MA 423
Apprentissage et applications financières

Partie 1

Jérémie Bigot

DMIA / ISAE

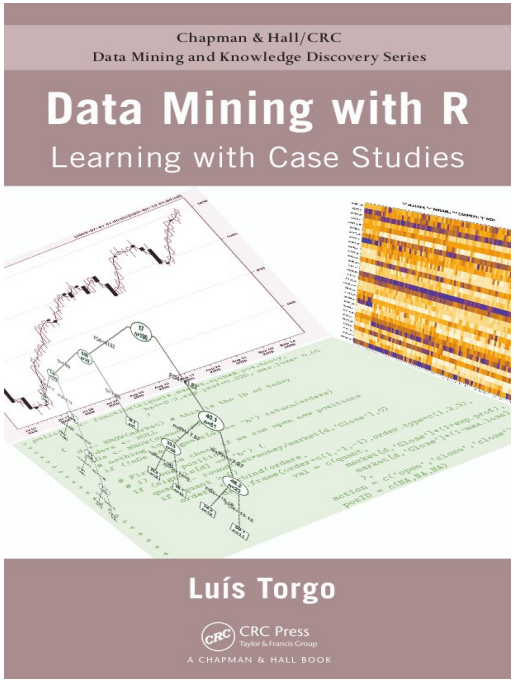
Janvier 2014

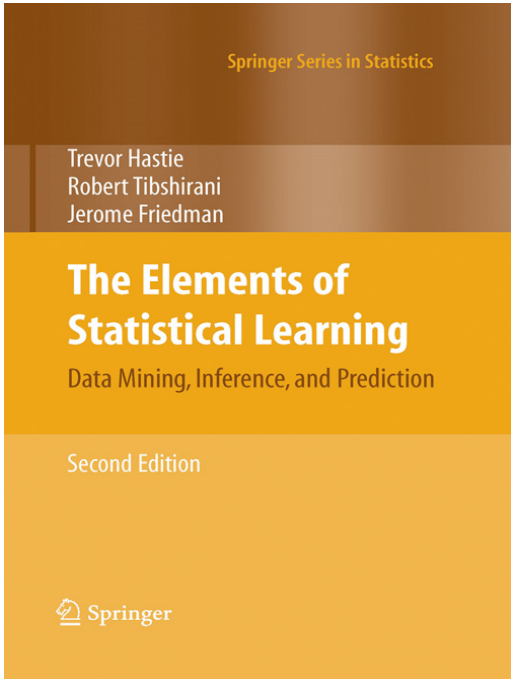
Plan du cours

- Exemple introductif : un problème d'apprentissage en finance
- Principes de base en apprentissage statistique
- Classification supervisée : règle de Bayes et analyse discriminante linéaire (LDA)
- Apprentissage par SVM (Support Vector Machine)
- Apprentissage par réseaux de neurones (Neural Networks)
- Principe de la validation croisée

Bibliographie & ressources logicielles

- Torgo L. (2010). *Data Mining with R : Learning with Case Studies*. Chapman and Hall/CRC.
- Hastie, T., Tibshirani, R. & Friedman, J. (2008). *The elements of statistical learning : data mining, inference and prediction*, Springer.





Bibliographie & ressources logicielles

- Logiciel libre R : The R Project for Statistical Computing
  `http://www.r-project.org`

- Packade *DMwR* : Data Mining with R - Chapitre 3 de Torgo (2010)

L'ensemble des expériences numériques présentées dans le cours sont disponibles sous la forme de codes R sur le LMS

Partie 1 - Introduction à l'apprentissage statistique

1. Un problème d'apprentissage en finance

2. Principes de base en apprentissage statistique
   - Compromis bias/variance en apprentissage
   - Principe de la pénalisation/régularisation

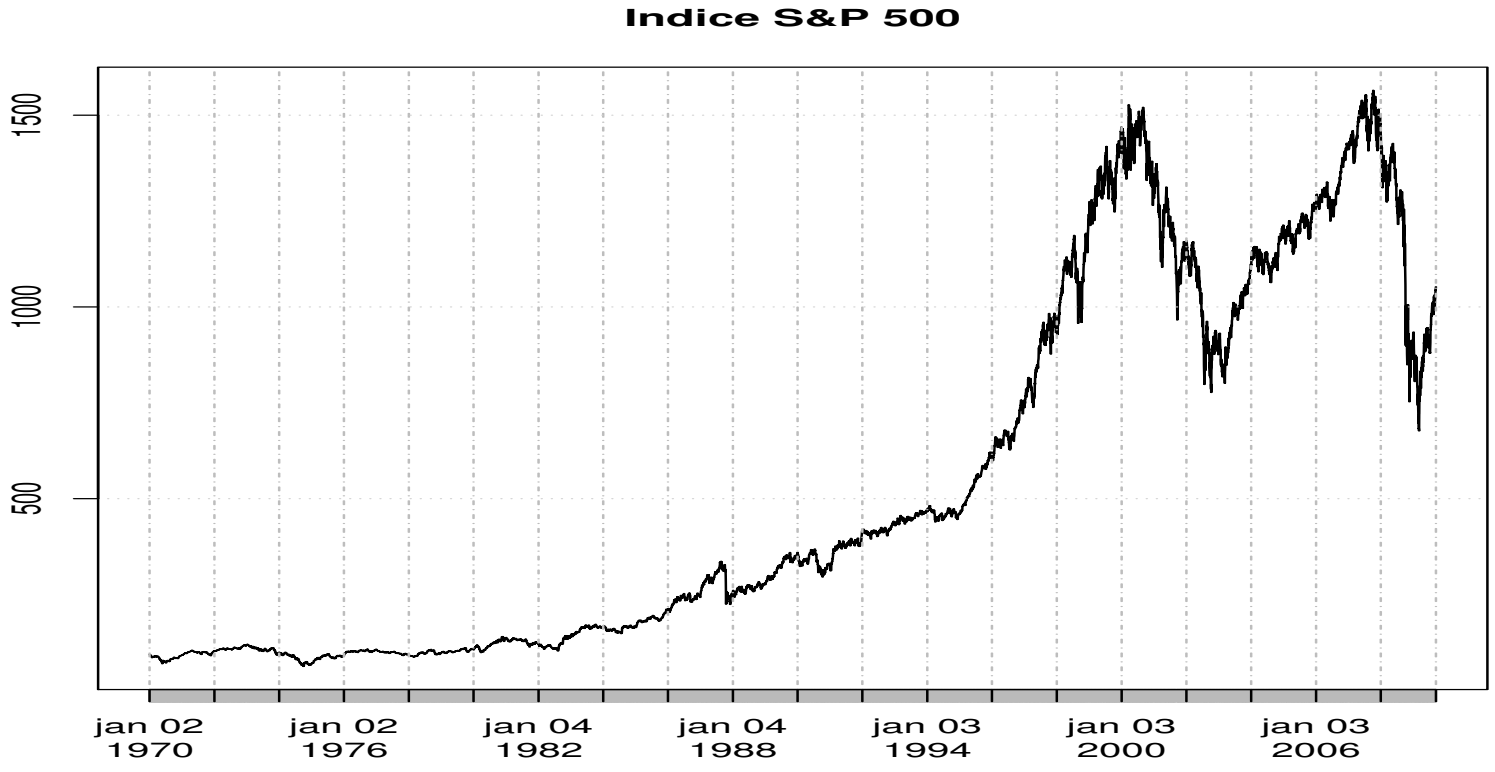
Problématique - Chapitre 3 de Torgo (2010)

But : définir une stratégie de trading à partir d'une prévision de l'évolution de l'indice S&P 500.

Indice S&P 500 :

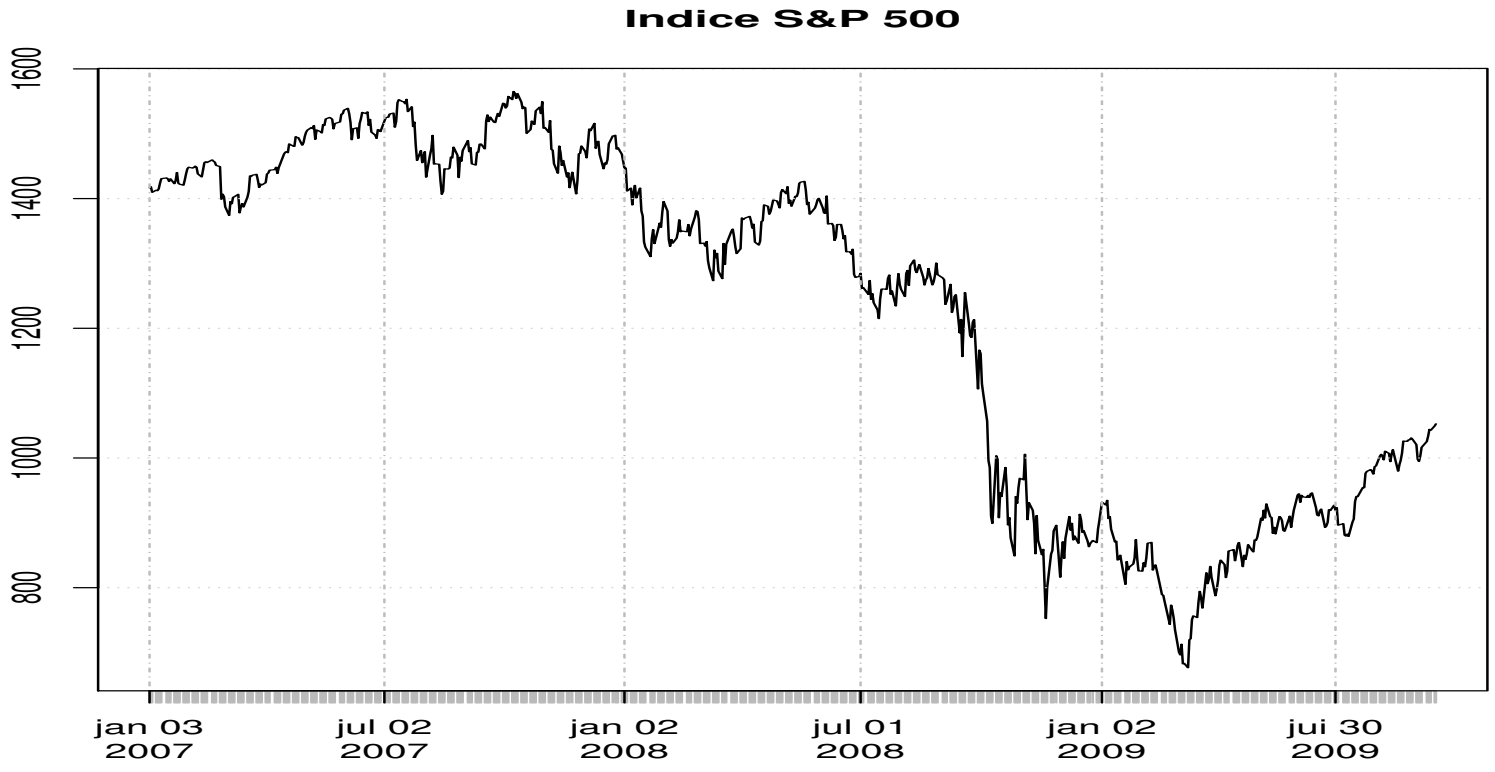
- publié par l'agence Standard & Poor's depuis mars 1957
- indice boursier pondéré des capitalisations boursières de 500 entreprises dont les titres sont les plus échangés sur les 2 principales places financières américaines (le NYSE et le NASDAQ)
- indice boursier représentatif des performances boursières des plus grandes entreprises américaines, et de l'ensemble de l'économie des USA
- de nombreux produits financiers dérivés sont adossés à sa cotation en continu

