Régression spline & GAM

Vincent Lefieux

Plan

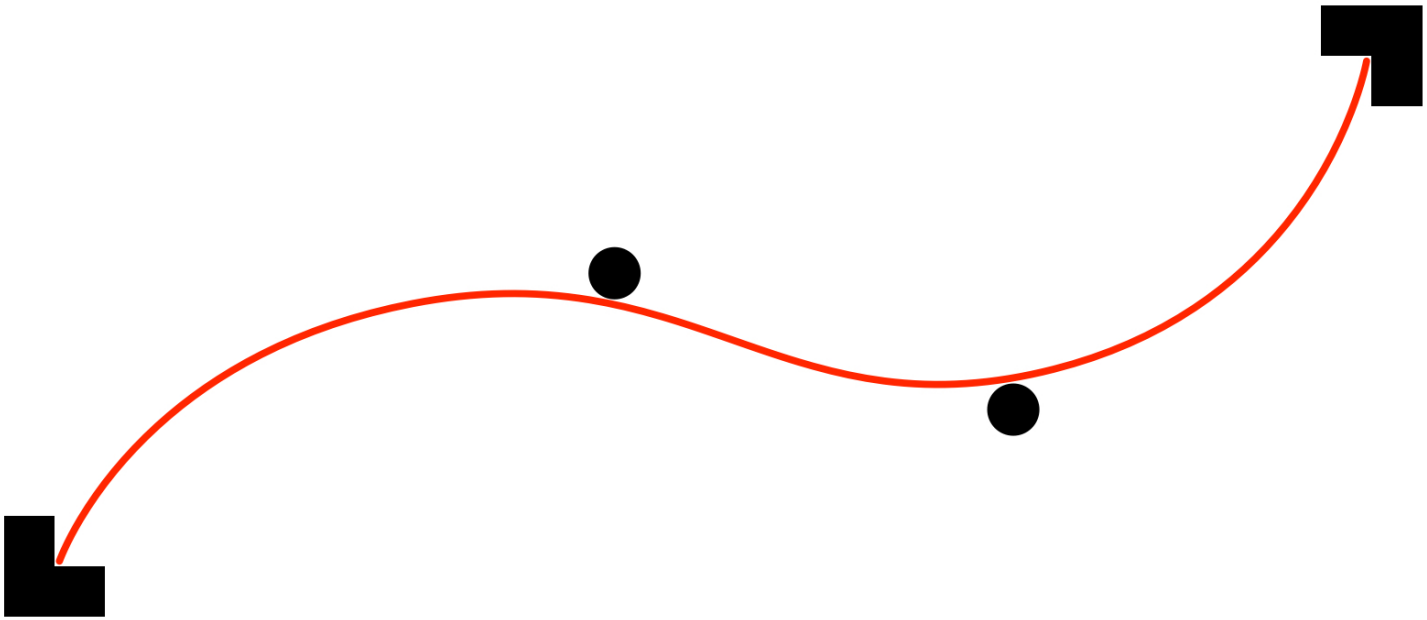
Plan

Splines

Point de vue physique I

- Le terme *spline* (cerce en français) désigne une latte en bois flexible utilisée par les dessinateurs industriels pour matérialiser des lignes à courbure variable passant par des points fixés à priori.
- L'enjeu était d'obtenir des courbes « lisses », d'éventuelles discontinuités pouvant être synonymes de ruptures potentielles à cause d'une faiblesse mécanique.
- Le tracé de la spline minimise l'énergie de déformation de la latte considérée.

Point de vue physique II

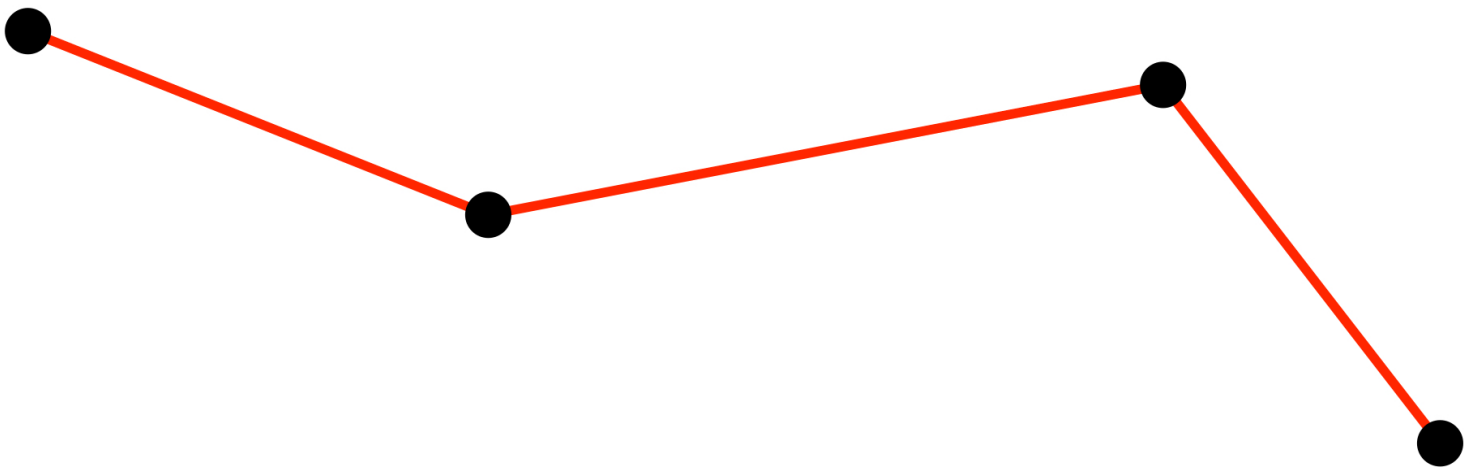


Point de vue mathématique

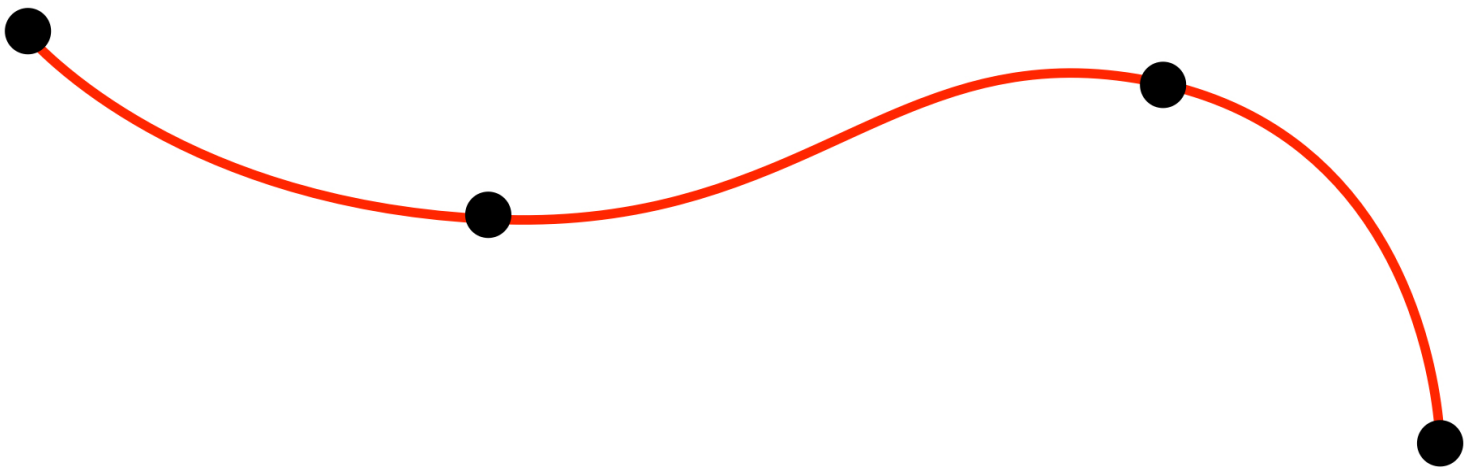
- Une spline est une fonction définie par morceaux par des polynômes.

