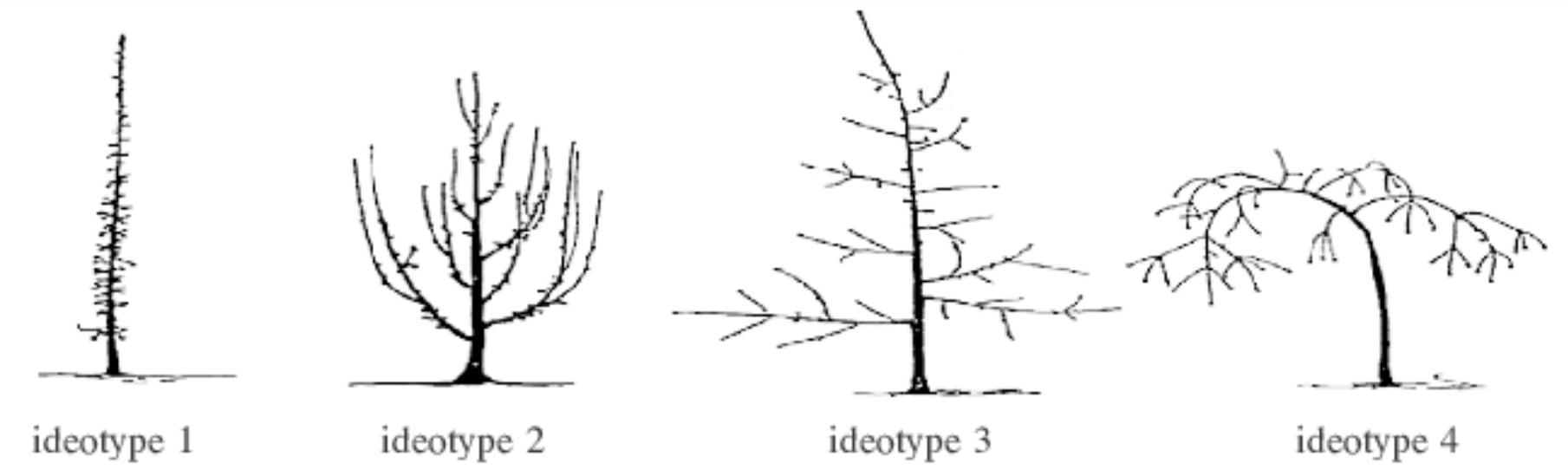
Métamodélisation et optimisation robuste

Application à la conception d'idéotypes sous incertitude climatique.

Léonard Torossian

Sous la direction de Robert Faivre, Aurélien Garivier et Victor Picheny

Présentation de l'optimisation séquentielle



Définition

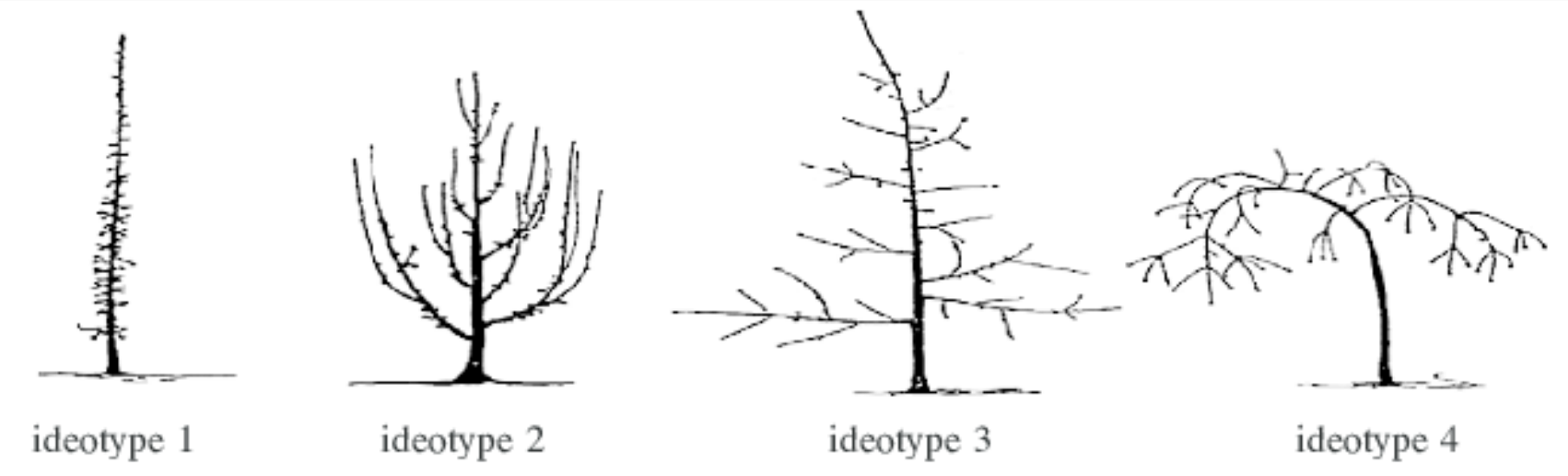
Un idéotype est une variété de plant de culture sélectionnée pour sa capacité à profiter de manière optimale d'un environnement donné.