Evolution de l'indice S&P 500



Niveau de l'indice en clôture de session du 02/01/1970 au 15/09/2009

Evolution de l'indice S&P 500



Niveau de l'indice en clôture de session du 02/01/2007 au 15/09/2009

Crise des subprimes !

Données : l'indice S&P 500

TABLE: 6 séries temporelles

Date	Open	High	Low	Close	Volume	Adjusted
1970-01-02	92.06	93.54	91.79	93.00	8050000	93.00
1970-01-05	93.00	94.25	92.53	93.46	11490000	93.46
1970-01-06	93.46	93.81	92.13	92.82	11460000	92.82
1970-01-07	92.82	93.38	91.93	92.63	10010000	92.63
1970-01-08	92.63	93.47	91.99	92.68	10670000	92.68
1970-01-09	92.68	93.25	91.82	92.40	9380000	92.40

Premières lignes du tableau `GSPC` dans `R`

Définition d'un problème de prédiction

Questions :

- que cherche-t-on à prédire ?
- quel horizon temporel pour la prédiction ?

Statégie de trading :

- choix de vendre ou d'acheter si les prix varient plus que $p\%$ (à la hausse ou à la baisse)

Modèles de prédiction :

- déterminer si cette marge de $\pm p\%$ est atteignable dans les k jours suivants = prévoir la **dynamique des prix** au cours des jours $t+1, t+2, \ldots, t+k$ et non pas seulement si le prix au temps $t+k$ est de $\pm p\%$!

Définition d'un problème de prédiction

Construction d'un indicateur de tendance des prix :

- prix moyen $\bar{P}_i = \frac{C_i+H_i+L_i}{3}$, où C_i, H_i, L_i sont les prix `Close`, `High` et `Low` pour le jour i.
- on définit ensuite le vecteur V_i de taille k par

$$V_i = \left\{ \frac{\bar{P}_{i+j} - C_i}{C_i} \right\}_{j=1}^{k}.$$

- l'indicateur considéré est alors

$$T_i = \sum_{v \in V_i} v \mathbb{1}_{\{ v>p\% \cup v<-p\% \}}$$

Interprétation : T_i grand = hausse des prix = choix d'acheter : *buy* si $T_i > \lambda$

Définition d'un problème de prédiction

Construction d'un indicateur de tendance des prix :

- prix moyen $\bar{P}_i = \frac{C_i+H_i+L_i}{3}$, où C_i, H_i, L_i sont les prix `Close`, `High` et `Low` pour le jour i.
- on définit ensuite le vecteur V_i de taille k par

$$V_i = \left\{ \frac{\bar{P}_{i+j} - C_i}{C_i} \right\}_{j=1}^{k}.$$

- l'indicateur considéré est alors

$$T_i = \sum_{v \in V_i} v \mathbb{1}_{\{ v>p\% \cup v<-p\% \}}$$

Interprétation : T_i fortement négatif = baisse des prix = choix de vendre : *sell* si $T_i < -\lambda$

Définition d'un problème de prédiction

Construction d'un indicateur de tendance des prix :

- prix moyen $\bar{P}_i = \frac{C_i+H_i+L_i}{3}$, où C_i, H_i, L_i sont les prix `Close`, `High` et `Low` pour le jour i.
- on définit ensuite le vecteur V_i de taille k par

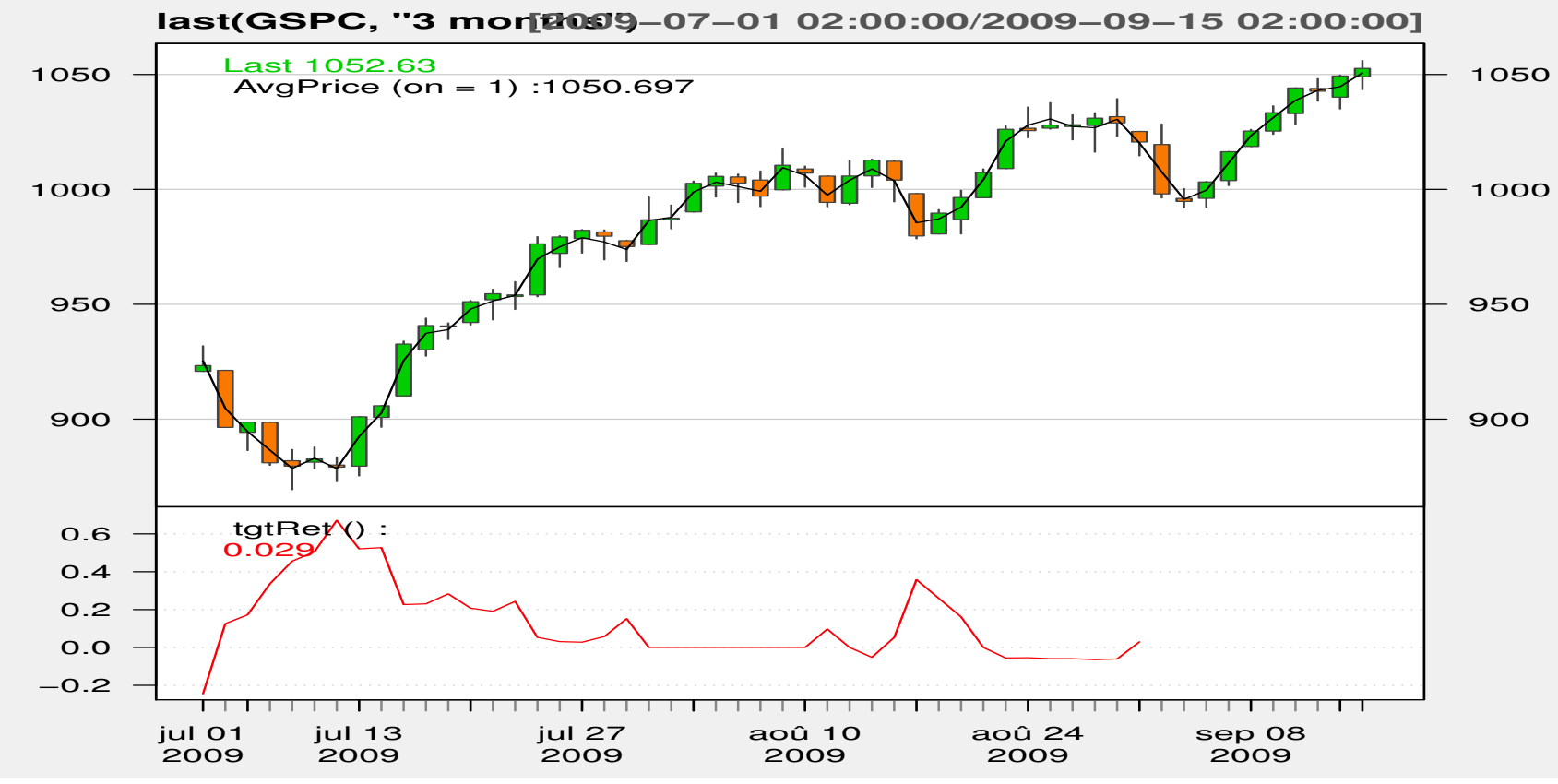
$$V_i = \left\{ \frac{\bar{P}_{i+j} - C_i}{C_i} \right\}_{j=1}^{k}.$$

- l'indicateur considéré est alors

$$T_i = \sum_{v \in V_i} v \mathbb{1}_{\{ v>p\% \cup v<-p\% \}}$$

Interprétation : T_i proche de zéro = stagnation des prix = choix de ne rien faire : *hold* si $|T_i| \leq \lambda$

Exemple d'évolution de l'indicateur T



Niveau de l'indicateur T du 01/07/2009 au 15/09/2009 pour $p = 2.5\%$ et $k = 10$

Choix des prédicteurs

Question : quelles variables explicatives (prédicteurs) peut-on utiliser pour prédire T_i à la fin du jour i ?

Hypothèse : il est possible de prédire l'évolution **future** du marché (ici l'indicateur T_i) à partir de l'observation du **comportement passé** du marché

Indicateurs du comportement passé :

- 1ère approche par des méthodes issues de l'analyse statistique de la série temporelle des T_i

Choix des prédicteurs

Question : quelles variables explicatives (prédicteurs) peut-on utiliser pour prédire T_i à la fin du jour i ?