Interpolation : exemple I

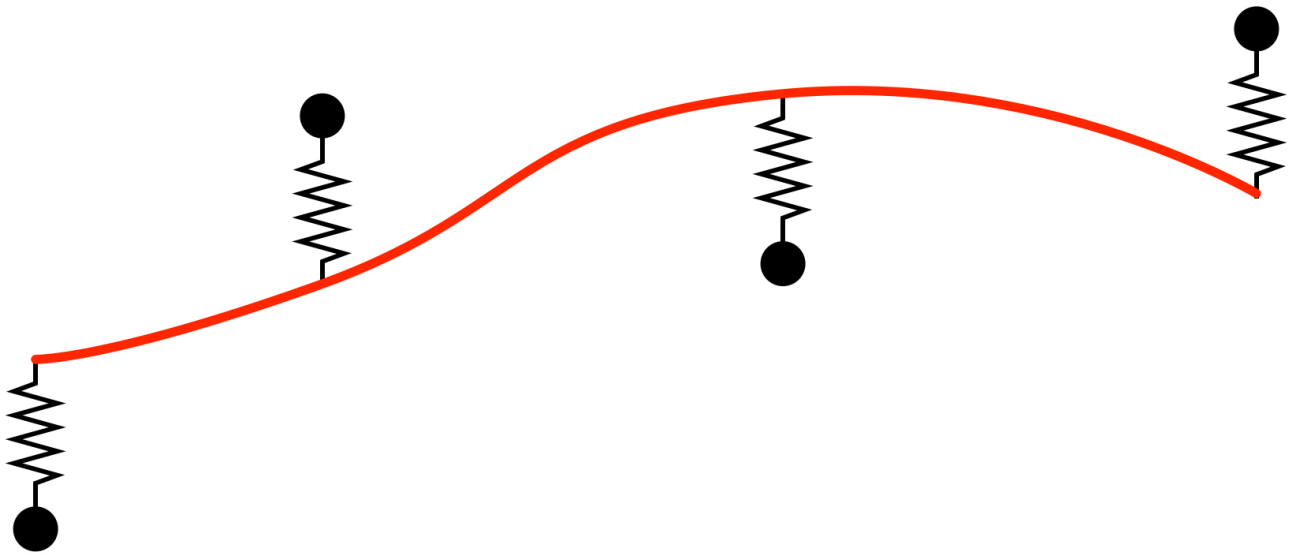
Interpolation : exemple II



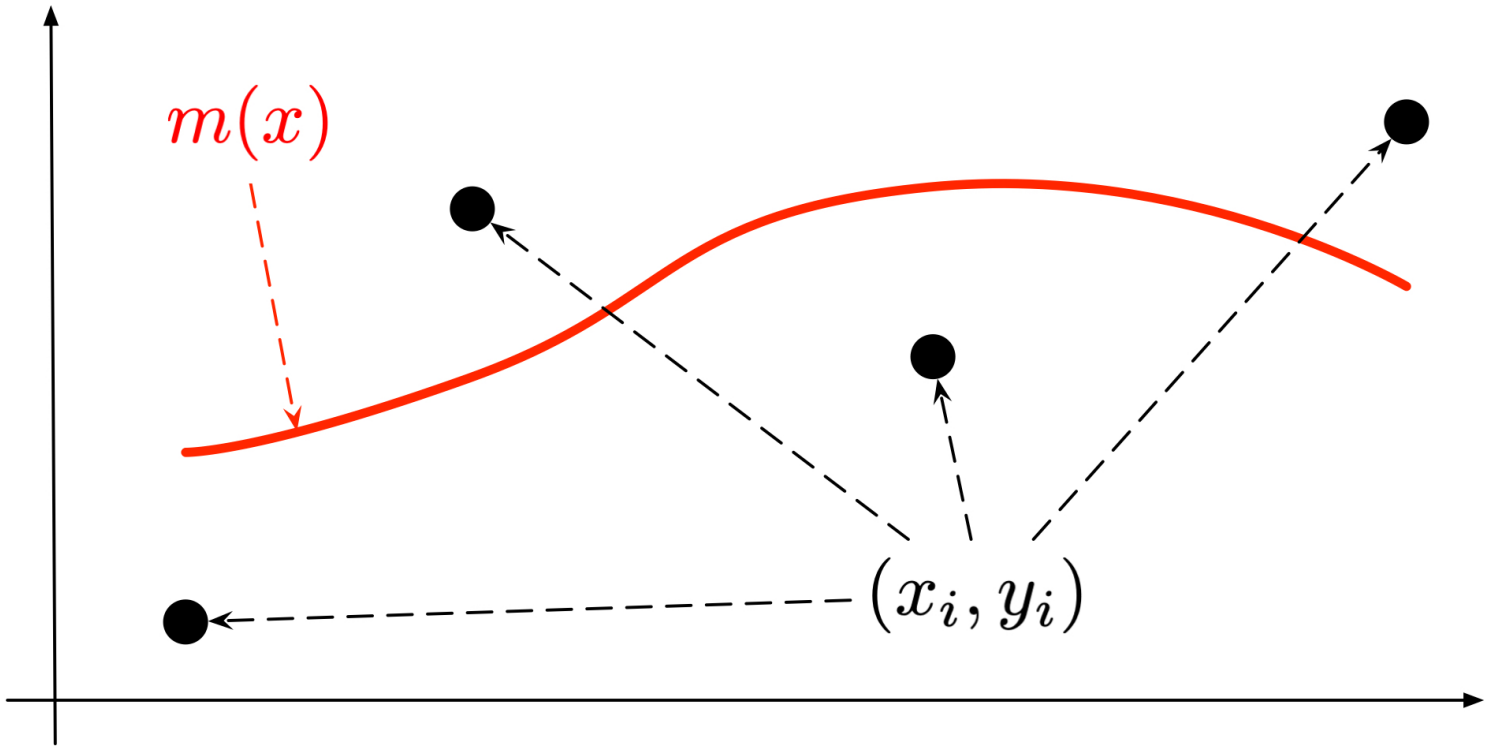
Interpolation : exemple III



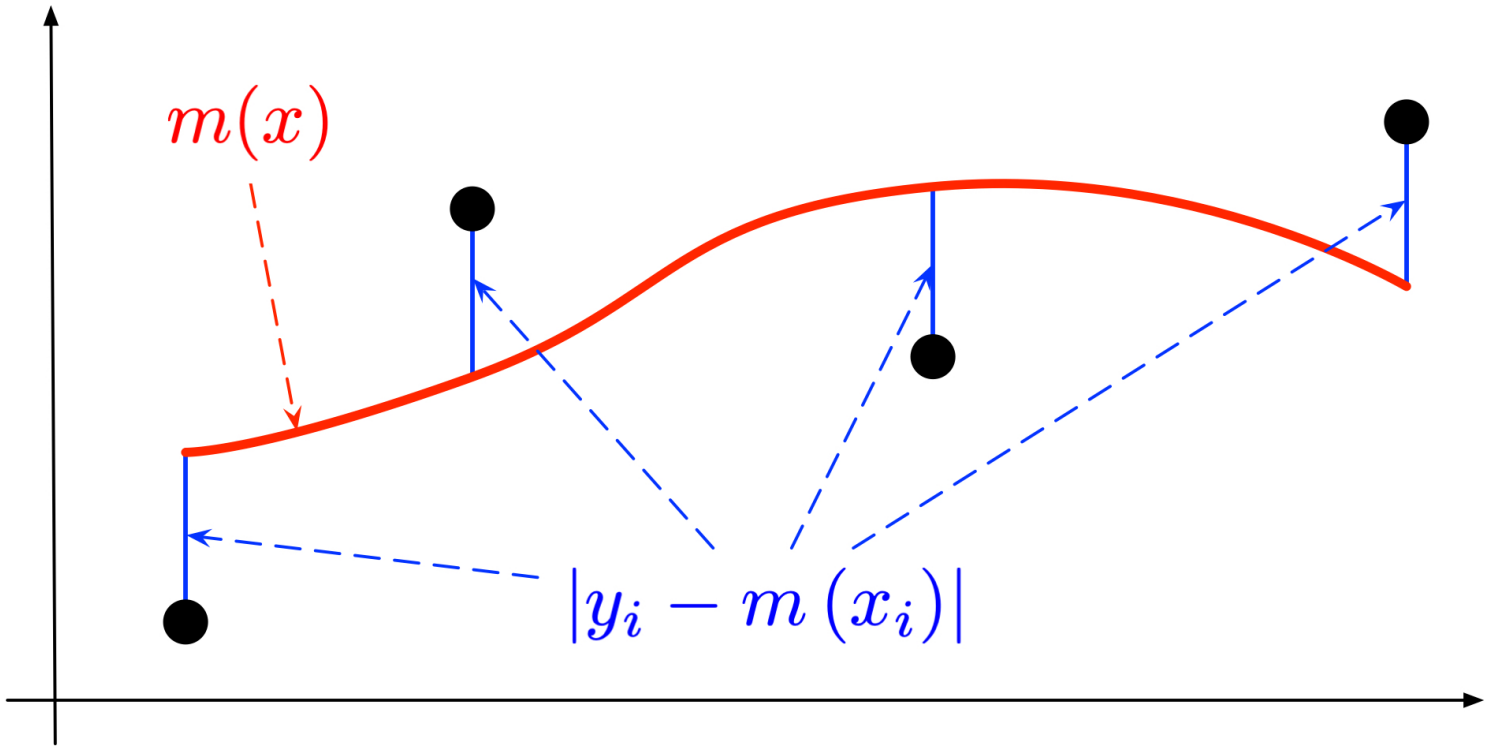
Lissage : point de vue physique I



Lissage : point de vue physique II



Lissage : point de vue physique III



Lissage : point de vue physique IV

- La latte, liée aux points via des ressorts, correspond à une courbe lissée.
- Sa forme est celle qui minimise l'énergie totale de déformation (allongement des ressorts et courbure de la latte) :

$$E(m, \lambda) = \sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int \left[m^{(2)}(x) \right]^2 \mathrm{d}x$$

où $(x_i, y_i)_{i \in \{1,\dots,n\}}$ est l'ensemble des points, m l'équation de la forme de la latte (avec $m^{(2)}$ comme dérivée seconde) et λ le rapport entre la raideur de la latte et celle des ressorts.

Données considérées dans le cas univarié

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\dots,n\}}$$

où $X \in \mathbb{R}$ et $Y \in \mathbb{R}$.

- On note :

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}} \ .$$

Splines d'interpolation, de moindres carrés et de lissage

- Pour l'interpolation, on utilise des splines d'interpolation pour ajuster m telle que :

$$\forall i \in \{1, \ldots, n\} : y_i = m(x_i) \ .$$

- Pour le modèle de régression :

$$\forall i \in \{1, \ldots, n\} : y_i = m(x_i) + \varepsilon_i \ ,$$

on utilise :

  - Des splines de moindres carrés qui minimisent :

$$\sum_{i=1}^{n} (y_i - m(x_i))^2 \ .$$

  - Des splines de lissage qui minimisent :

$$\sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int \left[m^{(2)}(x) \right]^2 \mathrm{d}x \ .$$

Données considérées dans le cas multivarié

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\dots,n\}}$$

où $X = \left(X^1, \dots, X^p\right)^\top \in \mathbb{R}^p$ et $Y \in \mathbb{R}$.

- On note :

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}} \ .$$

Cas de la régression multivariée

- Le modèle MARS (Multivariate Adaptive Regression Splines) : (Friedman, 1991).
- Le modèle GAM (Generalized Additive Models) : (Hastie et Tibshirani, 1986).

Splines polynomiales d'ordre d

- On considère K nœuds (*knots*) : $(\xi_1, \ldots, \xi_K)$ sur $[a, b]$:

$$a < \xi_1 < \ldots < \xi_K < b \ .$$

- Une spline polynomiale d'ordre $d \in \mathbb{N}^\star$ est une fonction :
  - continûment différentiable jusqu'à l'ordre :
    - $d - 2$ si d>1,
    - 0 si d=1 (simplement continue),
  - constituée de polynômes de degré (inférieur ou égal à) $(d-1)$ sur les intervalles inter-noeuds $[a, \xi_1]$, $[\xi_1, \xi_2], \ldots, [\xi_{K-1}, \xi_K]$, $[\xi_K, b]$.
- On parle de :
  - spline linéaire si $d = 2$,
  - spline cubique si $d = 4$.

Splines linéaires

- On considère K nœuds (*knots*) : $(\xi_1, \ldots, \xi_K)$ sur $[a, b]$:

$$a < \xi_1 < \ldots < \xi_K < b \ .$$

- Une spline linéaire (spline d'ordre 2) est une fonction :
  - continue,
  - constituée de droites sur les intervalles inter-noeuds $[a, \xi_1]$, $[\xi_1, \xi_2]$,. . . , $[\xi_{K-1}, \xi_K]$, $[\xi_K, b]$.

Splines cubiques

- On considère K nœuds (*knots*) : $(\xi_1, \ldots, \xi_K)$ sur $[a, b]$:

$$a < \xi_1 < \ldots < \xi_K < b \ .$$

- Une spline cubique (spline d'ordre 4) est une fonction :
  - continûment différentiable jusqu'à l'ordre 2,
  - constituée de polynômes de degré (inférieur ou égal à) 3 sur les intervalles inter-noeuds $[a, \xi_1]$, $[\xi_1, \xi_2]$,. . . , $[\xi_{K-1}, \xi_K]$, $[\xi_K, b]$.

