

# MS BGD

## MDI 720: Au delà du modèle linéaire

**Joseph Salmon**

<http://josephsalmon.eu>

Télécom Paristech, Institut Mines-Télécom

# Syllabus

## Modèle linéaire généralisés

Régression Polynomiale

Régression polynomiale locale / Splines

Modèles additifs (généralisés)

## Robustesse

Moindres déviations absolues (Least Absolute Deviations)

# Syllabus

## Modèle linéaire généralisés

Régression Polynomiale

Régression polynomiale locale / Splines

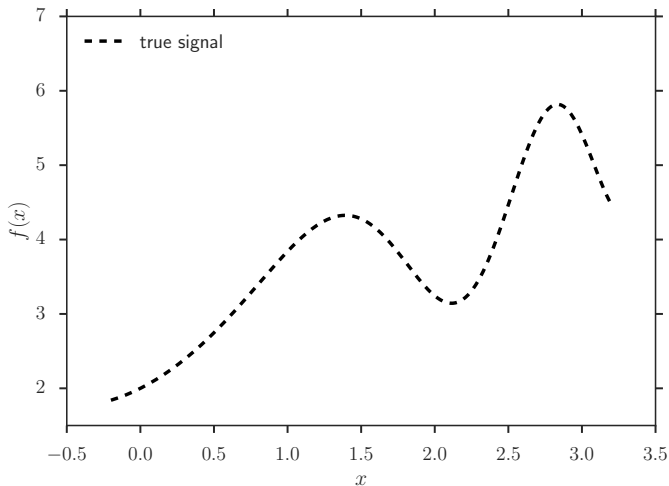
Modèles additifs (généralisés)

## Robustesse

Moindres déviations absolues (Least Absolute Deviations)