Présentation de l'optimisation séquentielle

- Un budget initial
- Identifier la meilleure machine en un nombre minimal de parties

MIA TOULOUSE

Présentation de l'optimisation séquentielle



Simulateur Sunflo

Modèle Sunflo

Modèle Sunflo

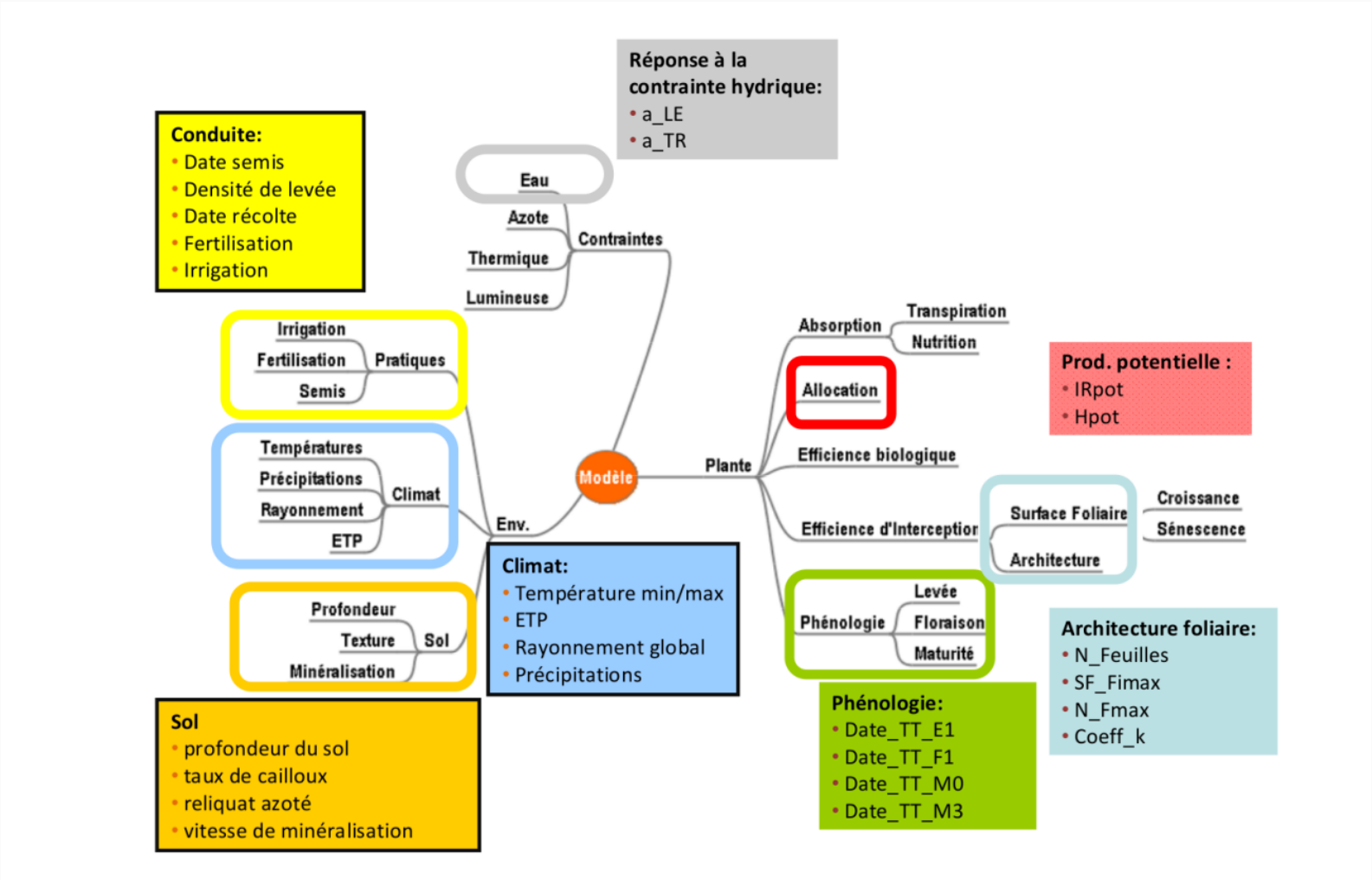