Espace des splines polynomiales d'ordre d

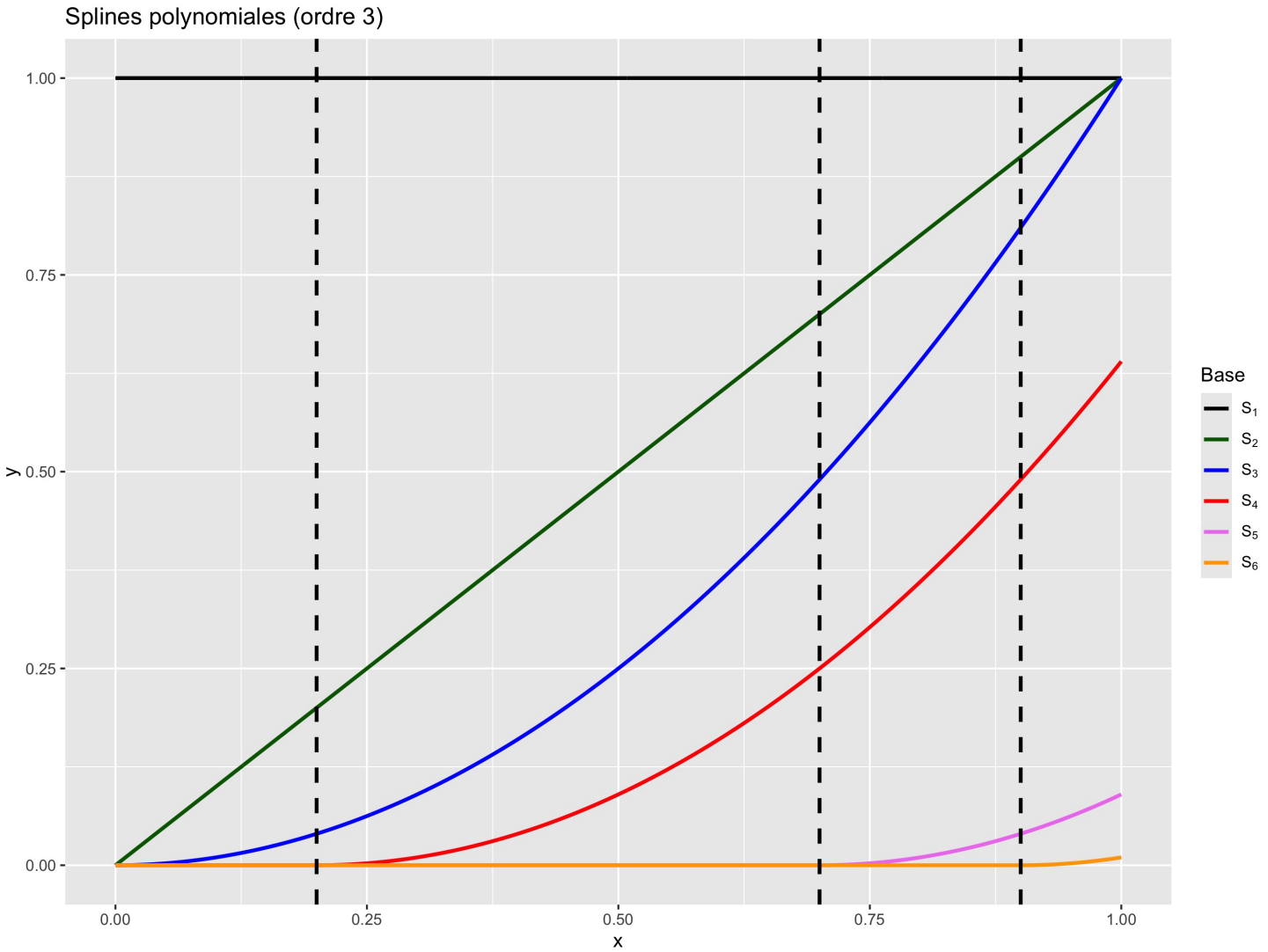
- On note $\mathcal{S}_d(\xi_1, \ldots, \xi_K)$ l'ensemble des splines polynomiales d'ordre d ayant pour nœuds $(\xi_1, \ldots, \xi_K)$.
- $\mathcal{S}_d(\xi_1, \ldots, \xi_K)$ est un sous-espace vectoriel de l'espace des fonctions dérivables jusqu'à l'ordre $(d-2)$ (si $d > 1$, 0 sinon), de dimension $d + K$.
- On peut considérer comme base :

$$S_1(x) = 1 \ ,$$
$$\vdots$$
$$S_d(x) = x^{d-1} \ ,$$
$$\forall k \in \{1, \ldots, K\} : S_{d+k}(x) = \left[(x - \xi_k)_+\right]^{d-1}$$

où :

$$x_+ = \max(x, 0) = \begin{cases} x & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases} \ .$$

Illustration



Remarques

- On trouve également les splines naturelles d'ordre pair : elles diffèrent des splines d'ordre d au niveau des intervalles $[a, \xi_1]$ et $[\xi_K, b]$ sur lesquels elles coincident avec un polynôme de degré $\frac{d}{2} - 1$.
  Des splines cubiques naturelles coincident donc avec des droites sur le premier et le dernier intervalle.
- La base $(S_1, \ldots, S_{d+K})$ de $\mathcal{S}_d(\xi_1, \ldots, \xi_K)$, définie précédemment, est simple d'un point de vue conceptuel mais est peu utilisée en pratique, à cause du support non compact des fonctions de la base, et des problèmes d'arrondis pouvant apparaitre pour de grandes valeurs de x.
- On lui préfère très souvent la base B-splines dans laquelle chaque fonction de base a un support fini.
  Les B-splines sont une généralisation des courbes de Bézier, et ont été généralisées par les NURBS (Non-Uniform Rational Basis Splines).

B-splines

- On considère K nœuds (*knots*) : $(\xi_1, \ldots, \xi_K)$ sur $[a, b] := [\xi_0, \xi_{K+1}]$:

$$a = \xi_0 \leq \xi_1 \leq \ldots \leq \xi_K \leq b = \xi_{K+1} \, .$$

Quand les nœuds sont équidistants, on parle de B-splines uniformes.

- On définit de manière récursive la base de B-splines d'ordre $d \in \mathbb{N}^\star$ avec $d \leq K$:

1. Pour $i \in \{1, \ldots, K-1\}$:

$$B_{i,1}(x) = \begin{cases} 1 & \text{si } x \in [\xi_i, \xi_{i+1}[\\ 0 & \text{sinon} \end{cases} .$$

Si $\xi_i = \xi_{i+1}$, on pose par convention $B_{i,1} = 0$.

2. Pour $i \in \{1, \ldots, K-d\}$:

$$B_{i,d}(x) = \frac{x - \xi_i}{\xi_{i+d-1} - \xi_i} B_{i,d-1}(x) + \frac{\xi_{i+d} - x}{\xi_{i+d} - \xi_{i+1}} B_{i+1,d-1}(x) \, .$$

Par convention, une fraction dont le dénominateur est nul, est considérée nulle.