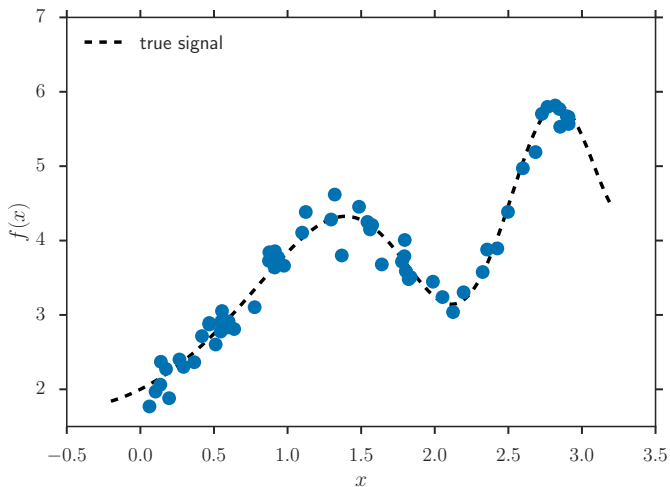
## Limites du modèle linéaire

Vrai signal:  $f(x_i)$  pour  $i = 1, \dots, n$



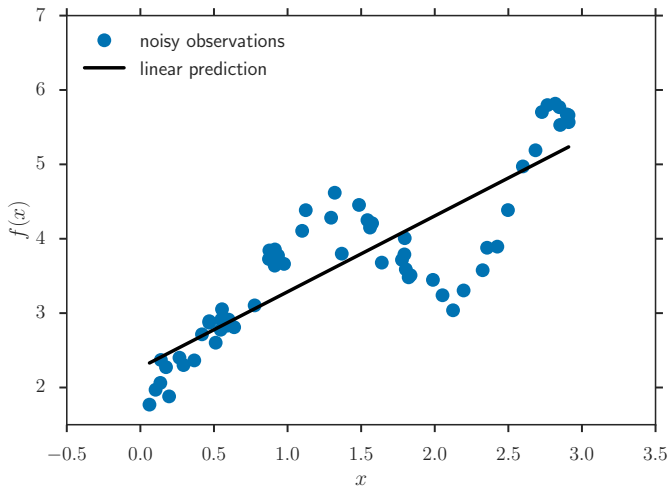
## Limites du modèle linéaire

Observations bruitées:  $y_i = f(x_i) + \varepsilon_i$  pour  $i = 1, \dots, n$



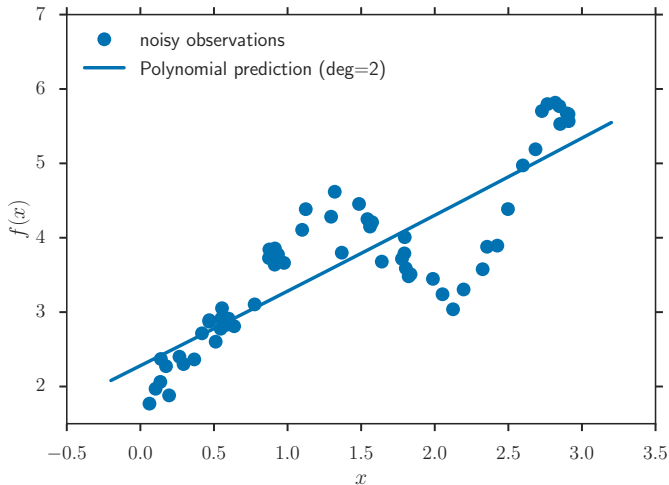
# Limites du modèle linéaire

Modèle linéaire: pas bien adapté ici



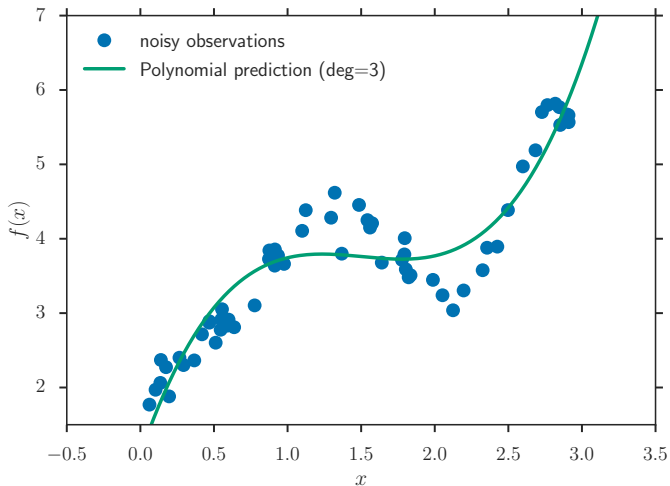
# Limites du modèle linéaire

Modèle linéaire: pas bien adapté ici



# Limites du modèle linéaire

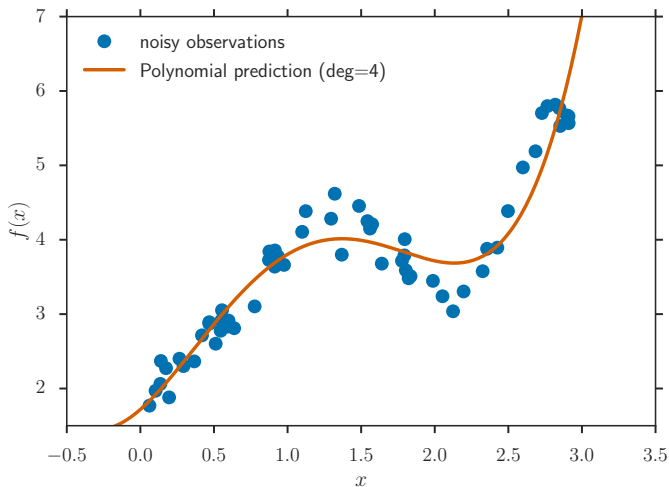
Modèle polynomial: mieux adapté ici





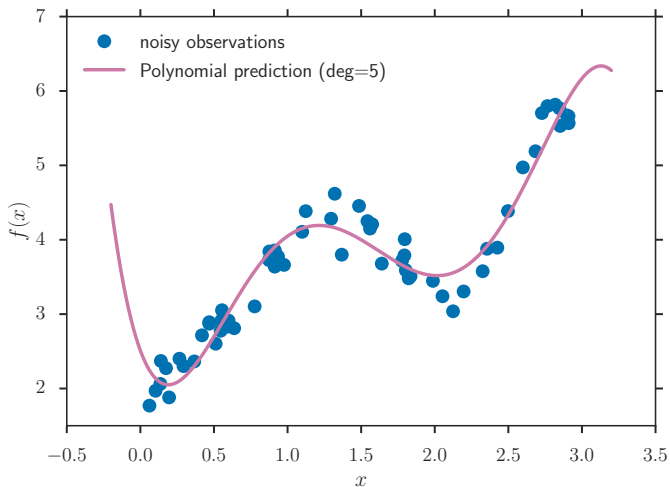
# Limites du modèle linéaire

Modèle polynomial: mieux adapté ici



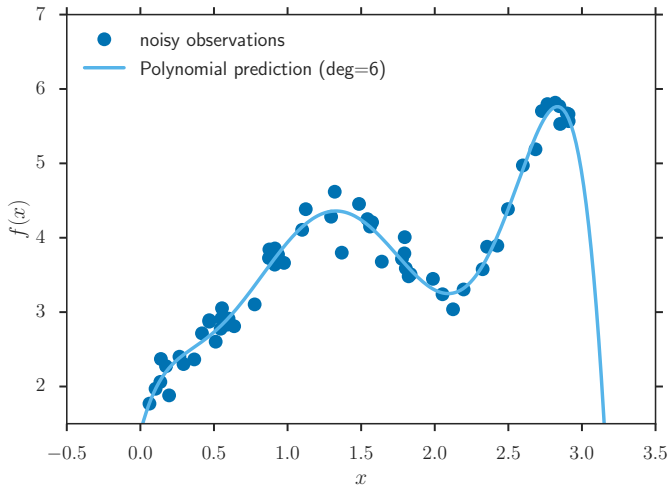
