline

Exercise 1 (Penalized regression)

solution

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

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

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

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

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

Exercise 1 (Penalized regression)

solution

slide 12

We denote $X = [X_1 \ldots X_n]^\top$, 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 $\hat{\beta}_j^R = c_{1,\lambda}\hat{\beta}_j$, $j = 1, \ldots, p$.

Exercise 1 (Penalized regression)

solution

slide 12

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

Solution of exercise 1

back to questions

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

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

2. This corresponds to ridge regression..

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

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

Solution of exercise 1

back to questions

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

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

Solution of exercise 1

back to questions

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:

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

Similarly, still assuming $x > 0$,

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

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.

Solution of exercise 1

back to questions

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

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

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

Solution of exercise 1

back to questions

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

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

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

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

Lecture outline

Lecture outline

Non-linearities in linear models...

If the empirical risk $\hat{\mathscr{R}}(\hat{h})$ 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: $x^{(1)}, \ldots, x^{(p)} \mapsto \tilde{x}^{(1)}, \ldots, \tilde{x}^{(q)}$

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

Examples

A few examples:

- scalar transformations: $\ln(x^{(j)})$, $\sqrt{x^{(j)}}$, $(x^{(j)})^k \ldots$
- 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 $\tilde{X}^{(m)} = \psi_m(X)$, $m = 1, \ldots, M$

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

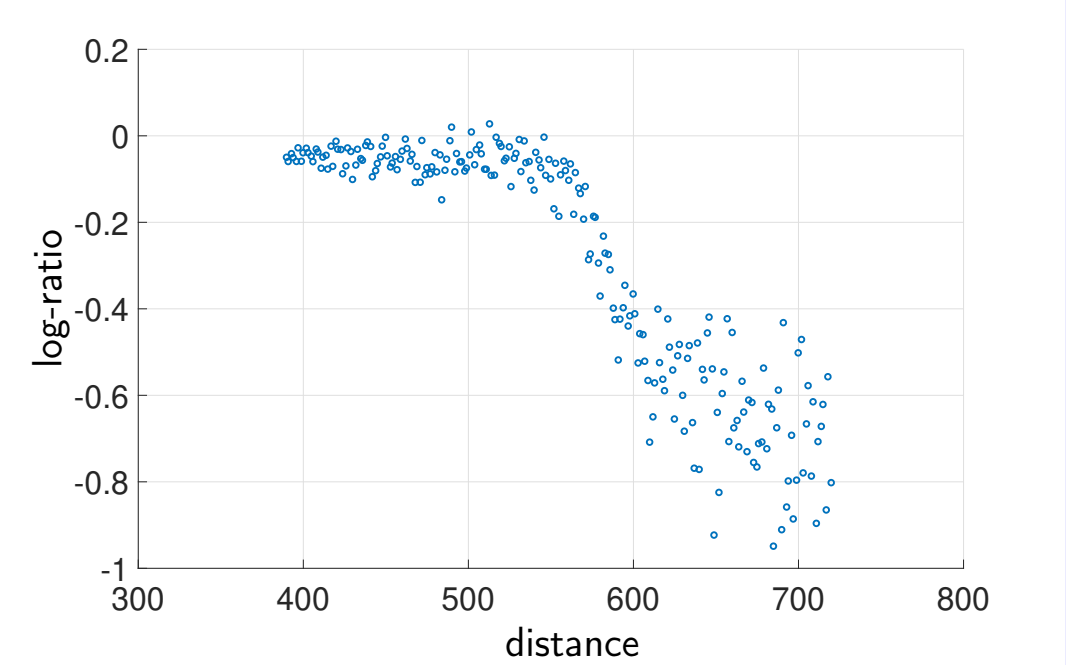
Examples of bases (preferably orthogonal):