Illustration

Splines naturelles d'ordre d

- On considère K nœuds (*knots*) : $(\xi_1, \ldots, \xi_K)$ sur $[a, b]$:

$$a < \xi_1 < \ldots < \xi_K < b \ .$$

- Une spline naturelle d'ordre $d \in \mathbb{N}^\star$ pair est une fonction :
  - continûment différentiable jusqu'à l'ordre $d-2$ $(d \geq 2)$,
  - constituée de polynômes de degré :
    - $\frac{d}{2} - 1$ sur les intervalles $[a, \xi_1]$ et $[\xi_K, b]$,
    - (inférieur ou égal à) $(d-1)$ sur les intervalles inter-noeuds $[\xi_1, \xi_2], \ldots, [\xi_{K-1}, \xi_K]$.
- On parle de spline cubique naturelle si $d = 4$.

Splines cubiques naturelles

- On considère K nœuds (*knots*) : $(\xi_1, \ldots, \xi_K)$ sur $[a, b]$:

$$a < \xi_1 < \ldots < \xi_K < b \ .$$

- Une spline cubique naturelle est une fonction :
  - continûment différentiable jusqu'à l'ordre 2,
  - constituée de polynômes de degré :
    - 1 sur les intervalles $[a, \xi_1]$ et $[\xi_K, b]$,
    - (inférieur ou égal à) 3 sur les intervalles inter-noeuds $[\xi_1, \xi_2], \ldots, [\xi_{K-1}, \xi_K]$.

Espace des splines naturelles d'ordre d

- On note $\mathcal{S}_d^\star(\xi_1, \ldots, \xi_K)$ l'ensemble des splines naturelles d'ordre d ayant pour nœuds $(\xi_1, \ldots, \xi_K)$.
- $\mathcal{S}_d(\xi_1, \ldots, \xi_K)$ est un sous-espace vectoriel de l'espace des fonctions dérivables jusqu'à l'ordre $(d-2)$, de dimension K.

Plan

Interpolation spline

Concepts généraux

- On dit qu'une fonction f est absolument continue s'il existe $a \in \mathbb{R}$ et une fonction g intégrable tels que :

$$\forall x \in \mathbb{R} : f(x) = \int_a^x g(y)\, dy \ .$$

- Soit $W^d(a, b)$ l'ensemble des fonctions f définies sur $[a, b]$ telles que :
  - $\left(f^{(0)}, \ldots, f^{(d-1)}\right)$ sont absolument continues et de carré intégrable,
  - $f^{(d)}$ est de carré intégrable.

- La régularité d'une fonction $f \in W^d(a, b)$ peut être mesurée par :

$$\int_a^b \left[f^{(d)}(x)\right]^2 \mathrm{d}x \ .$$

Splines d'interpolation I

- On dispose d'un échantillon de $(X, Y) \in \mathbb{R} \times \mathbb{R}$:

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}}$$

tel que les $(x_i)_{i \in \{1,\dots,n\}}$ sont distincts.

- Il existe une unique fonction $\widehat{m} \in W^d(a, b)$ vérifiant :

$$\forall i \in \{1, \dots, n\} : y_i = \widehat{m}(x_i) \ ,$$

$$\widehat{m} = \arg \min_{m \in W^d(a,b)} \int_a^b \left[m^{(d)}(x) \right]^2 \mathrm{d}x \ .$$

- La fonction $\widehat{m}$ est une spline naturelle d'ordre $2d$ ayant pour nœuds $(x_1, \dots, x_n)$.

Splines d'interpolation II

- Il est possible de relaxer les conditions d'interpolation, notamment dans le cas où les $(x_i)_{i\in\{1,\dots,n\}}$ ne sont pas distincts, par :

$$\sum_{i=1}^{n} (y_i - m(x_i))^2 \leq \varepsilon$$

où $\varepsilon \in \mathbb{R}^\star$ donne le niveau de la relaxation.

- On peut montrer que la fonction $\widehat{m}$ est alors la solution du problème d'optimisation suivant :

$$\widehat{m} = \arg \min_{m\in W^d(a,b)} \sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int_a^b \left[m^{(d)}(x)\right]^2 \mathrm{d}x$$

où λ dépend de ε (de manière complexe).

Plan

Régression spline

Données considérées

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\dots,n\}}$$

où $X \in \mathbb{R}$ et $Y \in \mathbb{R}$.

- On note :

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}} \ .$$

Splines de moindres carrés I

- On appelle spline de moindres carrés d'ordre d ayant comme nœuds $(\xi_1, \ldots, \xi_K)$, la fonction $\widehat{m}$ suivante :

$$\widehat{m} = \arg \min_{m \in \mathcal{S}_d(\xi_1, \ldots, \xi_K)} \sum_{i=1}^{n} \left(y_i - m\left(x_i\right)\right)^2 .$$

Splines de moindres carrés II

- On considère la base $(S_1, \ldots, S_{d+K})$ de $\mathcal{S}_d(\xi_1, \ldots, \xi_K)$.
- On peut écrire :

$$\widehat{m}(x) = \sum_{j=1}^{d+K} \widehat{\theta}_j S_j(x)$$

où $\left(\widehat{\theta}_1, \ldots, \widehat{\theta}_{d+K}\right)^\top \in \mathbb{R}^{d+K}$ minimisent :

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d+K} \theta_j S_j(x_i) \right)^2 .$$

Splines de moindres carrés III

- On note :

$$\mathbf{y} = (y_1, \ldots, y_n)^\top ,$$
$$\widehat{\mathbf{y}} = \left(\widehat{m}(x_1), \ldots, \widehat{m}(x_n)\right)^\top ,$$
$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d+K})^\top ,$$
$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \ldots, \widehat{\theta}_{d+K}\right)^\top ,$$
$$\mathbf{N} = \left[S_j(x_i)\right]_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,d+K\}} .$$

- On cherche $\widehat{\theta}$ qui minimise :

$$(\mathbf{y} - \mathbf{N}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{N}\boldsymbol{\theta}) .$$

- La solution est :

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{N}^\top \mathbf{N}\right)^{-1} \mathbf{N}\mathbf{y} .$$

Splines de moindres carrés IV

- On peut écrire :

$$\widehat{\mathbf{y}} = \mathbf{S}\,\mathbf{y}$$

où :

$$\mathbf{S} = \mathbf{N}\left(\mathbf{N}^\top \mathbf{N}\right)^{-1} \mathbf{N}^\top$$

est la matrice de lissage (*smoothing*).

Splines de moindres carrés V

- Si on considère une base de B-splines, la matrice $N^\top N$ est alors une matrice bande avec $2d-1$ diagonales non nulles.
- On considère usuellement :
  - un ordre $d = 4$,
  - un nombre de noeuds $K \in \{0, \ldots, n-d\}$,
  - des nœuds $(\xi_1, \ldots, \xi_K)$ équirépartis ou égaux à des quantiles empiriques de X.
- A d fixé, le nombre de nœuds K est un hyperparamètre modulant le lissage :
  - $K = 0$: l'estimateur correspond à une régression polynomiale de degré $d-1$.
  - $K = n-d$: l'estimateur correspond à une spline d'interpolation.

Des splines de moindres carrés aux splines de lissage

- On ajoute une pénalité afin de contrôler les variations de l'estimateur (importantes dans le cas des splines de moindres carrés).