# Limites du modèle linéaire

Modèle polynomial: mieux adapté ici



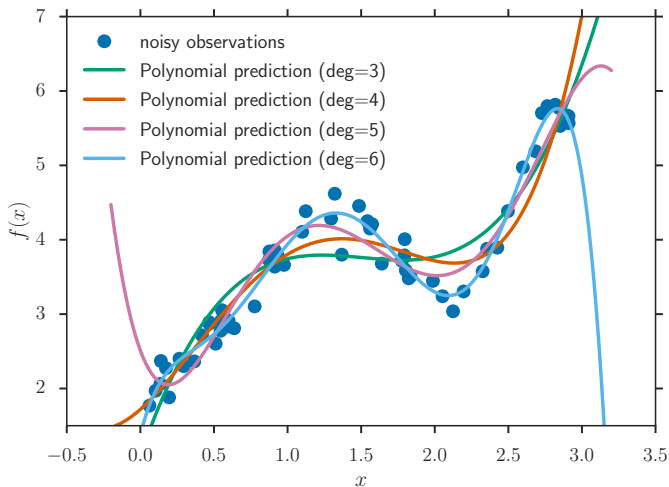
# Limites du modèle linéaire

Modèle polynomial: mieux adapté ici



# Limites du modèle linéaire

Modèle polynomial: mieux adapté ici



# Modèle polynomial

Soit  $D$  le degré du polynôme:

$$y_i = \theta_0^* + \sum_{d=1}^D \theta_d^* x_i^d + \varepsilon_i$$

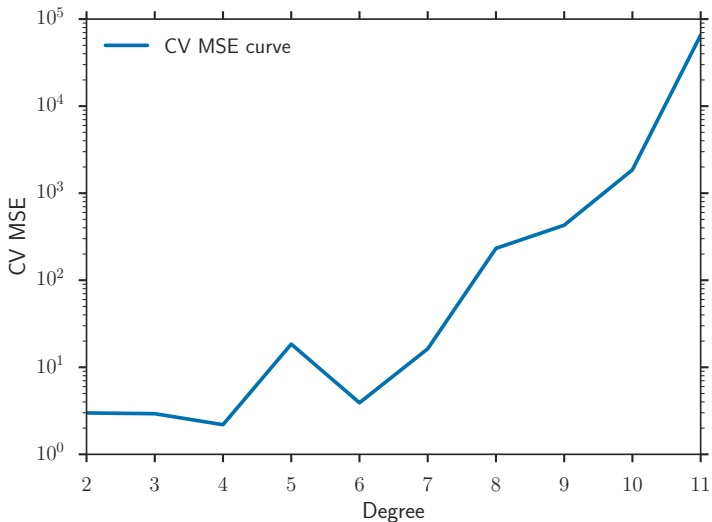
$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^D \\ 1 & x_2 & x_2^2 & \dots & x_2^D \\ 1 & x_3 & x_3^2 & \dots & x_3^D \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^D \end{pmatrix} \quad (\text{matrice de Vandermonde})$$

De manière équivalente  $X_{i,j} = x_i^{j-1}$  et  $\boldsymbol{\theta}^* = (\theta_0, \dots, \theta_D)^\top \in \mathbb{R}^{D+1}$   
et