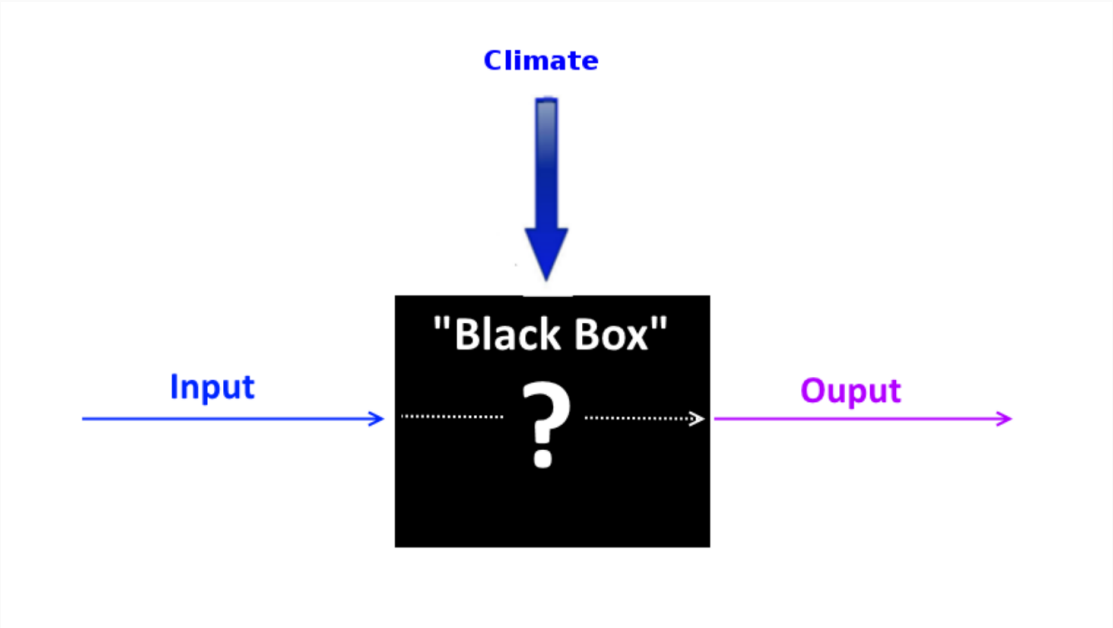


Figure 1: Modèle sunflo [Debaeke et al., 2011]

Modèle Sunflo

- Nous allons nous **affranchir** de cette **complexité**
- Incorporer toute la **complexité** du modèle sous la forme d'une **boîte noire stochastique**