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

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

➠ back to slide 28

Example: LIDAR data



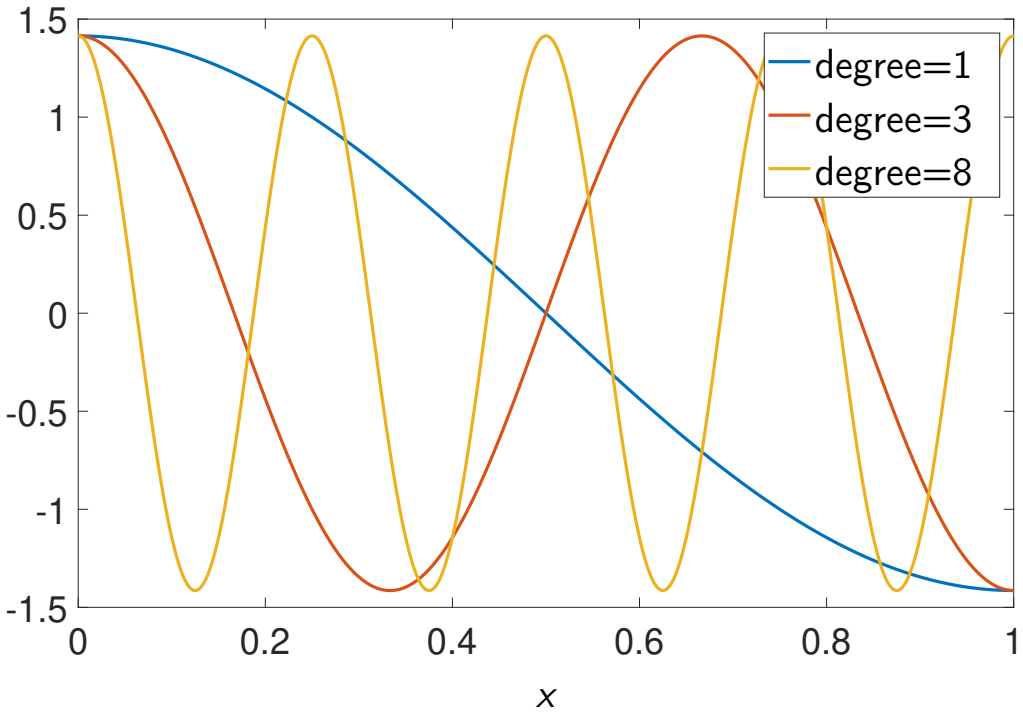
x-axis: distance travelled before the light is reflected back to its source
y-axis: logarithm of the ratio of received light from two laser sources

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

LIDAR: LIght Detection And Ranging

back to slide 28

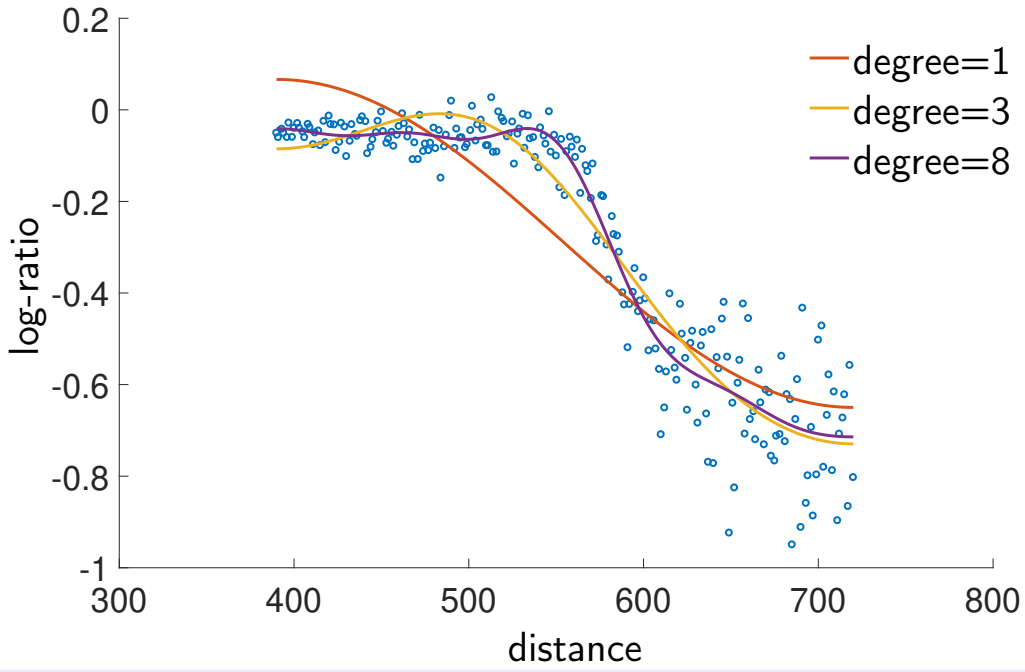
Basis of orthogonal cosines (basis of $L^2([0,1])$