Hypothèse : il est possible de prédire l'évolution **future** du marché (ici l'indicateur T_i) à partir de l'observation du **comportement passé** du marché

Indicateurs du comportement passé :

- ~~1ère approche par des méthodes issues de l'analyse statistique de la série temporelle des T_i~~
- 2ème approche : utilisation d'une série de 22 indicateurs techniques relatifs au comportement passé du marché

Choix des prédicteurs

Question : quelles variables explicatives (prédicteurs) peut-on utiliser pour prédire T_i à la fin du jour i ?

Hypothèse : il est possible de prédire l'évolution **future** du marché (ici l'indicateur T_i) à partir de l'observation du **comportement passé** du marché

Indicateurs du comportement passé :

- ~~1ère approche par des méthodes issues de l'analyse statistique de la série temporelle des T_i~~
- 2ème approche : utilisation d'une série de 27 indicateurs techniques relatifs au comportement passé du marché
  - 10 indicateurs définis par : $R_{i-h} = \frac{C_i - C_{i-h}}{C_{i-h}}$ (variation des prix lors des $1 \leq h \leq 10$ jours précédents)
  - 17 autres indicateurs issus du package `TTR` de `R` : `ATR, SMI, ADX, Aroon, BB, ChaikinVol, CLV, EMV, MACD, MFI, SAR, Volat,...`

Base de données prédicteurs/variable à prédire

TABLE: Les données sur 3 jours

Date	T.ind.GSPC.	Delt.1.arithmetic	Delt.2.arithmetic	Delt.3.arithmetic	Delt.4.arithmetic	Delt.5.arithmetic
1970-11-11	-0.02536360	0.002830522	0.004251801	0.009617668	0.0110582640	0.0075838369
1970-11-12	0.00000000	-0.010349288	-0.007548060	-0.006141490	-0.0008311565	0.0005945303
1970-11-13	0.06868978	-0.009269162	-0.019522521	-0.016747258	-0.0153537262	-0.0100926146
Date	Delt.6.arithmetic	Delt.7.arithmetic	Delt.8.arithmetic	Delt.9.arithmetic	Delt.10.arithmetic	atr
1970-11-11	0.009617668	0.0182014130	0.021381381	0.020033589	0.019177754	1.459246
1970-11-12	-0.002843939	-0.0008311565	0.007663753	0.010810811	0.009476967	1.478586
1970-11-13	-0.008680143	-0.0120867401	-0.010092615	-0.001676446	0.001441441	1.473687
Date	SMI	ADX	oscillator	pctB	Delt.1.arithmetic.1	clv.EMA.10
1970-11-11	3.157942	17.49590	50	1.0652366	-0.5310337	-0.08581241
1970-11-12	3.616957	17.33069	50	0.7387661	-11.1688511	-0.18056223
1970-11-13	1.262925	16.09806	50	0.3666488	0.1488204	-0.21349675
Date	CMO.Cl.GSPC.	Delt.1.EMA.10	maEMV	myVolat.GSPC.	signal	mfi
1970-11-11	42.416452	0.0018259548	2.853887e-04	0.1843347	0.2735797	73.60305
1970-11-12	8.675799	-0.0003877258	2.046560e-04	0.1877988	0.2865031	65.81745
1970-11-13	1.276596	-0.0020025324	4.055646e-05	0.1871654	0.2750557	64.68935
Date	Close.EMA.14	mySAR.GSPC.	runMean.Cl.GSPC.	runSD.Cl.GSPC.		
1970-11-11	62.19452	83.58937	84.154	0.6117225		
1970-11-12	52.06064	83.85444	84.233	0.5451819		
1970-11-13	45.05338	86.24000	84.245	0.5219674		

Base de données prédicteurs/variable à prédire

Pour chaque jour $1 \leq i \leq n$ on dispose de :

- 27 prédicteurs stockés sous la forme d'un vecteur $X_i \in \mathbb{R}^p$ avec $p = 27$ où

$$X_i = \begin{pmatrix} R_{i-1} \\ \vdots \\ R_{i-10} \\ ATR_i \\ \vdots \\ Volat_i \\ \vdots \end{pmatrix} \in \mathbb{R}^p$$

- d'une variable à expliquer/prédire Y_i de deux types possibles

  $Y_i = T_i \in \mathbb{R}$, problème de **régression**

  $Y_i \in \{buy; sell; hold\}$, problème de **classification supervisée**

  où *buy* si $T_i > \lambda$, *sell* si $T_i < -\lambda$, et *hold* si $|T_i| < \lambda$.

Base de données prédicteurs/variable à prédire

Le problème de l'apprentissage statistique est alors le suivant :

étant donné une base d'apprentissage $(X_i, Y_i)_{1 \leq i \leq n}$ où $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$ (**régression**) ou $Y_i \in \{1; 2; \ldots; K\}$ (**classification**), on souhaite :

- déterminer un modèle qui permet de lier l'entrée X_i à la sortie Y_i pour tout $1 \leq i \leq n$
- pour $i_0 \notin \{1, \ldots, n\}$ on veut déterminer $\hat{Y}_{i_0}$ prédiction de Y_{i_0} (**non-observé**) au vu de l'observation de X_{i_0}

1 Un problème d'apprentissage en finance

2 Principes de base en apprentissage statistique
- Compromis bias/variance en apprentissage
- Principe de la pénalisation/régularisation

Modèle de régression en statistique

Base d'apprentissage $(X_i, Y_i)_{1 \leq i \leq n}$ où $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$ (**régression**)

Modèle de régression standard

$$Y_i = f^*(X_i) + \epsilon_i, \ i = 1, \ldots, n,$$
$$\epsilon_1, \ldots, \epsilon_n \text{ iid avec } \mathbb{E}(\epsilon_i) = 0, \ \text{Var}(\epsilon_i) = \sigma^2,$$

et $f^* : \mathbb{R}^p \to \mathbb{R}$ est une fonction **inconnue**.

But de la régression : estimer $f^*(\mathbf{x})$ en tout point $\mathbf{x} \in \mathbb{R}^p$ à partir de la base d'apprentissage.

Principales hypothèses

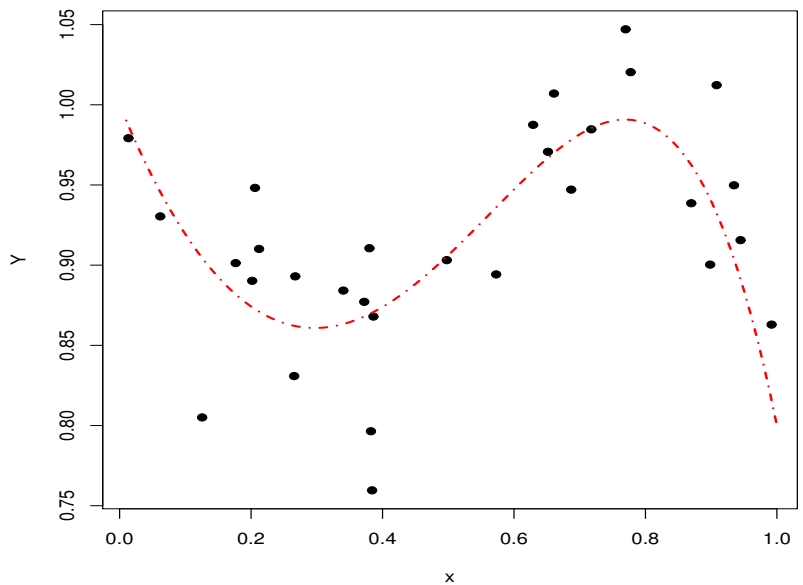
- f^* appartient à une classe fonctionnelle $\mathcal{F} \subset \{f : \mathbb{R}^p \to \mathbb{R}\}$ de fonctions mesurables
- $\epsilon_1, \ldots, \epsilon_n \sim_{iid} \mathcal{N}(0, \sigma^2)$

L'exemple de la régression polynomiale

Données : observations bruitées d'un polynôme de degré $m^* \in \mathbb{N}^*$ **inconnu**

$$Y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \ldots + \theta_{m^*} x_i^{m^*} + \epsilon_i, \ i = 1, \ldots, n,$$

où $x_i \in \mathbb{R}$ est **déterministe** (instant d'observation), et $(\theta_0, \theta_1, \ldots, \theta_{m^*})$ sont des coefficients réels **inconnus**.



$n = 30$ observations et polynôme de degré $m^* = 5$ (trait pointillé)

L'exemple de la régression polynomiale

Données : observations bruitées d'un polynôme de degré $m^* \in \mathbb{N}^*$ **inconnu**

$$Y_i = \theta_0^* + \theta_1^* x_i + \theta_2^* x_i^2 + \ldots + \theta_{m^*}^* x_i^{m^*} + \epsilon_i, \ i = 1, \ldots, n,$$

où $x_i \in \mathbb{R}$ est **déterministe** (instant d'observation), et $(\theta_0^*, \theta_1^*, \ldots, \theta_{m^*}^*)$ sont des coefficients réels **inconnus**.

Base d'apprentissage $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$ avec $\mathbf{x}_i = (1, x_i, x_i^2, \ldots, x_i^M)^t \in \mathbb{R}^p$ où $M \geq m^*$ est un entier suffisamment grand et $p = M + 1$.

Modèle de **régression linéaire**

$$Y_i = f^*(\mathbf{x}_i) + \epsilon_i, \ i = 1, \ldots, n,$$

avec $f^*(\mathbf{x}_i) = \langle \mathbf{x}_i, \theta \rangle$ avec $\theta \in \mathbb{R}^p$.

L'exemple de la régression polynomiale

Modèle de **régression linéaire**

$$Y_i = f^*(\mathbf{x}_i) + \epsilon_i,\ i = 1, \ldots, n,$$

avec $f^*(\mathbf{x}_i) = \langle \mathbf{x}_i, \theta \rangle$ avec $\theta \in \mathbb{R}^p$ et $\mathbf{x}_i = (1, x_i, x_i^2, \ldots, x_i^M)^t \in \mathbb{R}^p$.

Problème : quelle classe fonctionnelle choisir pour f^* ?