Modèle Sunflo



- Une boîte noire stochastique est une fonction $f : \mathbb{X} \subset \mathbb{R}^d \times \Omega \to \mathbb{R}$
- L'entrée $x \in \mathbb{X}$ représente un paramètre fixé, il peut être déterministe ou stochastique.
- L'entrée ω représente le caractère stochastique du code.
- **L'objectif est d'obtenir des informations sur la loi de**

$$Y(x) = f(x, \omega) \quad \text{pour tout} \quad x \in \mathbb{X}$$

Le quantile conditionnel

Le quantile conditionnel

Définition

Soient (X, Y) une paire de variable aléatoires et $\alpha \in (0, 1)$. Le quantile conditionnel de Y sachant $X = x$ de niveau α , $q_\alpha(x)$ est défini par la fonction $q_\alpha : \mathcal{X} \to \mathbb{R}$ qui point par point, minimise en q

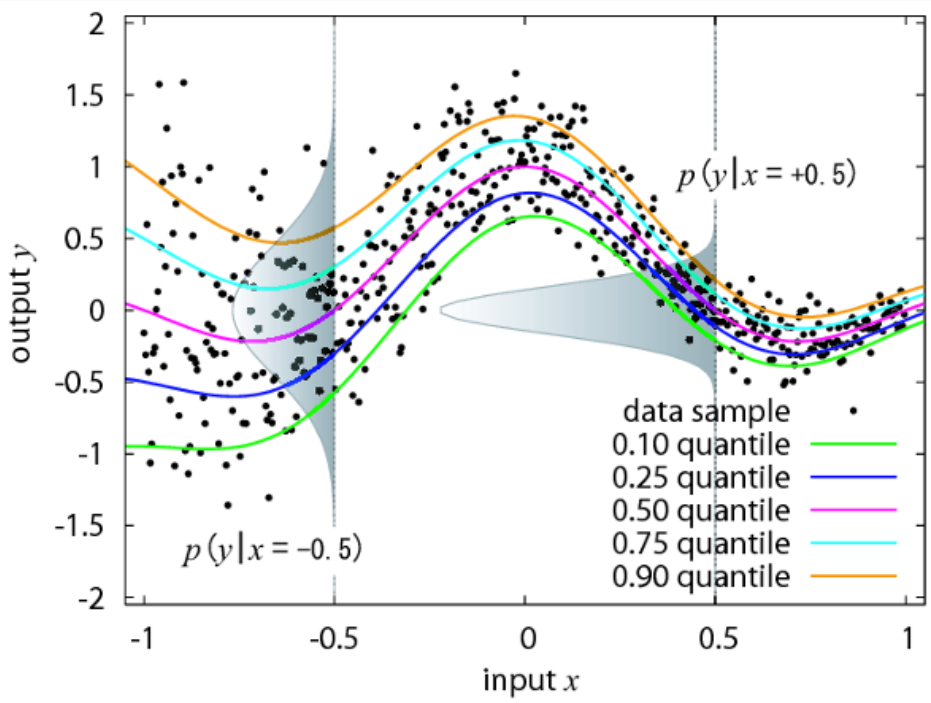
$$\mathbb{P}(Y \leq q | X = x) \geq \alpha$$