Splines de lissage I

- On appelle spline de lissage d'ordre $2d$ ayant comme nœuds $(\xi_1, \ldots, \xi_K)$, la fonction $\widehat{m}$ suivante :

$$\widehat{m} = \arg \min_{m \in W^d(a,b)} \sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int_a^b \left[m^{(d)}(x) \right]^2 \mathrm{d}x$$

où $\lambda \in \mathbb{R}^{+\star}$ caractérise le compromis entre l'ajustement et le caractère lisse de la fonction.

Splines de lissage II

- Si on considère la base $(S_1, \ldots, S_n)$ de $\mathcal{S}_{2d}^{\star}(x_1, \ldots, x_n)$ (splines naturelles), on peut alors écrire :

$$\widehat{m}(x) = \sum_{j=1}^{n} \widehat{\theta}_j \, S_j(x)$$

où $\left(\widehat{\theta}_1, \ldots, \widehat{\theta}_n\right) \in \mathbb{R}^n$ minimisent :

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \theta_j \, S_j(x_i) \right)^2 + \lambda \int_a^b \left(\sum_{j=1}^{n} \theta_j \, S_j^{(d)}(x) \right)^2 \mathrm{d}x \, .$$

Splines de lissage III

- On note :

$$\mathbf{y} = (y_1, \ldots, y_n)^\top ,$$
$$\widehat{\mathbf{y}} = \left(\widehat{m}(x_1), \ldots, \widehat{m}(x_n)\right)^\top ,$$
$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top ,$$
$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \ldots, \widehat{\theta}_n\right)^\top ,$$
$$\mathbf{N} = \left[S_j(x_i)\right]_{i\in\{1,\ldots,n\}, j\in\{1,\ldots,n\}} ,$$
$$\boldsymbol{\Omega} = \left[\int_a^b S_i^{(d)}(x)\, S_j^{(d)}(x)\, \mathrm{d}x\right]_{i\in\{1,\ldots,n\}, j\in\{1,\ldots,n\}} .$$

- On cherche $\widehat{\theta}$ qui minimise :

$$(\mathbf{y} - \mathbf{N}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{N}\boldsymbol{\theta}) + \lambda\, \boldsymbol{\theta}^\top \boldsymbol{\Omega}\, \boldsymbol{\theta} .$$

- La solution est :

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{N}^\top \mathbf{N} + \lambda\, \boldsymbol{\Omega}\right)^{-1} \mathbf{N}\, \mathbf{y} .$$

Splines de lissage IV

- On peut écrire :

$$\widehat{\mathbf{y}} = \mathbf{S}\,\mathbf{y}$$

où :

$$\mathbf{S} = \mathbf{N}\left(\mathbf{N}^\top\mathbf{N} + \lambda\,\mathbf{\Omega}\right)^{-1}\mathbf{N}^\top$$

est la matrice de lissage (*smoothing*).