$$\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$$

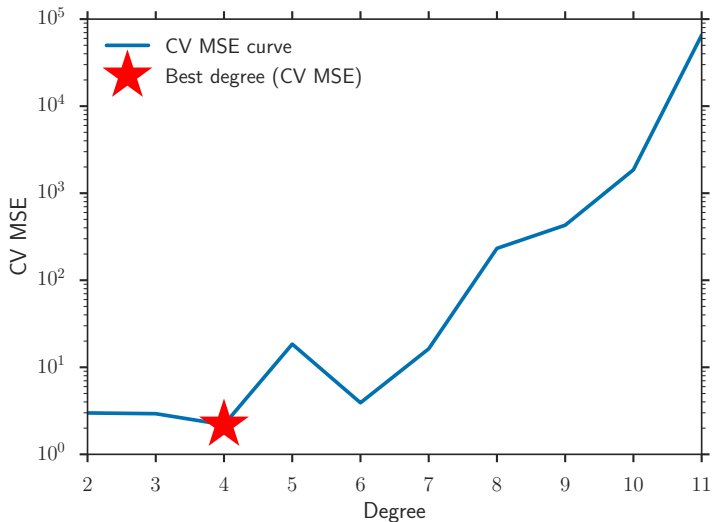
## Choix du degré

On peut utiliser la Validation Croisée (CV) pour choisir le degré



# Choix du degré

On peut utiliser la Validation Croisée (CV) pour choisir le degré



# Avantages/inconvénients de la régression polynomiale

## Avantages

- ▶ flexibilité pour de faible degrés
- ▶ utile en estimation non-paramétrique  
*cf. Green et Silverman (1994) Fan et Gijbels(1996)*

## Inconvénients

- ▶ les polynômes sont des fonctions globales (non localisées)
- ▶ le nombre de paramètres à estimer augmente vite avec la dimension et le degré



## Plusieurs covariables: $p = 2$ et $D = 2$

Considérons le cas  $x_i \in \mathbb{R}^2$

Ainsi  $x_i = [a_i, b_i]$ . Un polynôme d'ordre 2 requiert de fixer :

$$[1, a_i, b_i, a_i^2, a_i b_i, b_i^2]$$

Les coefficients  $a_i b_i$  représentent les interactions entre les variables  $x_1$  et  $x_2$ .

Cela peut se modéliser de manière compacte:

$$y_i = \theta_0 + \theta^\top x_i + \frac{1}{2} x_i^\top \Theta x_i + \varepsilon_i$$

$$y_i = \theta_0 + \sum_{j=1}^p \theta_j x_{i,j} + \frac{1}{2} \sum_{1 \leq j \leq k \leq p} \Theta_{j,k} x_{i,j} x_{i,k} + \varepsilon_i$$

où  $\Theta$  est une matrice (symétrique)  $p \times p$

## Plusieurs covariables: $p = 2$ et $D = 3$

Considérons le cas  $x_i \in \mathbb{R}^2$

Ainsi  $x_i = [a_i, b_i]$ . Un polynôme d'ordre 2 requiert de fixer :

$$[1, a_i, b_i, a_i^2, a_i b_i, b_i^2, a_i^3, a_i^2 b_i, a_i b_i^2, b_i^3]$$

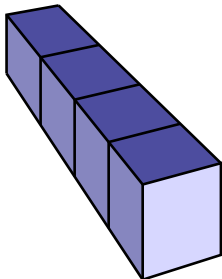
Les coefficients  $a_i b_i c_i$  représentent les interactions entre les variables  $x_1, x_2$  et  $x_3$ .

Cela peut se modéliser de manière compacte:

$$y_i = \theta_0 + \sum_{j=1}^p \theta_j x_{i,j} + \frac{1}{2} \sum_{1 \leq j \leq k \leq p} \Theta_{j,k} \cdot x_{i,j} x_{i,k} \\ + \frac{1}{6} \sum_{1 \leq j \leq k \leq \ell \leq p} \Theta_{j,k,\ell} \cdot x_{i,j} x_{i,k} x_{i,\ell} + \varepsilon_i$$

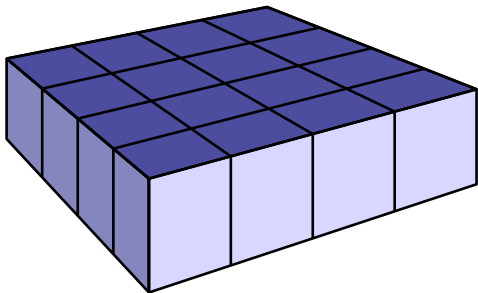
où  $\Theta$  est une matrice (symétrique)  $p \times p$ , et  $\Theta$  est un tenseur (symétrique)  $p \times p \times p$

# Représentation de tenseur 1D



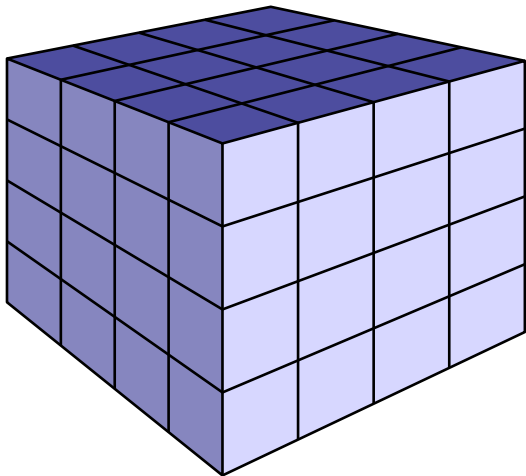
Cas vectoriel

# Représentation de tenseur 2D



Cas matriciel

# Représentation de tenseur 3D



Cas tensoriel

# Syllabus

## Modèle linéaire généralisés

Régression Polynomiale

Régression polynomiale locale / Splines

Modèles additifs (généralisés)