Définissons pour $1 \leq m \leq M$:

$$\mathcal{F}_m = \left\{ f \ : \ f(x) = \sum_{k=0}^{m} \theta_k x^k,\ x \in \mathbb{R}, (\theta_k)_{0 \leq k \leq m} \in \mathbb{R}^{m+1} \right\}$$

Remarque : $f(x) \equiv f(\mathbf{x}) = \langle \mathbf{x}, \theta \rangle$ avec $\mathbf{x} = (1, x, \ldots, x^m)^t$ et $\theta \in \mathbb{R}^{m+1}$.

L'exemple de la régression polynomiale

Modèle de **régression linéaire**

$$Y_i = f^*(\mathbf{x}_i) + \epsilon_i, \ i = 1, \ldots, n,$$

Etape 1 : choix d'un estimateur $\hat{f}_m \in \mathcal{F}_m$ par minimisation des moindres carrés

$$\hat{f}_m = \underset{f \in \mathcal{F}_m}{\arg\min} \left\{ \sum_{i=1}^{n} (Y_i - f(\mathbf{x}_i))^2 \right\}$$

Remarque : dans un modèle gaussien, c'est équivalent à l'estimateur du maximum de vraisemblance.

L'exemple de la régression polynomiale

Pour $f \in \mathcal{F}_m, f(\mathbf{x}_i) = \langle \mathbf{x}_i, \theta \rangle$ avec $\mathbf{x}_i = (1, x_i, \ldots, x_i^m)^t$ et $\theta \in \mathbb{R}^{m+1}$

Etape 1 : revient à la résolution des moindres carrés classiques

$$\hat{\boldsymbol{\theta}}^{(m)} = \underset{\theta \in \mathbb{R}^{m+1}}{\arg\min} \left\{ \|\mathbf{Y} - \mathbf{X}_m \theta\|^2 \right\},$$

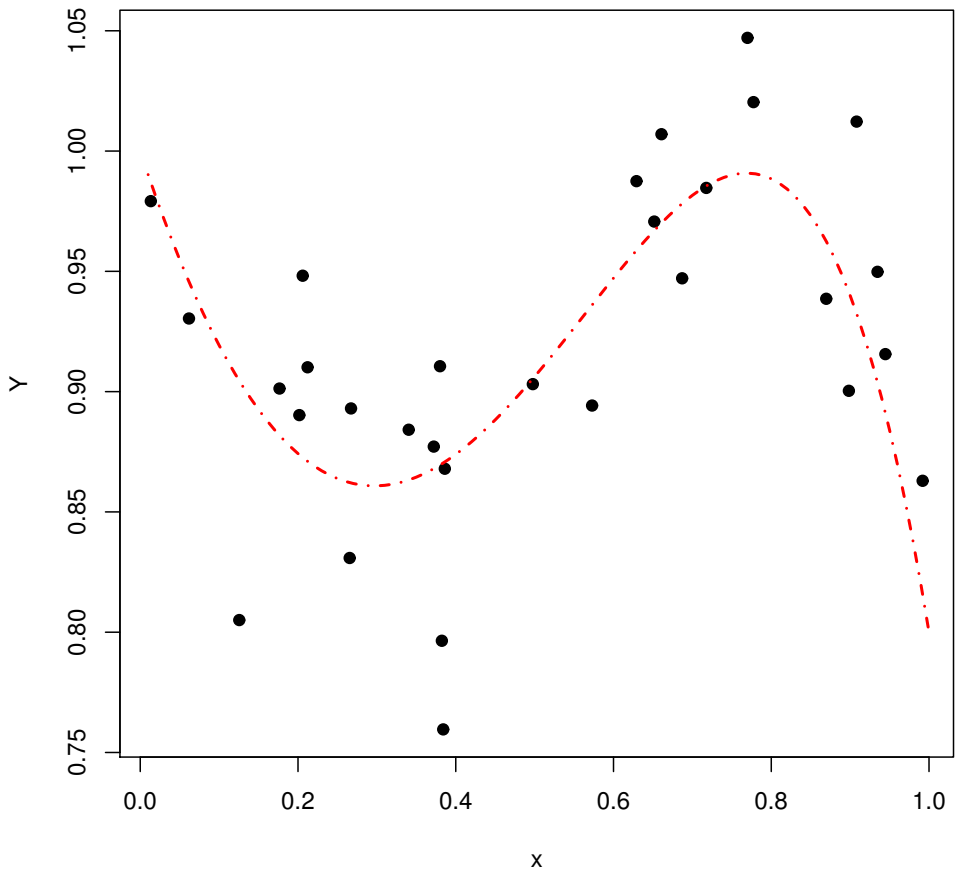
avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^t \in \mathbb{R}^n$ et $\mathbf{X}_m$ matrice de taille $n \times (m+1)$ qui s'écrit sous la forme :

$$\mathbf{X}_m = \begin{pmatrix} 1 & x_1 & \ldots & x_1^m \\ 1 & x_2 & \ldots & x_2^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^m \end{pmatrix}$$

L'estimation par moindres carrés conduit à :

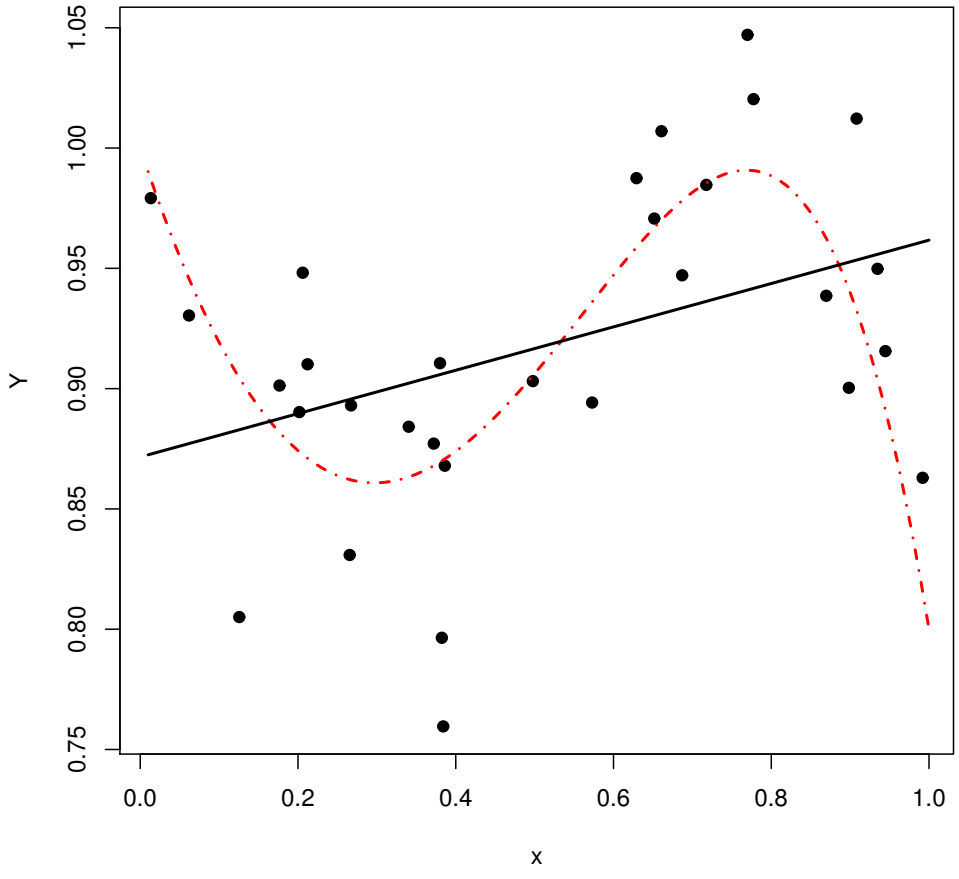
$$\hat{\boldsymbol{\theta}}^{(m)} = (\mathbf{X}_m^t \mathbf{X}_m)^{-1} \mathbf{X}_m^t \mathbf{Y} \quad \text{et} \quad \hat{f}_m(\mathbf{x}) = \langle \mathbf{x}, \hat{\boldsymbol{\theta}}^{(m)} \rangle$$

L'exemple de la régression polynomiale



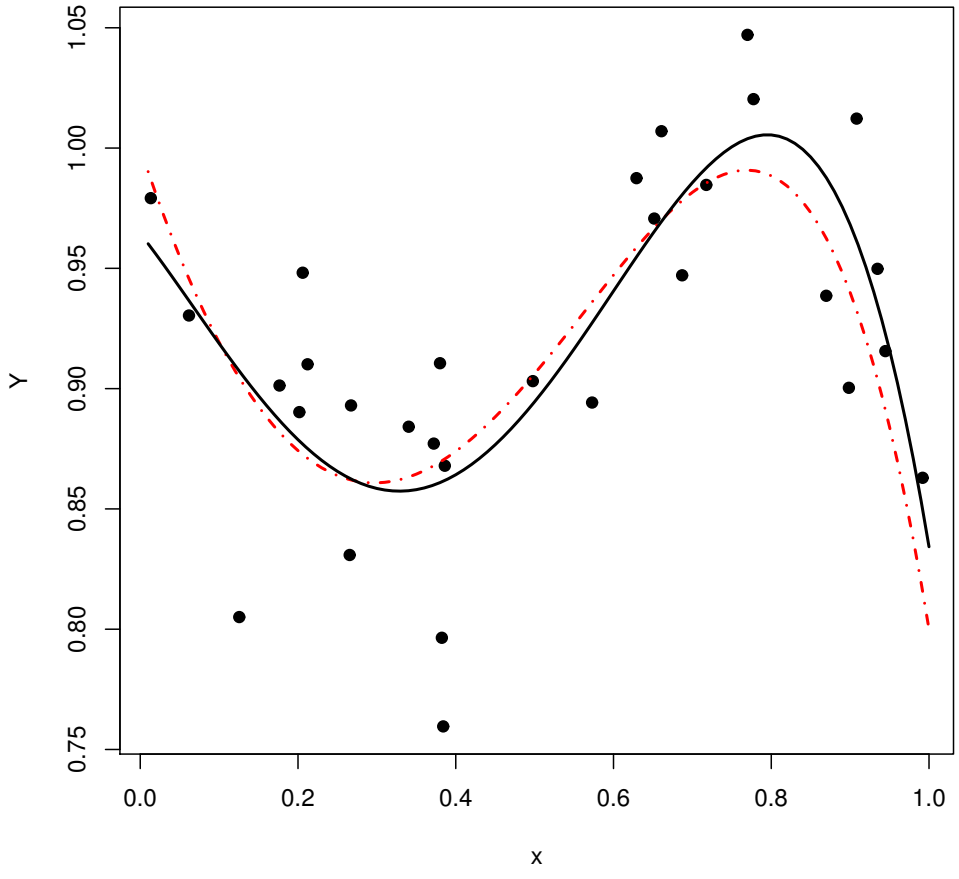
$n = 30$ observations et polynôme de degré $m^* = 5$ (trait pointillé)