Illustration I

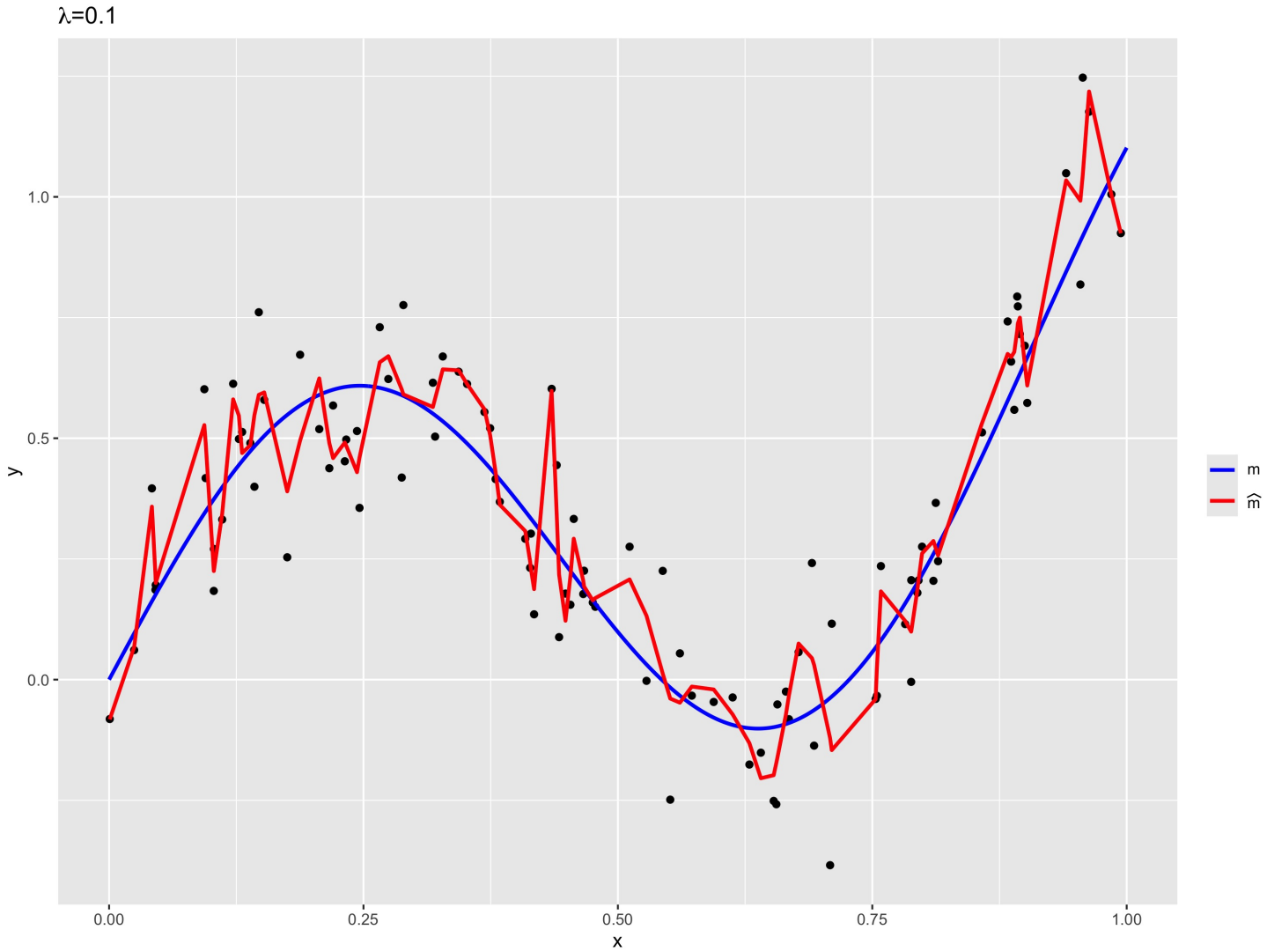


Illustration II

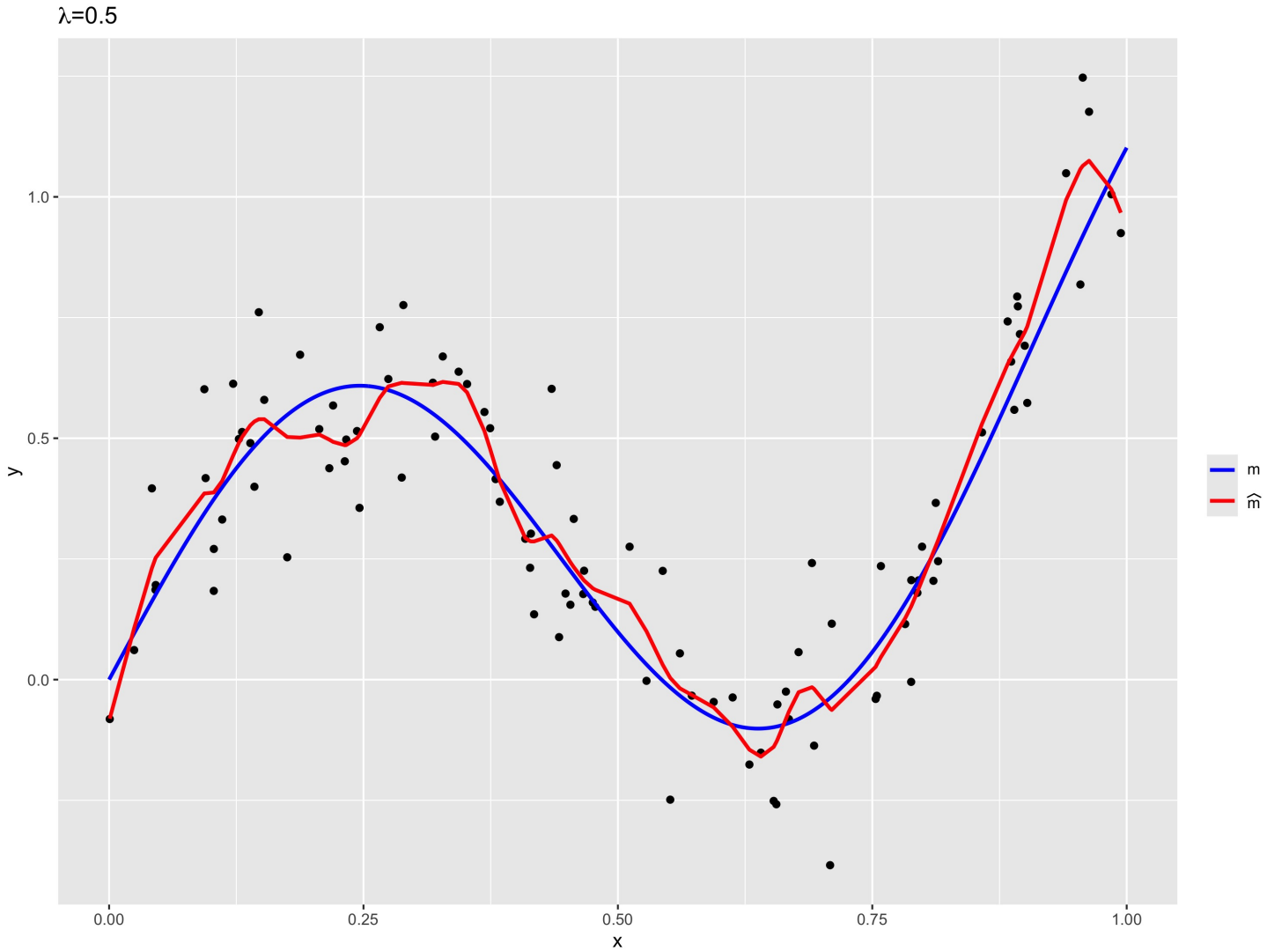