## Robustesse

Moindres déviations absolues (Least Absolute Deviations)

## Splines ( : cerces )

### Définition:

Un **spline**  $f$  est une fonction polynomiale par morceaux sur un intervalle  $[a, b]$ ,  $f : [a, b] \rightarrow \mathbb{R}$ , composé de  $n$  sous-intervalles  $[x_{i-1}, x_i]$  avec  $a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$ . La restriction de  $f$  sur chaque intervalle  $[x_{i-1}, x_i]$  est un polynôme

$P_i : [x_{i-1}, x_i] \rightarrow \mathbb{R}$ , ainsi

$$f(x) = P_1(x), \quad x_0 \leq t < x_1$$

$$f(x) = P_2(x), \quad x_1 \leq t < x_2$$

$\vdots$

$$f(x) = P_i(x), \quad x_{n-1} \leq t \leq x_n.$$

Le plus haut degré des polynômes  $P_i$  est appelé l'**ordre** du spline  $f$ , et les  $x_i$  sont appelé les **nœuds** (  : *knots* )

Rem: les plus populaires sont les splines (cubiques) d'ordre 3

Rem: on privilégie des splines lisses :  $C^0, C^1, C^2$ , etc.

# Utilisation

- ▶ statistiques
- ▶ computer vision, *cf.* courbes de Bézier dans [Inkscape](#) et autre logiciel de dessin vectoriel



# Utilisation

- ▶ statistiques
- ▶ computer vision, *cf.* courbes de Bézier dans [Inkscape](#) et autre logiciel de dessin vectoriel
- ▶ analyse numérique

# Utilisation

- ▶ statistiques
- ▶ computer vision, *cf.* courbes de Bézier dans [Inkscape](#) et autre logiciel de dessin vectoriel
- ▶ analyse numérique
- ▶ etc.

# Utilisation

- ▶ statistiques
- ▶ computer vision, *cf.* courbes de Bézier dans [Inkscape](#) et autre logiciel de dessin vectoriel
- ▶ analyse numérique
- ▶ etc.

# Algorithmes

Approches standards pour ajuster des splines quand on observe des points  $(x_i, y_i)$  pour  $i = 1, \dots, n$  : chercher le spline avec courbure minimum, *i.e.*, résoudre:

$$\hat{f} \triangleq SP_\lambda(\mathbf{y}) \in \arg \min_{f \text{ est un spline}} \left( \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \int_a^b |f''(t)|^2 dt \right)$$

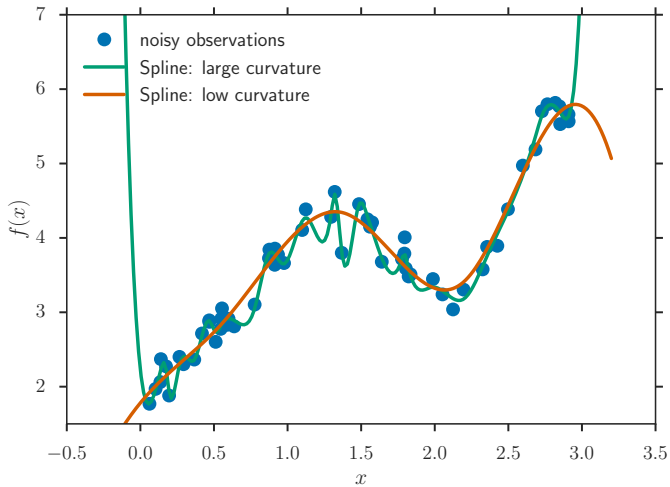
Fait: la solution est atteinte pour un spline cubique, et peut être obtenue par un moindre carré régularisé, avec  $\Omega \in \mathbb{R}^{n \times n}$

$$\arg \min_g \|\mathbf{y} - g\|^2 + \lambda g^\top \Omega g$$

voir détails dans [Ch. 2, Green and Silverman \(1994\)](#)

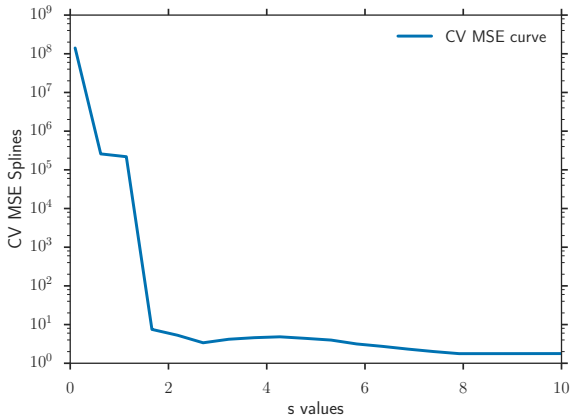
Note: avec cette régularisation les splines ont pour nœuds les  $x_i$