back to slide 28

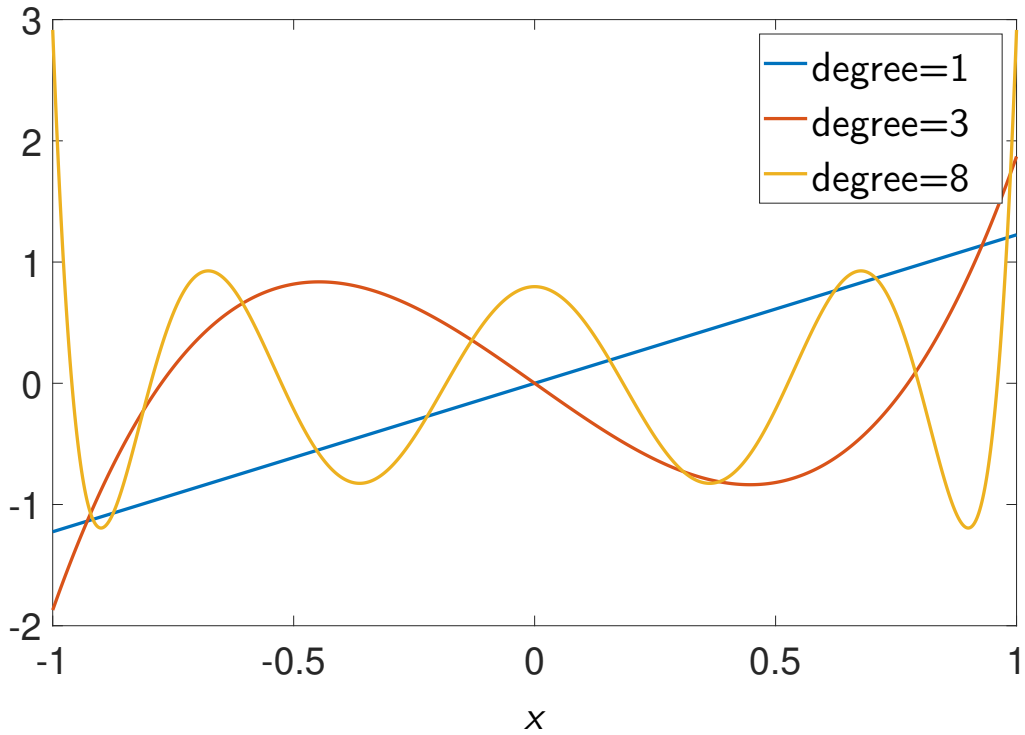
Example: LIDAR data (cont'd)