Le quantile conditionnel

Définition

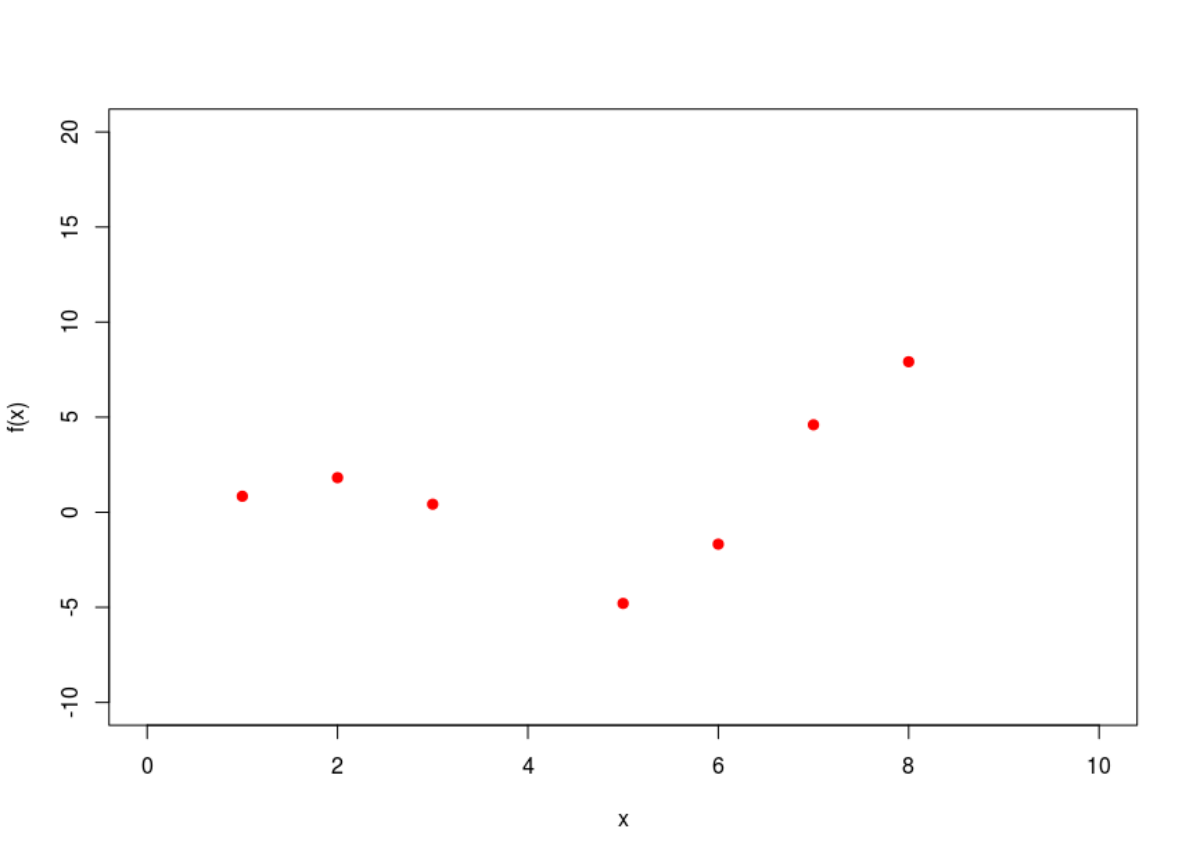
Soient (X, Y) une paire de variable aléatoires et $\alpha \in (0, 1)$. Le quantile conditionnel de Y sachant $X = x$ de niveau α , $q_\alpha(x)$ est défini par la fonction $q_\alpha : \mathcal{X} \to \mathbb{R}$ qui point par point, minimise en q