# Visual



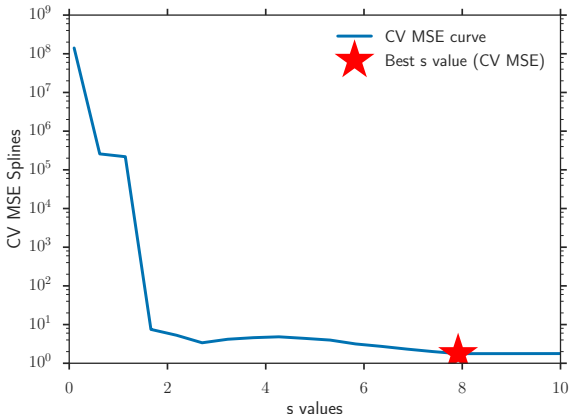
# Choix du paramètre de lissage

On peut utiliser la Validation Croisée (CV) pour choisir le niveau de lissage



# Choix du paramètre de lissage

On peut utiliser la Validation Croisée (CV) pour choisir le niveau de lissage



MSE Spline = 0.2498 vs. MSE Polynomials = 2.1899

# Syllabus

## Modèle linéaire généralisés

Régression Polynomiale

Régression polynomiale locale / Splines

Modèles additifs (généralisés)

## Robustesse

Moindres déviations absolues (Least Absolute Deviations)



## Modèles additifs pour la régression

Avec les des fonctions réelles, *i.e.*,  $f_j : \mathbb{R} \rightarrow \mathbb{R}$ , le modèle s'écrit

$$y_i = \sum_{j=1}^p f_j(x_{i,j}) + \varepsilon_i$$

Cela peut être résumé comme suit:

$$\mathbf{y} = \sum_{j=1}^p f_j(\mathbf{x}_j) + \varepsilon$$

avec la convention  $f_j(\mathbf{x}_j) = \begin{pmatrix} f_j(x_{1,j}) \\ \vdots \\ f_j(x_{n,j}) \end{pmatrix}$  avec  $\mathbf{x}_j = \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}$

Rem: potentiellement un des  $f_j$  encode la variable constante

Rem: GAM (Generalized Additive Models): extension aux modèles linéaires généralisés, *e.g.*, régression logistique,

$g(y_i) = \sum_{j=1}^p f_j(x_{i,j})$ , avec  $g$  une fonction

# Rétro-ajustement (Backfitting)

---

**Algorithm:** Rétro-ajustement d'un modèle additif

---

**Input** :  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$

---

# Rétro-ajustement (Backfitting)

---

**Algorithm:** Rétro-ajustement d'un modèle additif

---

**Input** :  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$

Initialize:  $f_1 \equiv 0, \dots, f_p \equiv 0$  and  $\mathbf{r} = \mathbf{y}$  (residual)

---

# Rétro-ajustement (Backfitting)

---

**Algorithm:** Rétro-ajustement d'un modèle additif

---

**Input** :  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$

Initialize:  $f_1 \equiv 0, \dots, f_p \equiv 0$  and  $\mathbf{r} = \mathbf{y}$  (residual)

**while** *not converged* **do**

|

---

# Rétro-ajustement (Backfitting)

---

**Algorithm:** Rétro-ajustement d'un modèle additif

---

**Input** :  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$

Initialize:  $f_1 \equiv 0, \dots, f_p \equiv 0$  and  $\mathbf{r} = \mathbf{y}$  (residual)

**while** *not converged* **do**

**for**  $j = 1, \dots, p$  **do**

        |

    |

# Rétro-ajustement (Backfitting)

---

**Algorithm:** Rétro-ajustement d'un modèle additif

---

**Input** :  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$

Initialize:  $f_1 \equiv 0, \dots, f_p \equiv 0$  and  $\mathbf{r} = \mathbf{y}$  (residual)

**while** *not converged* **do**

**for**  $j = 1, \dots, p$  **do**

$\mathbf{r} \leftarrow \mathbf{r} + f_j(\mathbf{x}_j)$

        // Partial residual update

# Rétro-ajustement (Backfitting)

---

**Algorithm:** Rétro-ajustement d'un modèle additif

---

**Input** :  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$

Initialize:  $f_1 \equiv 0, \dots, f_p \equiv 0$  and  $\mathbf{r} = \mathbf{y}$  (residual)