Quadratic loss + basis of cosines



back to slide 28

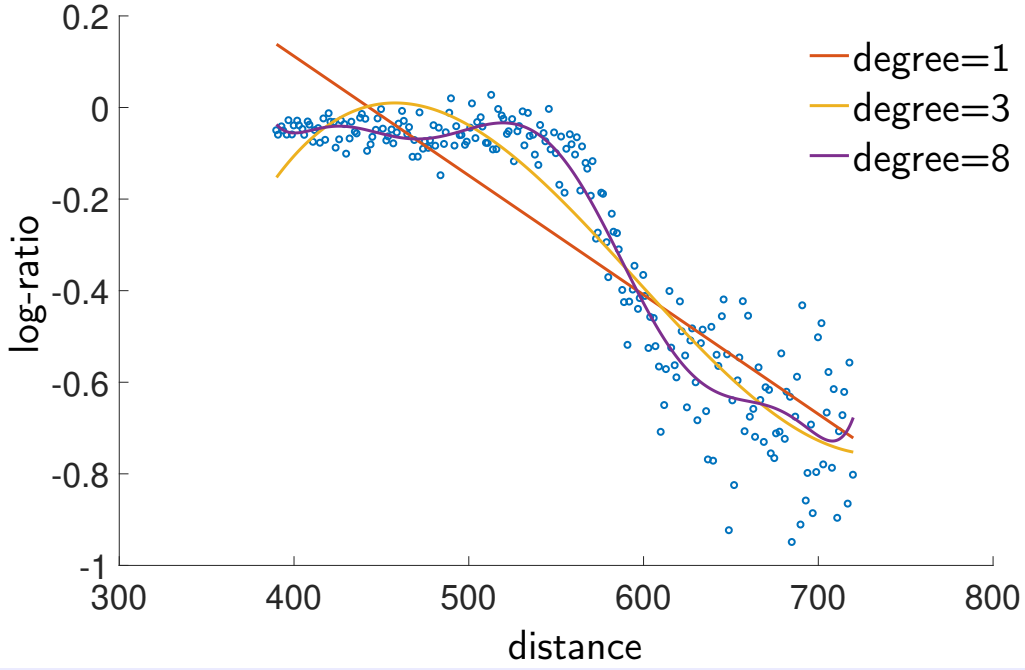
Legendre polynomials (orthonormal basis of $L^2([-1,1])$)



back to slide 28

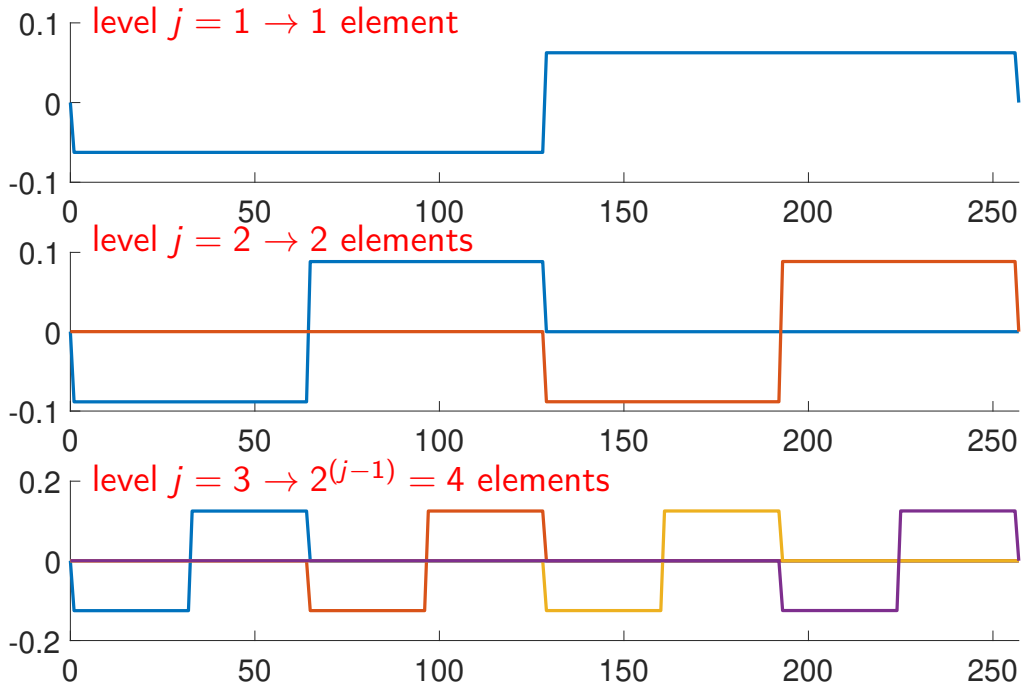
Example: LIDAR data (cont'd)