L'exemple de la régression polynomiale



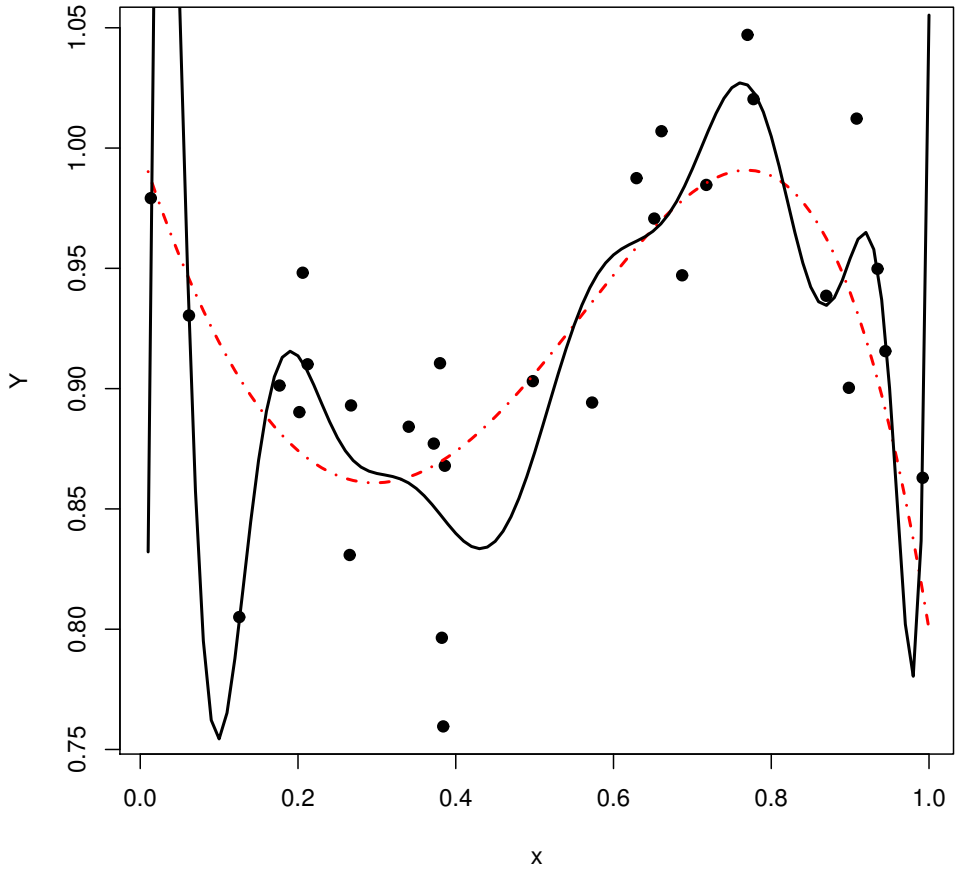
Estimation par un polynôme de degré $m = 1$

L'exemple de la régression polynomiale



Estimation par un polynôme de degré $m = 4$

L'exemple de la régression polynomiale



Estimation par un polynôme de degré $m = 15$

Décomposition biais/variance d'un estimateur

Modèle de régression standard

$$Y_i = f^*(\mathbf{x}_i) + \epsilon_i, \text{ avec } \mathbb{E}(\epsilon_i) = 0, \text{ Var}(\epsilon_i) = \sigma^2, \ i = 1, \ldots, n,$$

où $f^* : \mathbb{R}^p \to \mathbb{R}$ est une fonction **inconnue**, et $\mathbf{x}_i \in \mathbb{R}^p$ sont des vecteurs non-aléatoires.

Definition

On appelle **risque quadratique** d'un estimateur $\hat{f}$ de f^* la quantité

$$R(\hat{f}, f^*) = \sum_{i=1}^{n} \mathbb{E}\left(f^*(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right)^2$$

Sélection d'un estimateur : choisir $\hat{f}$ qui minimise $R(\hat{f}, f^*)$.

Question : comment estimer $R(\hat{f}, f^*)$ qui est **a priori inconnu** à partir de la base d'apprentissage ?

Décomposition biais/variance d'un estimateur

Proposition

Soit $\hat{f}$ un estimateur de la fonction de régression f^*. On définit le biais et la variance de l'estimateur $\hat{f}$ par

$$\begin{aligned} \mathit{Biais}^2(\hat{f}) &= \sum_{i=1}^{n} \left(\mathbb{E}[\hat{f}(\mathbf{x}_i)] - f^*(\mathbf{x}_i) \right)^2 \\ \mathit{Var}(\hat{f}) &= \sum_{i=1}^{n} \mathit{Var}[\hat{f}(\mathbf{x}_i)] = \sum_{i=1}^{n} \mathbb{E}\left(\hat{f}(\mathbf{x}_i) - \mathbb{E}[\hat{f}(\mathbf{x}_i)] \right)^2 \end{aligned}$$

Le risque quadratique se décompose alors de la façon suivante :

$$R(\hat{f}, f^*) = \mathit{Biais}^2(\hat{f}) + \mathit{Var}(\hat{f})$$

Décomposition biais/variance d'un estimateur

Exemple de la régression polynomiale. Notations

$$\underline{\mathbf{f}}^* = \begin{pmatrix} f^*(\mathbf{x}_1) \\ \vdots \\ f^*(\mathbf{x}_n) \end{pmatrix} \quad \text{et} \quad \hat{\underline{\mathbf{f}}}_m = \begin{pmatrix} \hat{f}_m(\mathbf{x}_1) \\ \vdots \\ \hat{f}_m(\mathbf{x}_n) \end{pmatrix}.$$

On a alors que

$$\hat{\underline{\mathbf{f}}}_m = \mathbf{X}_m \hat{\boldsymbol{\theta}}^{(m)} = \mathbf{X}_m (\mathbf{X}_m^t \mathbf{X}_m)^{-1} \mathbf{X}_m^t \mathbf{Y} = P_m \mathbf{Y}$$

où P_m est la matrice de projection orthogonale sur

$$V_m = \left\{ \mathbf{X}_m \theta ; \theta \in \mathbb{R}^{m+1} \right\}.$$

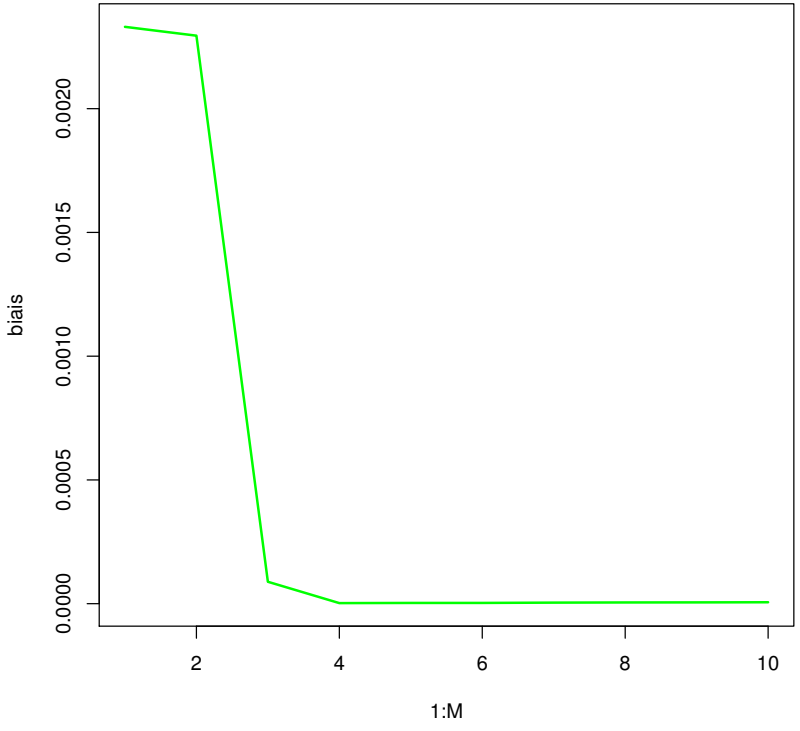
Décomposition biais/variance en régression polynomiale

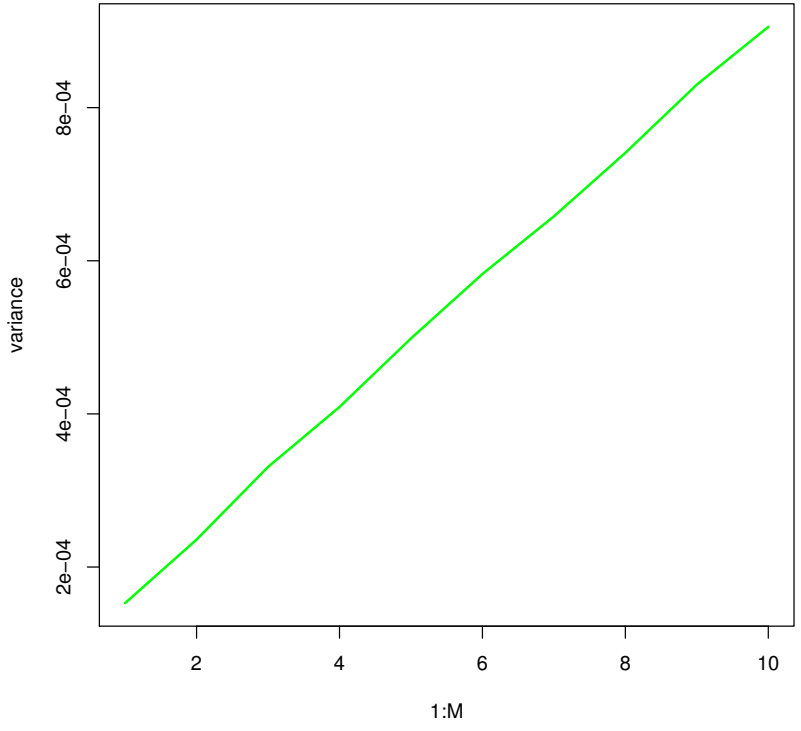
$$\begin{aligned} \text{Biais}^2(\hat{\underline{\mathbf{f}}}_m) &= \|P_m \underline{\mathbf{f}}^* - \underline{\mathbf{f}}^*\|^2 \\ \text{Var}(\hat{\underline{\mathbf{f}}}_m) &= \sigma^2 (m+1) \\ R(\underline{\mathbf{f}}^*, \hat{\underline{\mathbf{f}}}_m) &= \|P_m \underline{\mathbf{f}}^* - \underline{\mathbf{f}}^*\|^2 + \sigma^2 (m+1). \end{aligned}$$