Illustration III

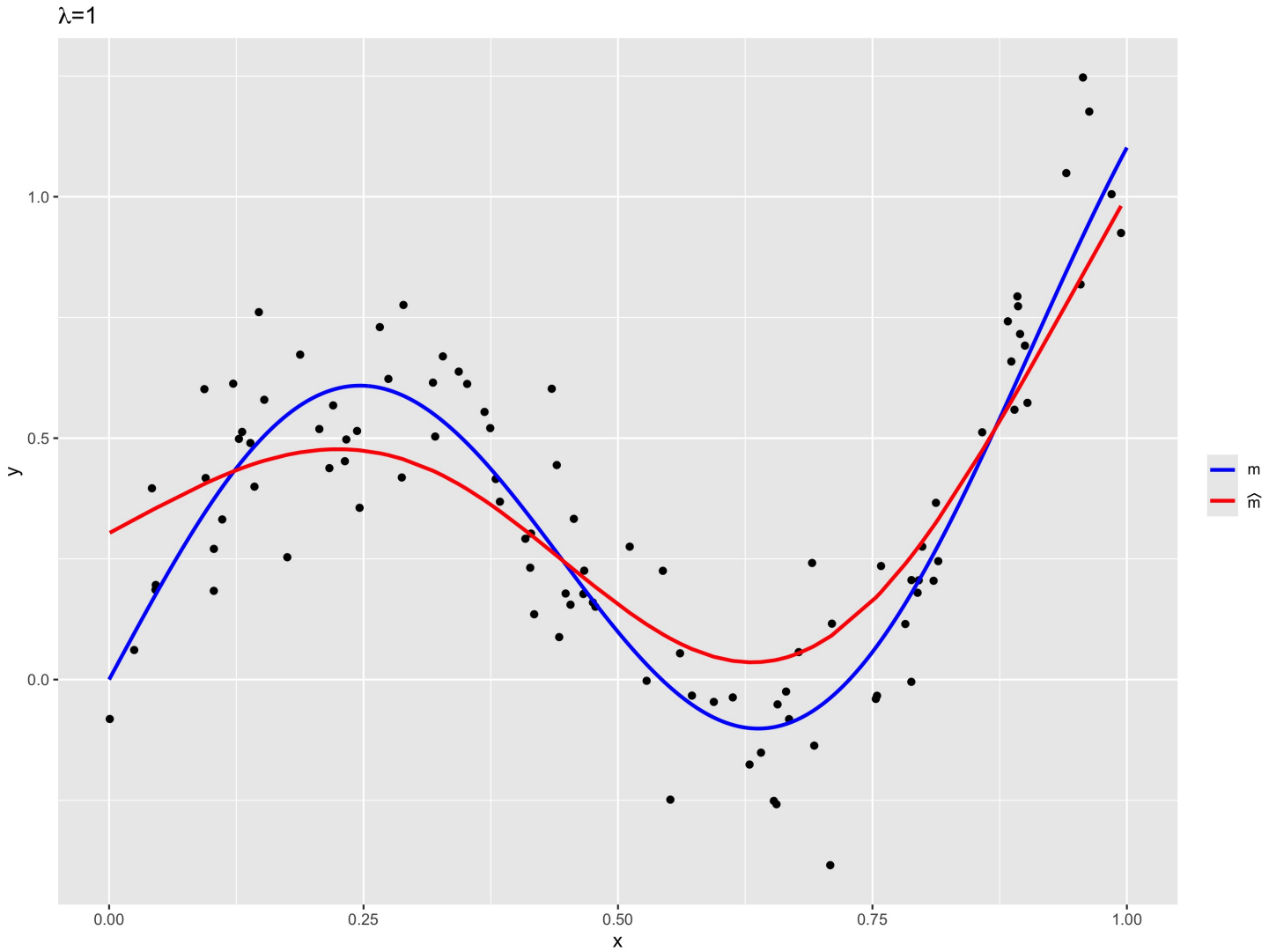
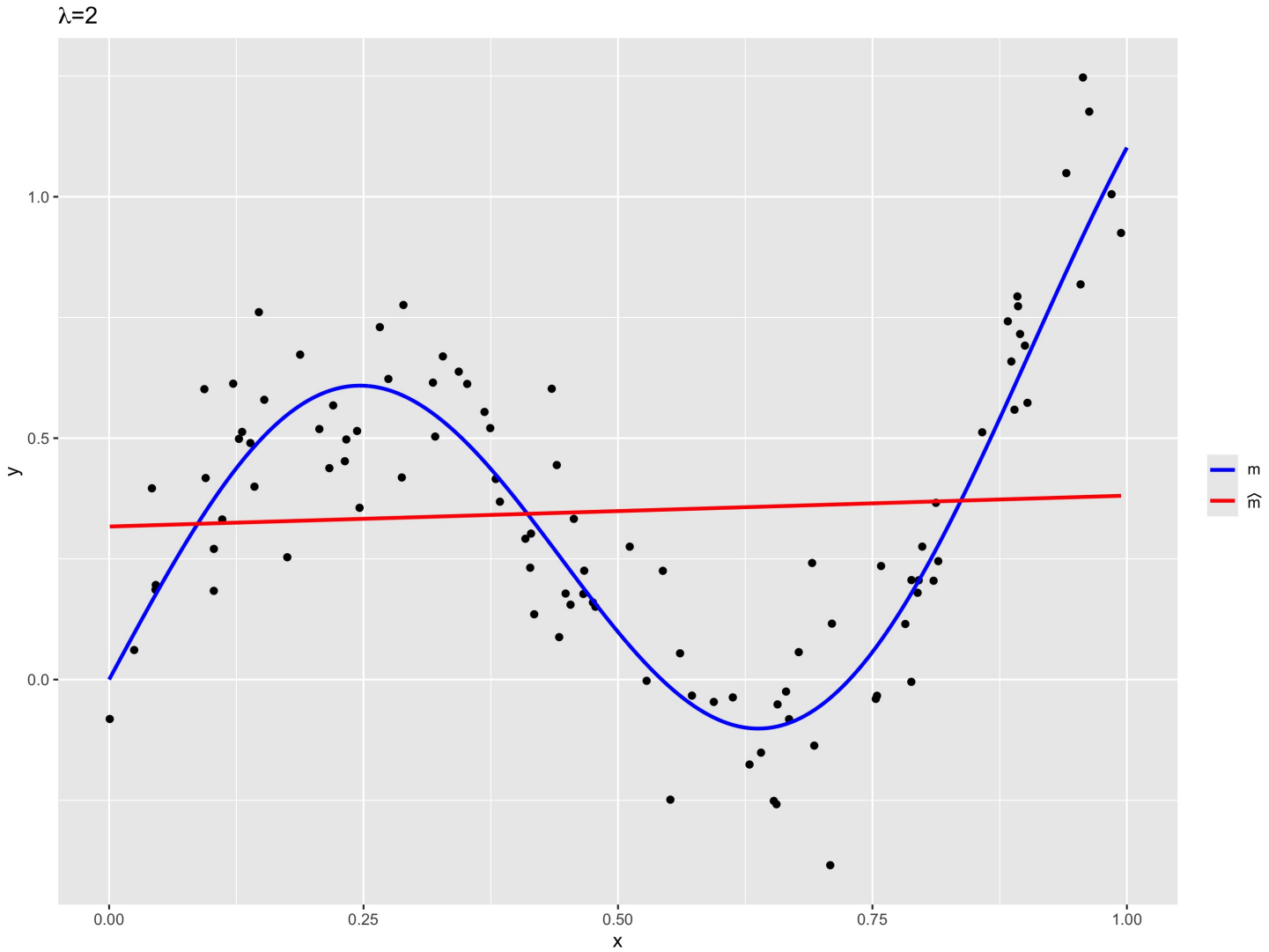