$$\mathbb{P}(Y \leq q | X = x) \geq \alpha$$



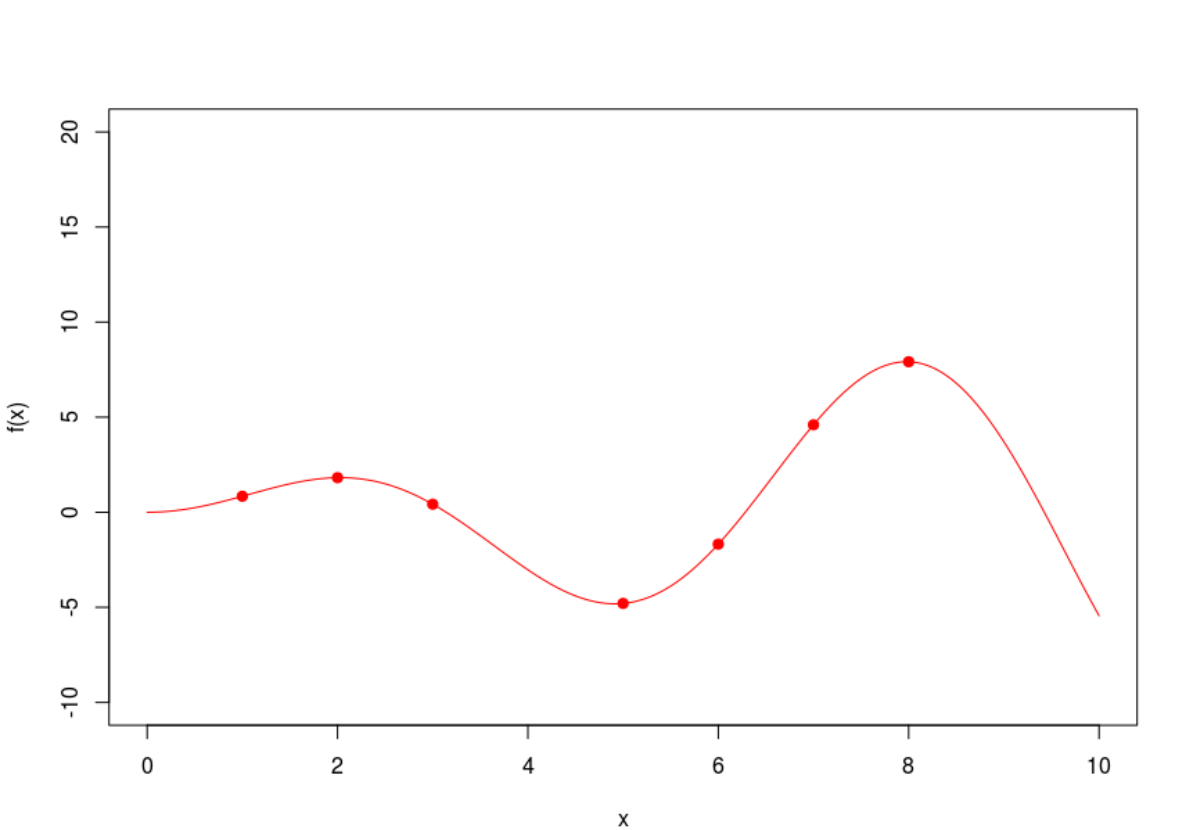
Métamodélisation

Boîte noire déterministe



On se donne un échantillon $(x_i, y_i)_{1 \leq i \leq n}$

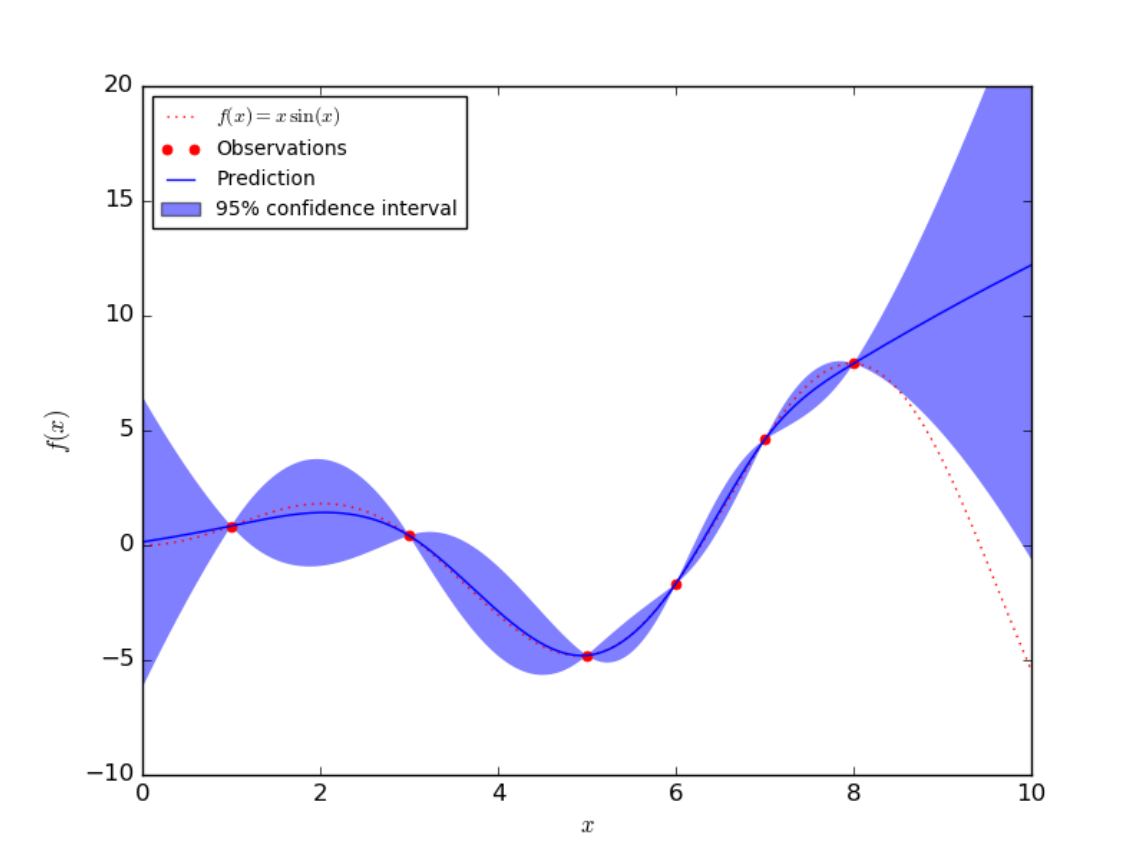