Décomposition biais/variance d'un estimateur

Exemple de la régression polynomiale

Biais / Variance





Décomposition biais/variance d'un estimateur

Exemple de la régression polynomiale

Risque quadratique

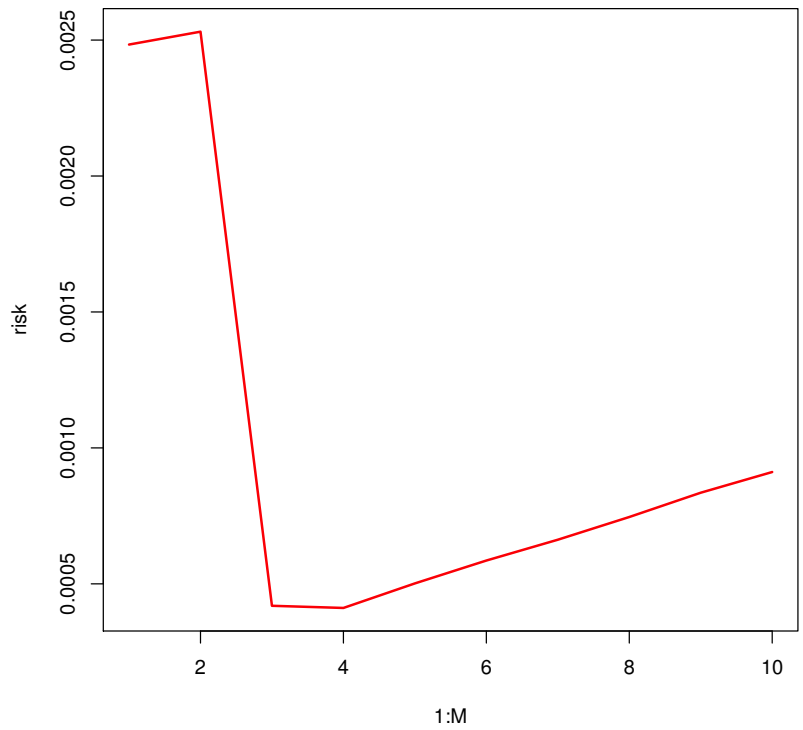
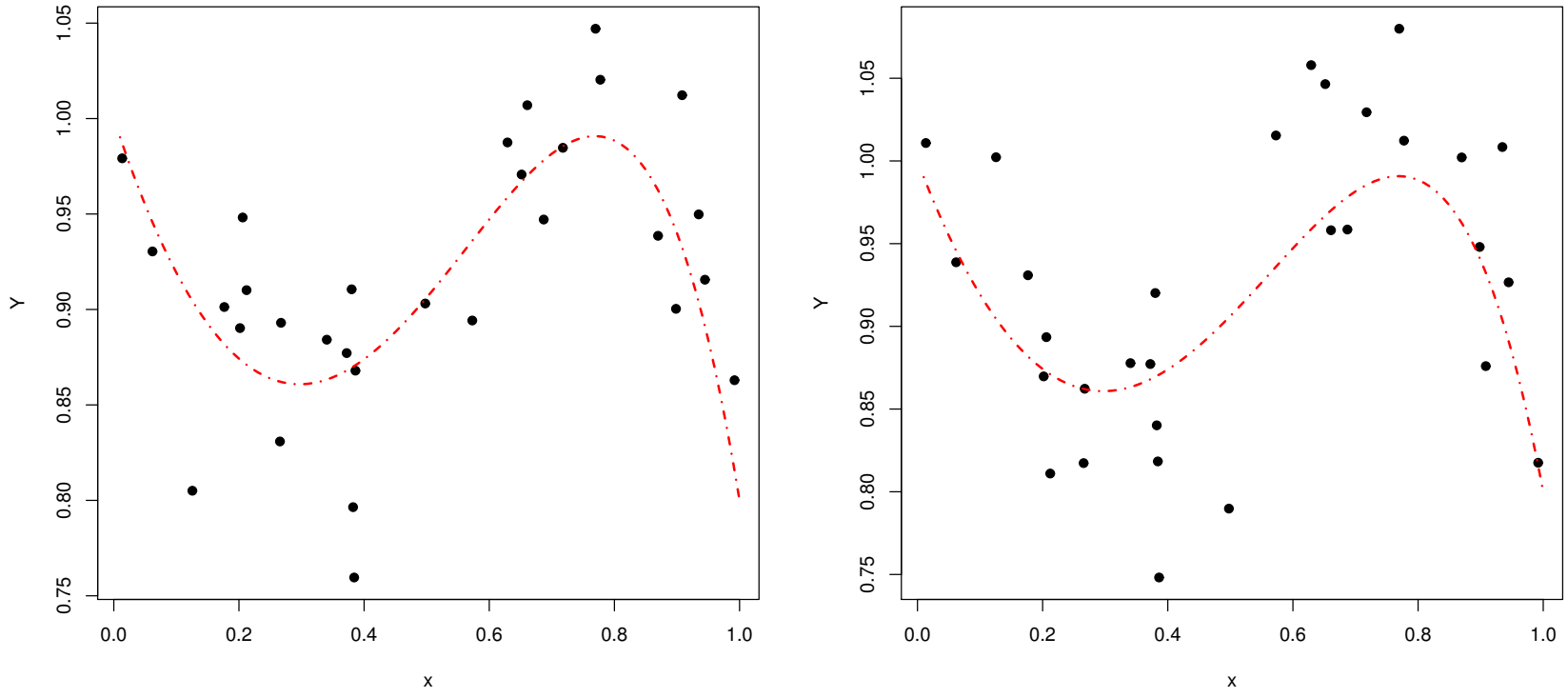


Illustration du compromis biais/variance

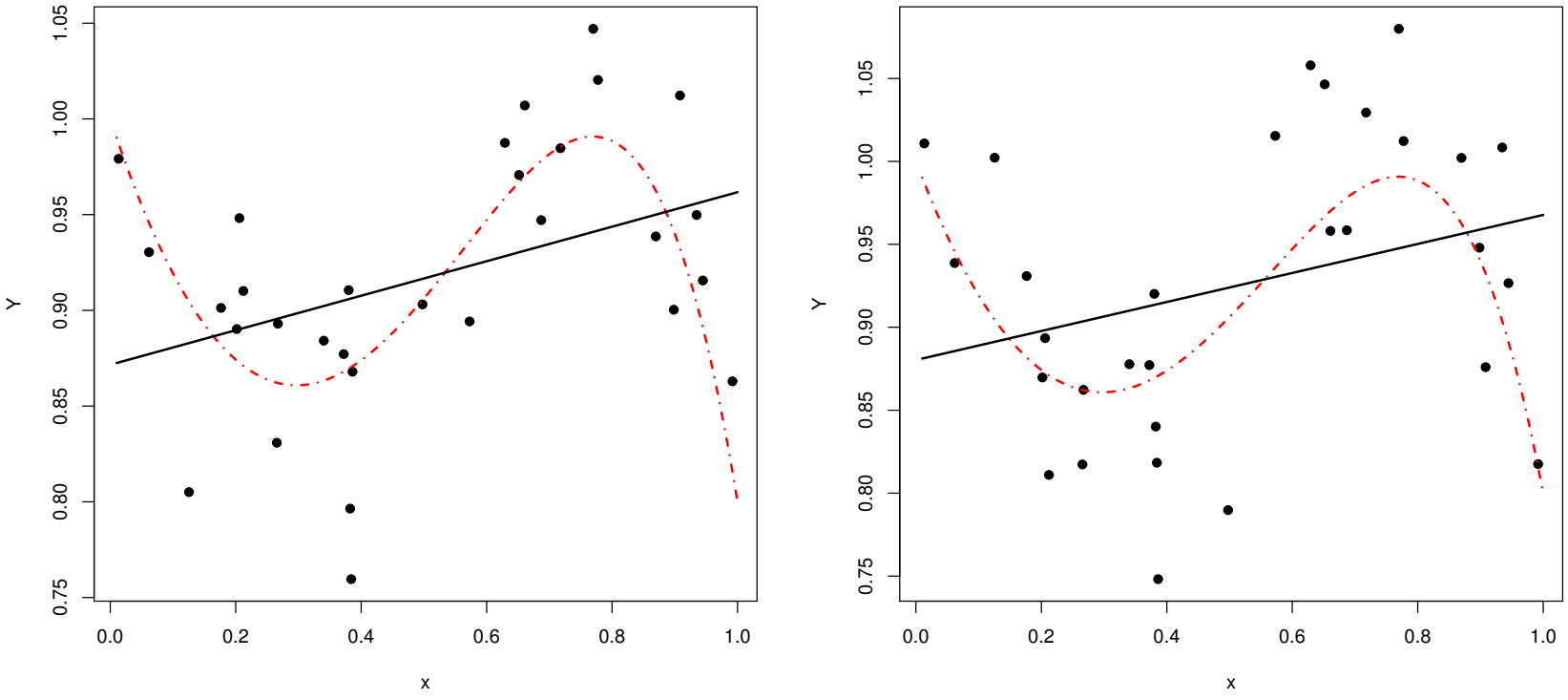
Deux jeux de données (variabilité de l'erreur)