Quadratic loss + Legendre polynomials



back to slide 28

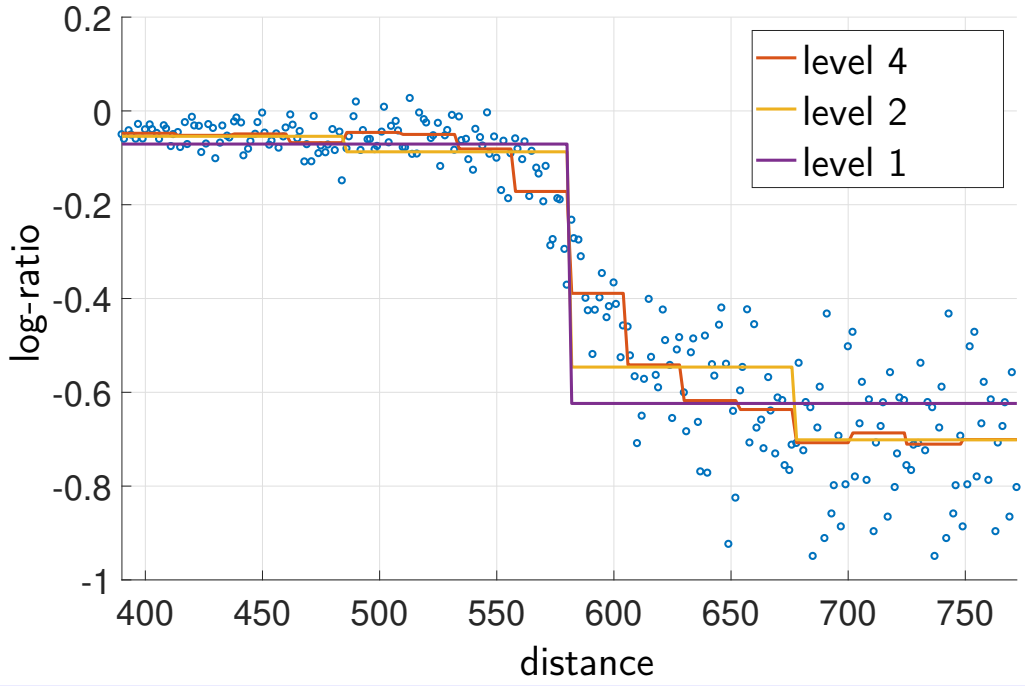
Haar wavelet basis



back to slide 28

Example: LIDAR data (cont'd)