Boîte noire déterministe



Nous souhaitons trouver une fonction g telle que :

$$y_i = g(x_i), \quad \forall i, 1 \leq i \leq n$$

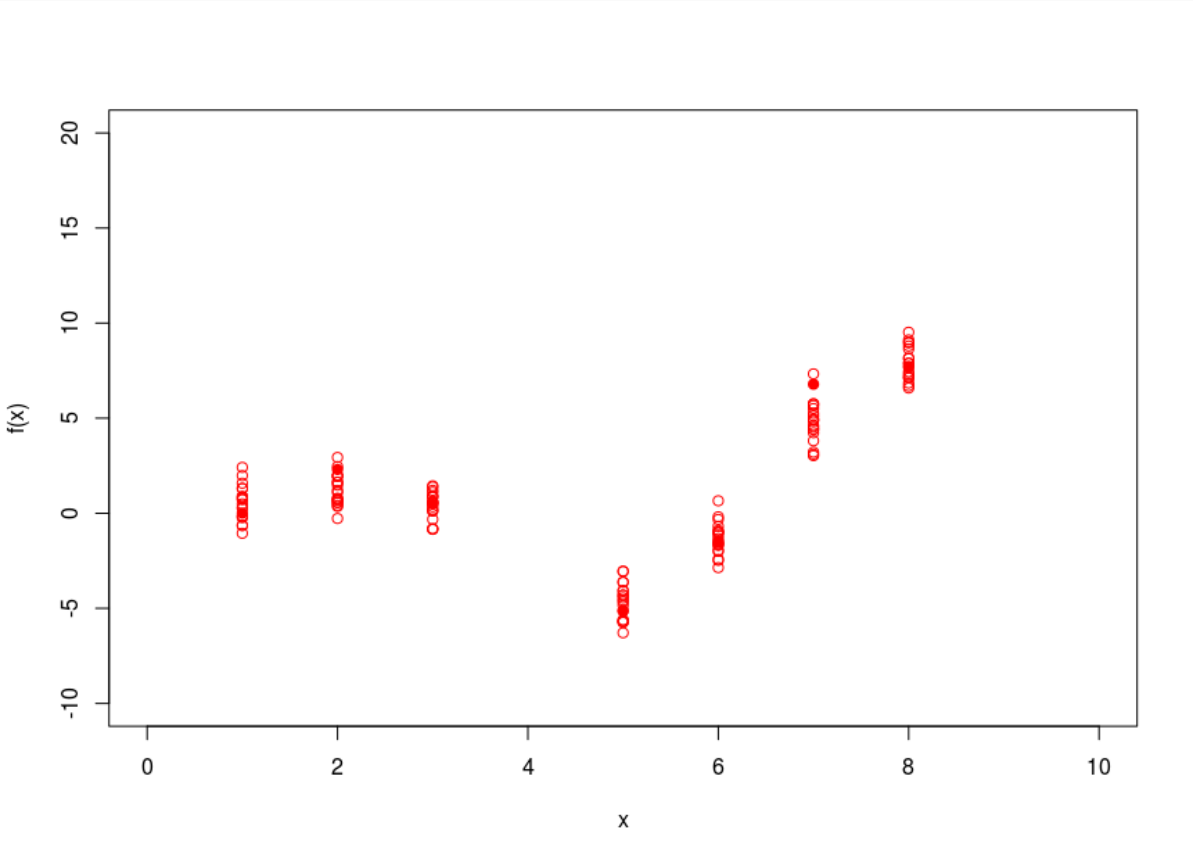
Boîte noire déterministe



- A priori gaussien sur la fonction visée.