$n = 30$ observations et polynôme de degré $m^* = 5$ (trait pointillé)

Illustration du compromis biais/variance

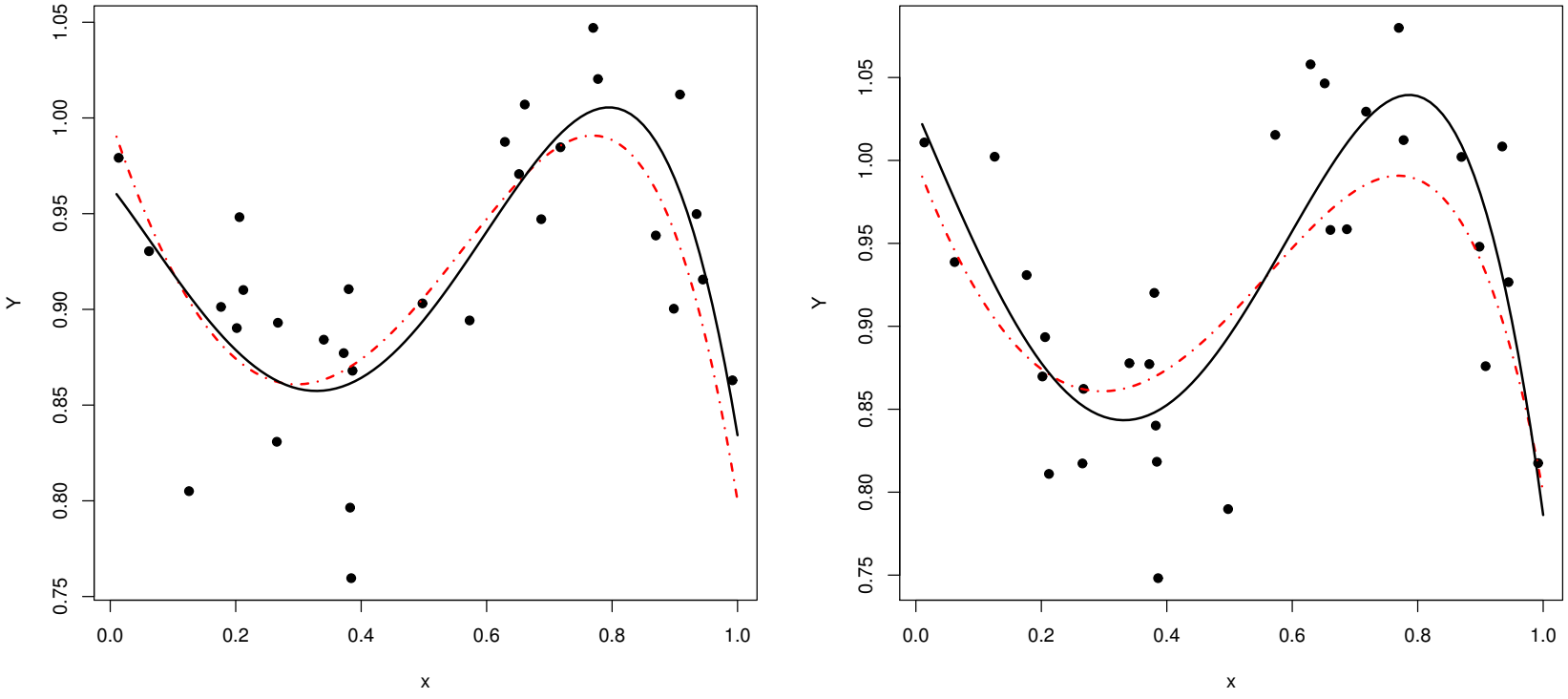
Deux jeux de données (variabilité de l'erreur)



Estimation par un polynôme de degré $m = 1$

Illustration du compromis biais/variance

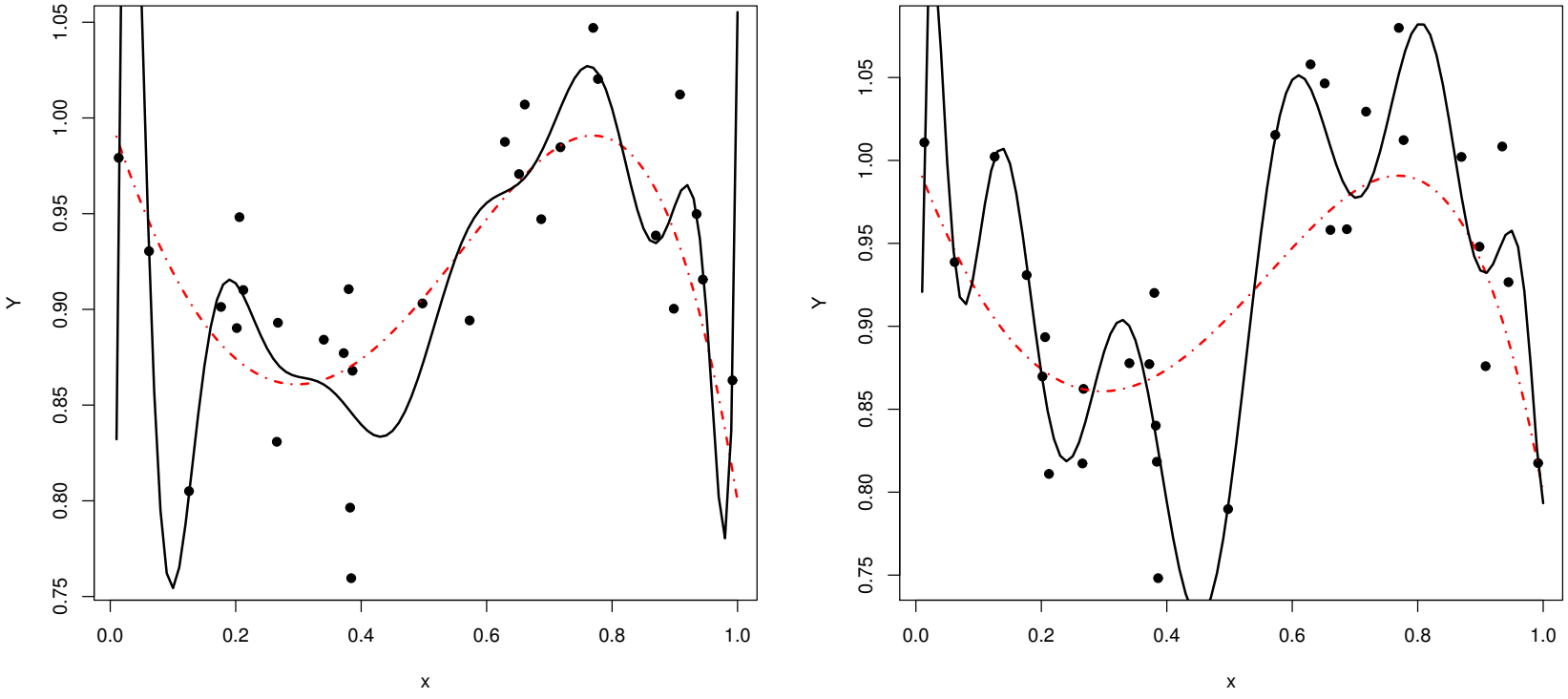
Deux jeux de données (variabilité de l'erreur)



Estimation par un polynôme de degré $m = 4$

Illustration du compromis biais/variance

Deux jeux de données (variabilité de l'erreur)



Estimation par un polynôme de degré $m = 15$

1 Un problème d'apprentissage en finance

2 Principes de base en apprentissage statistique
- Compromis bias/variance en apprentissage
- Principe de la pénalisation/régularisation

Choix d'un estimateur ?

Base d'apprentissage $(\mathbf{x}_i, Y_i)_{1\leq i\leq n}$ où $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$ (**régression**)

Risque quadratique : $R(\hat{f}, f^*) = \sum_{i=1}^{n} \mathbb{E}\left(f^*(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right)^2$

Sélection d'un estimateur : choisir $\hat{f}$ qui minimise $R(\hat{f}, f^*)$ parmi $\hat{f} \in \mathcal{C}$.

Question : comment estimer $R(\hat{f}, f^*)$ qui est **a priori inconnu** à partir de la base d'apprentissage pour tout $\hat{f} \in \mathcal{C}$?

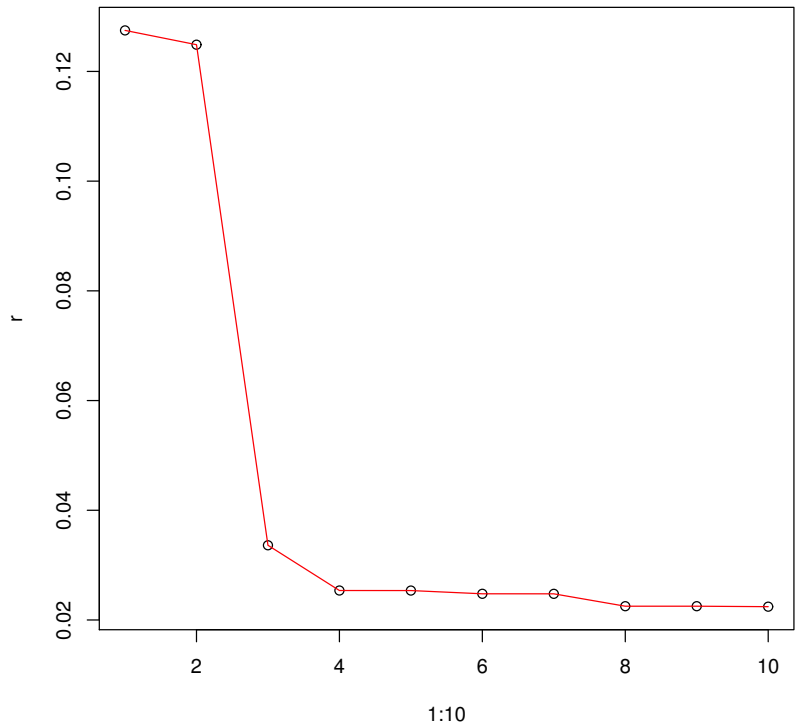
Minimisation de l'erreur des moindres carrés (risque empirique, attache aux données) ?

$$\min_{\hat{f}\in\mathcal{C}} \underbrace{\left\{\sum_{i=1}^{n}\left(Y_i - \hat{f}(\mathbf{x}_i)\right)^2\right\}}_{\text{Bon estimateur de } R(\hat{f}, f^*) \text{ ?}}$$

Choix d'un estimateur ?