**while** *not converged* **do**

**for**  $j = 1, \dots, p$  **do**

$\mathbf{r} \leftarrow \mathbf{r} + f_j(\mathbf{x}_j)$

        // Partial residual update

$f_j \leftarrow SP_{\lambda_j}(\mathbf{r})$

        // update with spline (param.  $\lambda_j$ )

# Rétro-ajustement (Backfitting)

---

**Algorithm:** Rétro-ajustement d'un modèle additif

---

**Input** :  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$

Initialize:  $f_1 \equiv 0, \dots, f_p \equiv 0$  and  $\mathbf{r} = \mathbf{y}$  (residual)

**while** *not converged* **do**

**for**  $j = 1, \dots, p$  **do**

$\mathbf{r} \leftarrow \mathbf{r} + f_j(\mathbf{x}_j)$

  // Partial residual update

$f_j \leftarrow SP_{\lambda_j}(\mathbf{r})$

  // update with spline (param.  $\lambda_j$ )

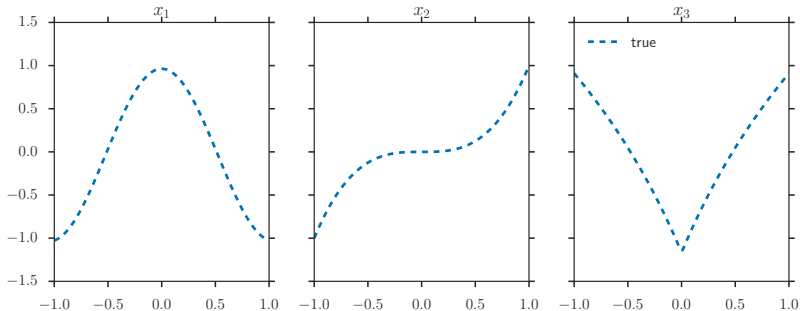
$\mathbf{r} \leftarrow \mathbf{r} - f_j(\mathbf{x}_j)$

  // Partial residual un-update





# GAM en action



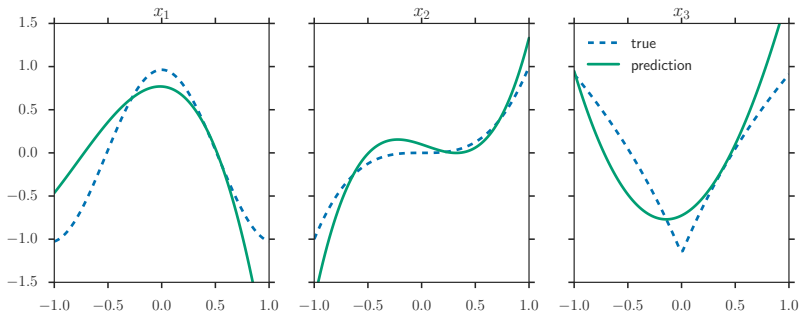
où  $y = f(\mathbf{x}) + \varepsilon$  avec  $f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + f_3(x_3)$  et

$$f_1(x) = \cos(3x)$$

$$f_2(x) = x^3$$

$$f_3(x) = 3 \log(1 + |x|)$$

# GAM en action



où  $y = f(\mathbf{x}) + \varepsilon$  avec  $f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + f_3(x_3)$  et

$$f_1(x) = \cos(3x)$$

$$f_2(x) = x^3$$

$$f_3(x) = 3 \log(1 + |x|)$$

# Pros and cons of GAM

## Pros

- ▶ can model non-linear effect automatically
- ▶ interpretation is possible : functions are 1D (can visualize!)
- ▶ can be extended to second order interactions of features (for small  $p$ )

## Cons

- ▶ Stopping is not so simple (non-convex nature...)
- ▶ Proper tuning is hard: at least one parameter by feature

## Plus de détails sur les GAM:

- ▶ Vidéo: <https://vimeo.com/125940125>
- ▶ Livre: Hastie and Tibshirani (1990)

## Plus de détails sur les GAM:

- ▶ Vidéo: <https://vimeo.com/125940125>
- ▶ Livre: Hastie and Tibshirani (1990)

# Syllabus

## Modèle linéaire généralisés

Régression Polynomiale

Régression polynomiale locale / Splines

Modèles additifs (généralisés)

## Robustesse

Moindres déviations absolues (Least Absolute Deviations)

## Les moindres carrés historiquement



(a) **Adrien-Marie Legendre**: "Nouvelles méthodes pour la détermination des orbites des comètes", 1805




(b) **Carl Friedrich Gauss**: "Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium" 1809



## et juste avant ...


### Définition

L'estimateur des Moindres Déviations Absolues ( : *Least Absolute Deviations (LAD)*) est donné par:

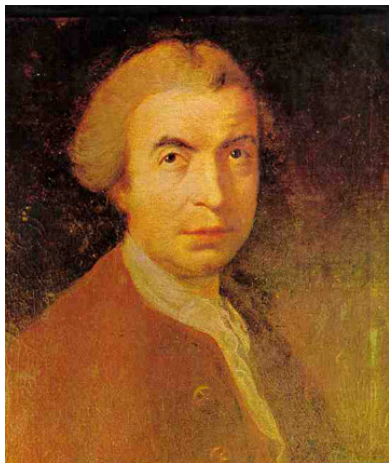
$$(\hat{\theta}) \in \arg \min_{\theta \in \mathbb{R}^p} \sum_{i=1}^n |y_i - x_i^\top \theta|$$

avec  $X = [x_1, \dots, x_n]^\top$  (description par ligne)

Rem: problème d'optimisation plus difficile que les moindres carrés; nécessite un algorithme d'optimisation adapté à l'optimisation non lisse (*i.e.*, fonctions non différentiables)

Rem: cet estimateur est moins sensible aux éléments atypiques ( : *outliers*), *e.g.*, observations ayant un  $\varepsilon_i$  important

# Paternité des Moindres déviations absolues

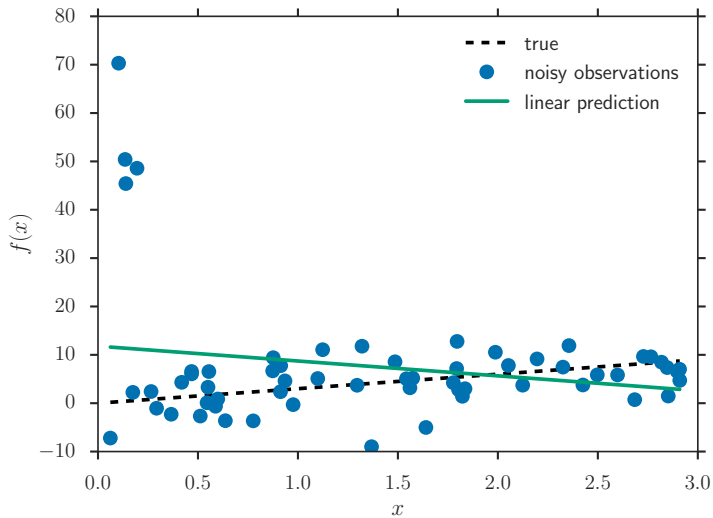


(c) **Ruđer Josip Bošković**: "???",  
1757

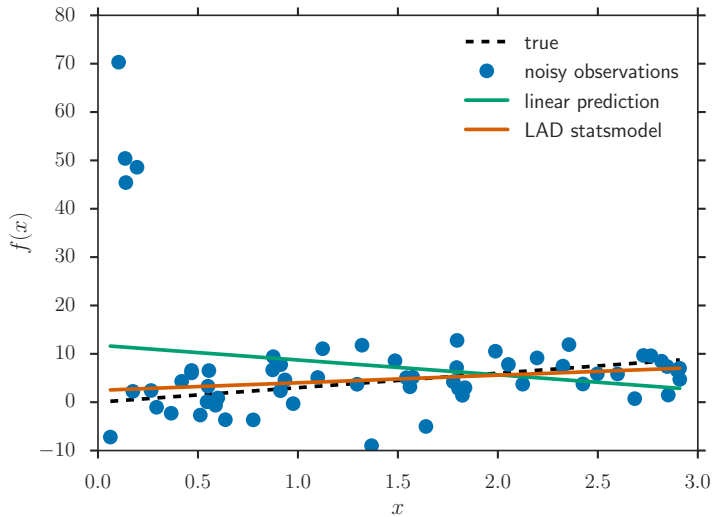


(d) **Pierre-Simon de Laplace**: "Traité  
de mécanique céleste", 1799


# LAD in action



# LAD in action



## Points non abordés: extensions possibles

- ▶ Robustesse: attache aux données  $\ell_1$ , régression quantile, moyennes tronquées, etc.
- ▶ Méthodes gloutones ( : *greedy*)
- ▶ boosting/bagging
- ▶ Point de vue bayésien
- ▶ Arbres / forêts (classification surtout)
- ▶  $K$ -plus proches voisins
- ▶ SVM (classification)
- ▶ Réseaux de neurones (classification surtout)

# References I

- ▶ J. Fan and I. Gijbels.

*Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*.

Chapman & Hall, London, 1996.

- ▶ P. J. Green and B. W. Silverman.

*Nonparametric regression and generalized linear models*, volume 58 of *Monographs on Statistics and Applied Probability*.

Chapman & Hall, London, 1994.

A roughness penalty approach.

- ▶ T. J. Hastie and R. J. Tibshirani.

*Generalized additive models*, volume 43.

CRC press, 1990.