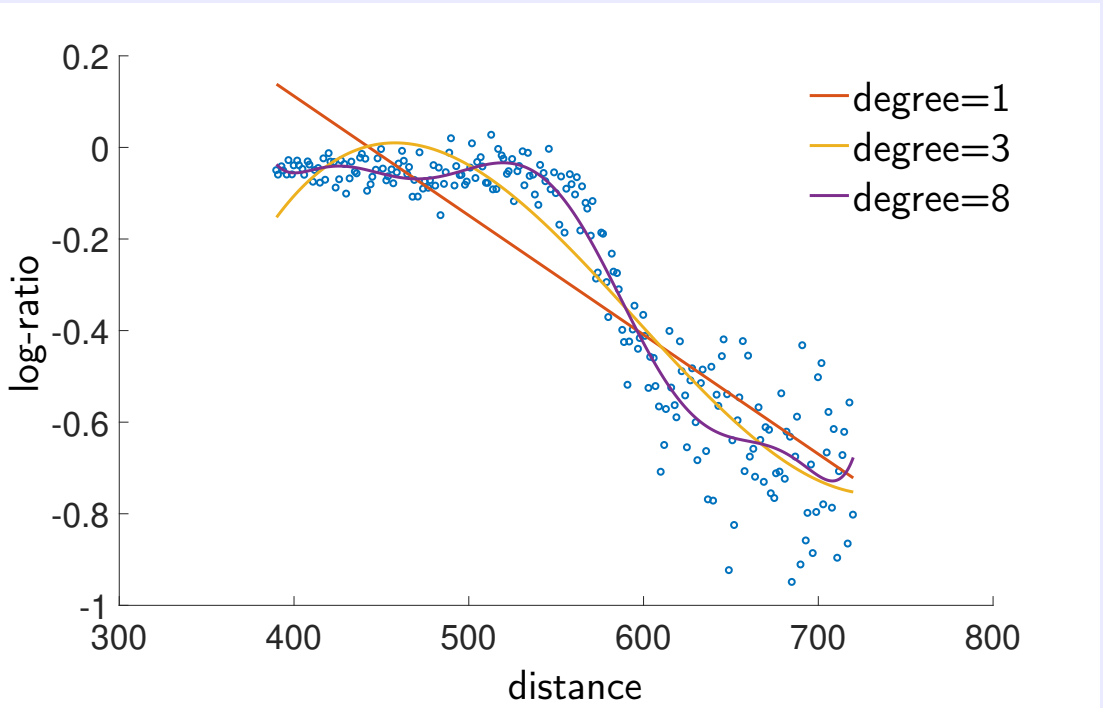
Quadratic loss + Haar wavelets



back to slide 28

Example: LIDAR data (cont'd)

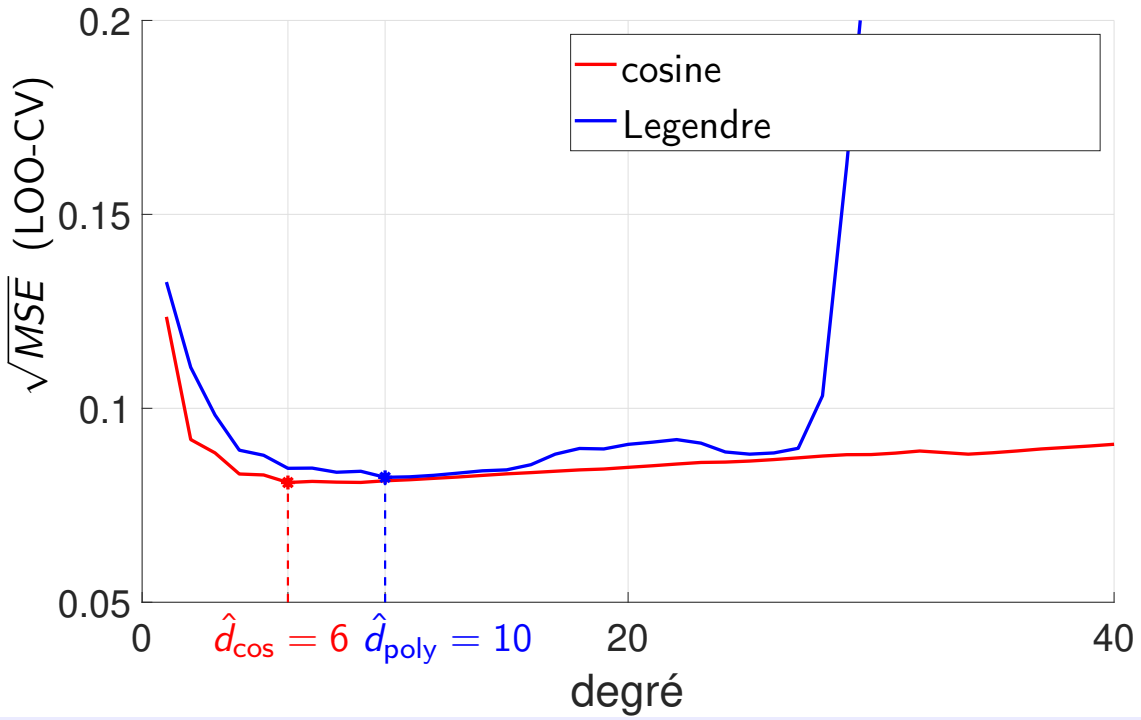
Quadratic loss + Legendre polynomials



back to slide 28

Example: LIDAR data (cont'd)

Model selection



back to slide 28