Exemple de la régression polynomiale

Erreur des moindres carrés en fonction du degré m



$$m \mapsto \sum_{i=1}^{n} \left(Y_i - \hat{f}_m(\mathbf{x}_i) \right)^2 : \textbf{ choix de } m = 10 \textbf{ !}$$

Choix d'un estimateur ?

ATTENTION - **La minimisation du risque empirique conduit toujours à sélectionner l'estimateur le plus complexe !**

Phénomène de SUR-APPRENTISSAGE (Overfitting) - **On ne peut pas évaluer la qualité/le risque d'un estimateur en cherchant à mininimiser l'attache aux données !**

Explication : $\hat{R}_{emp}(\hat{f}) = \sum_{i=1}^{n} \left(Y_i - \hat{f}(\mathbf{x}_i) \right)^2$ **n'est pas un bon estimateur** de $R(\hat{f}, f^*) = \sum_{i=1}^{n} \mathbb{E} \left(f^*(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i) \right)^2$

Estimation du risque quadratique d'un estimateur

Dans l'exemple de la régression polynomiale, on a que

$$\mathbb{E}\left(\underbrace{\sum_{i=1}^{n}\left(Y_i-\hat{f}_m(\mathbf{x}_i)\right)^2}_{\hat{R}_{emp}(\hat{f}_m)}\right)=R(\hat{f}_m,f^*)-\sigma^2\left(2(m+1)-n\right),$$

et donc $\hat{R}_{emp}(\hat{f}_m)$ **est un estimateur biaisé** de $R(\hat{f}_m,f^*)$!

Un estimateur **non-biaisé** de $R(\hat{f}_m,f^*)$ est donné par

$$\hat{R}_n(\hat{f}_m)=\sum_{i=1}^{n}\left(Y_i-\hat{f}_m(\mathbf{x}_i)\right)^2+2\sigma^2(m+1)-n\sigma^2$$

i.e. qui vérifie $\mathbb{E}(\hat{R}_n(\hat{f}_m))=R(\hat{f}_m,f^*)$.

Preuve

En utilisant le théorème de Pythagore dans l'espace V_m, on obtient que

$$\mathbb{E}\|\mathbf{Y} - \hat{\underline{\mathbf{f}}}_m\|^2 = \mathbb{E}\|\mathbf{Y} - P_m\underline{\mathbf{f}}^*\|^2 - \mathbb{E}\|P_m\underline{\mathbf{f}}^* - \hat{\underline{\mathbf{f}}}_m\|^2$$

Ce qui implique que

$$\begin{aligned}
\mathbb{E}\|\mathbf{Y} - \hat{\underline{\mathbf{f}}}_m\|^2 &= \mathbb{E}\|\mathbf{Y} - P_m\underline{\mathbf{f}}^*\|^2 - \mathbb{E}\|P_m\left(\underline{\mathbf{f}}^* - \mathbf{Y}\right)\|^2 \\
&= \mathbb{E}\|\mathbf{Y} - \underline{\mathbf{f}}^*\|^2 + \|\underline{\mathbf{f}}^* - P_m\underline{\mathbf{f}}^*\|^2 - \sigma^2(m+1) \\
&= \|\underline{\mathbf{f}}^* - P_m\underline{\mathbf{f}}^*\|^2 + \sigma^2(n-m-1) \\
&= \underbrace{\|P_m\underline{\mathbf{f}}^* - \underline{\mathbf{f}}^*\|^2 + \sigma^2(m+1)}_{R(\hat{f}_m, f^*)} - \sigma^2(2(m+1)-n).
\end{aligned}$$

Principe de pénalisation du risque empirique

Choix du degré en régression polynomiale :

Critère du C_p de Mallows

$$\hat{m} = \underset{1 \leq m \leq M}{\arg\min} \left\{ \underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{f}_m(\mathbf{x}_i) \right)^2}_{\text{attache aux données}} + \underbrace{2\sigma^2(m+1)}_{\text{terme de pénalisation}} \right\}$$

- Propriété importante du terme d'attache aux données :

  il est décroissant avec m

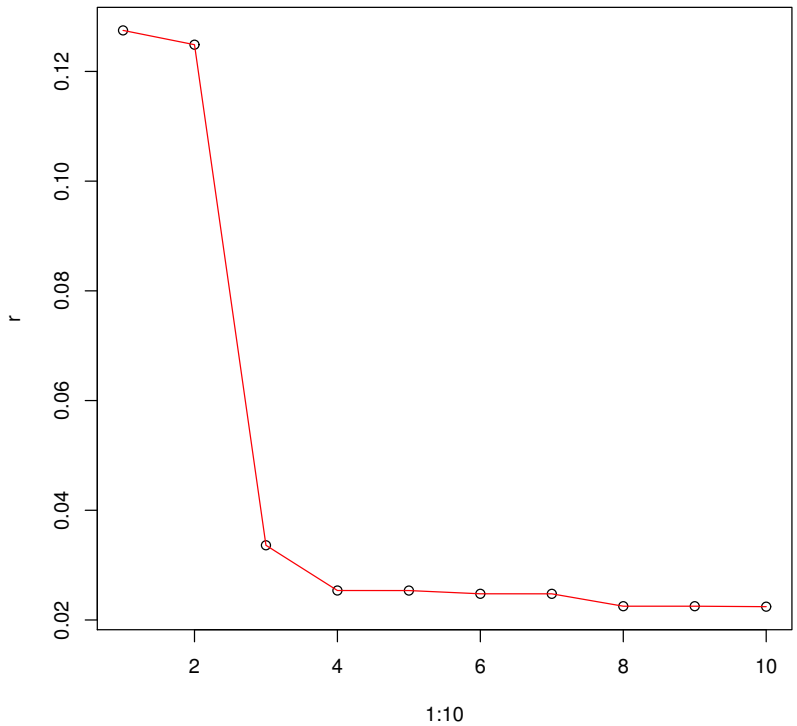
- Propriété importante du terme de pénalisation :

  il est croissant avec m

Choix du degré par le C_p de Mallows

Exemple de la régression polynomiale

Erreur des moindres carrés en fonction du degré m

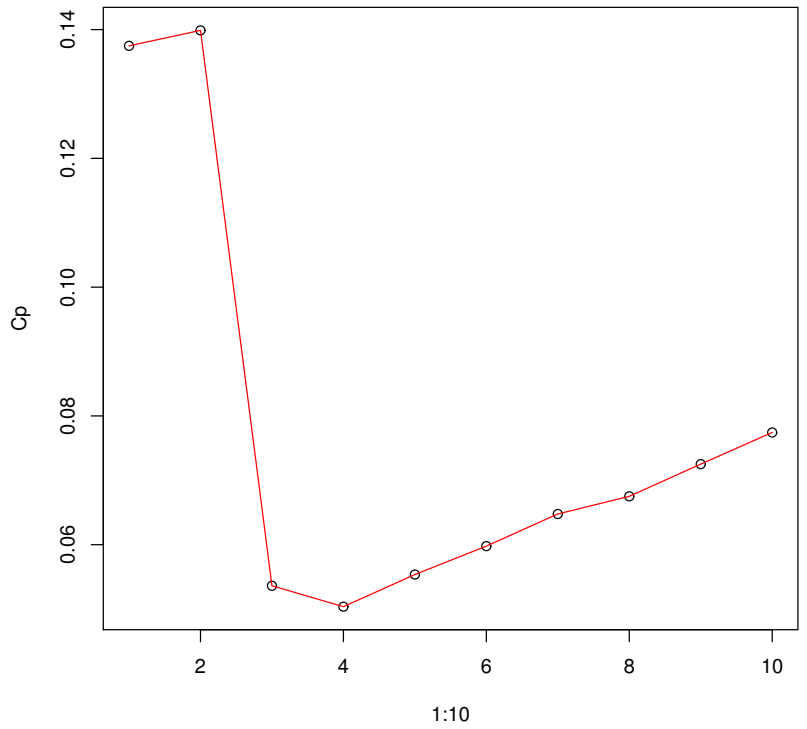


$$m \mapsto \sum_{i=1}^{n} \left(Y_i - \hat{f}_m(\mathbf{x}_i) \right)^2 : \textbf{ choix de } m = 10 \textbf{ !}$$

Choix du degré par le C_p de Mallows

Exemple de la régression polynomiale

Erreur des moindres carrés **penalisée** en fonction du degré m



$$m \mapsto \sum_{i=1}^{n} \left(Y_i - \hat{f}_m(\mathbf{x}_i) \right)^2 + 2\sigma^2(m+1) : \textbf{ choix de } \hat{m} = 4 \textbf{ !}$$

Principe de pénalisation du risque empirique

Choix d'un estimateur en régression

$$Y_i = f^*(X_i) + \epsilon_i,\ i = 1, \ldots, n,$$

$$\bar{f} = \underset{\hat{f} \in \mathcal{C}}{\arg\min} \left\{ \underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i) \right)^2}_{\text{attache aux données}} + \text{Pen}(\hat{f}) \right\}$$

- Cadre des modèles paramétriques :

$$\text{Pen}(\hat{f}) = 2 \times \sigma^2 \times \text{ nombre de paramètres de } \hat{f}$$

- Cadre des modèles non-paramétriques :

$$\text{Pen}(\hat{f}) = \lambda \times \|\hat{f}\|^2$$

où $\lambda > 0$ est un **paramètre de régularisation** et $\|\cdot\|$ une semi-norme sur l'espace $\mathcal{F}$ qui mesure la complexité de $\hat{f}$