Illustration IV



Plan

GAM

Données considérées

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\dots,n\}}$$

où $X = \left(X^1, \dots, X^p\right)^\top \in \mathbb{R}^p$ et $Y \in \mathbb{R}$.

- On note :

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}} \ .$$

Le modèle

- Le modèle additif généralisé (GAM : *Generalized Additive Model*) suppose que :

$$Y = c + \sum_{j=1}^{p} g_j\left(X^j\right) + \varepsilon$$

où $g_j : \mathbb{R} \to \mathbb{R}$ sont p fonctions inconnues.

- Pour assurer l'existence, il faut imposer une contrainte, par exemple :

$$\forall j \in \{1, \ldots, p\} : \int g_j(x)\, \mathrm{d}x = 0 \ .$$

- On peut notamment estimer les fonctions $(g_j)_{j\in\{1,\ldots,p\}}$ par noyau, par polynômes locaux, par projection sur des bases orthogonales et par ajustement spline.

- La méthode des splines est la plus couramment employée pour les modèles GAM.

Le critère d'estimation

- On suppose ici que chaque fonction g_j est estimée à l'aide de fonctions splines.
- On cherche à minimiser :

$$\sum_{i=1}^{n}\left(y_i - c - \sum_{j=1}^{p} g_j\left(x_i^j\right)\right)^2 + \sum_{j=1}^{p}\lambda_j \int \left[g_j^{(2)}(x)\right]^2 \mathrm{d}x$$

où $(\lambda_j)_{j\in\{1,\dots,p\}} \in \mathbb{R}^{+p}$ sont des hyperparamètres de régularisation.

- Les solutions sont des splines cubiques, chaque fonction g_j ayant pour noeuds les $\left(x_i^j\right)_{i\in\{1,\dots,n\}}$.
- On impose comme contrainte d'unicité :

$$\forall j \in \{1,\dots,p\} : \sum_{i=1}^{n} g_j\left(x_i^j\right) = 0 \ .$$

Estimation : méthode de backfitting

1. Initialisation :

$$\widehat{c} = \bar{y} \ ,$$
$$\forall j \in \{1, \ldots, p\} : \widehat{g}_j(x) = 0 \ .$$

2. Pour $k \in \{1, \ldots, p\}$:

   2.1 On estime g_k en fixant tous les autres (étape de backfitting).
   Le problème à minimiser est :

$$\min_{g_k} \sum_{i=1}^{n} \left(y_i - \widehat{c} - \sum_{j=1, j \neq k}^{p} \widehat{g}_j\left(x_i^j\right) - g_k\left(x_i^k\right) \right)^2 + \lambda_k \int \left[{g_k}^{(2)}(x) \right]^2 \mathrm{d}x \ .$$

   2.2 On centre $\widehat{g}_k$ en lui soustrayant :

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{g}_k\left(x_i^k\right) .$$

On itère l'étape 2 jusqu'à stabilisation de l'optimisation.

Remarques

- Il existe d'autre méthodes d'estimation pour les modèles GAM.
- On peut réunir certaines covariables (en veillant au fléau de la dimension).
- On peut utiliser les modèles sur des séries temporelles.

Mise en œuvre logicielle

- Sous Python :
  - Package `pygam`.
- Sous R :
  - Package `gam` proposé par Hastie.
  - Package `mgcv` proposé par Wood.

Références

Friedman, J. H. 1991, «Multivariate adaptive regression splines», *The Annals of Statistics*, vol. 19, n° 1, p. 1–67.

Hastie, T. et R. Tibshirani. 1986, «Generalized additive models», *Statistical Science*, vol. 1, n° 3, p. 295–318.