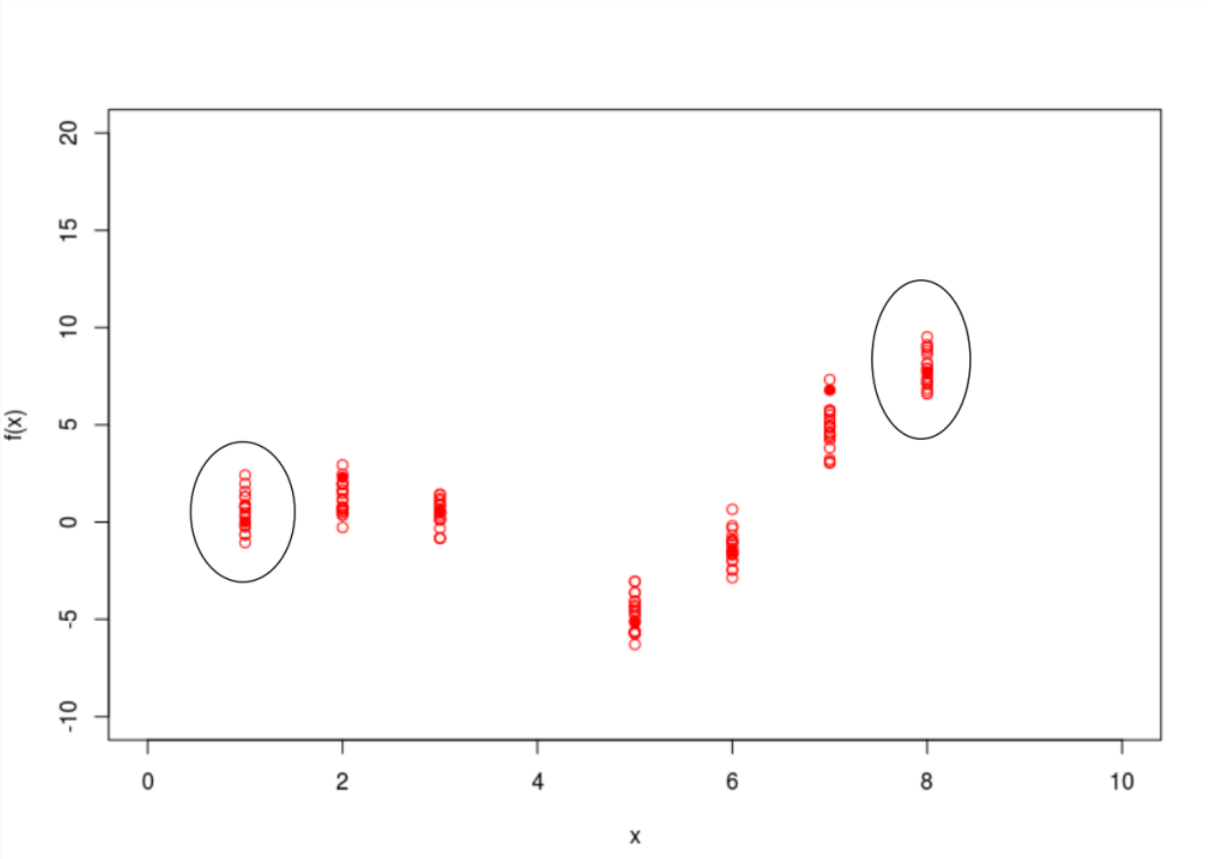
Boîte noire stochastique

Homoscédasticité : La variance est la même pour chaque observation

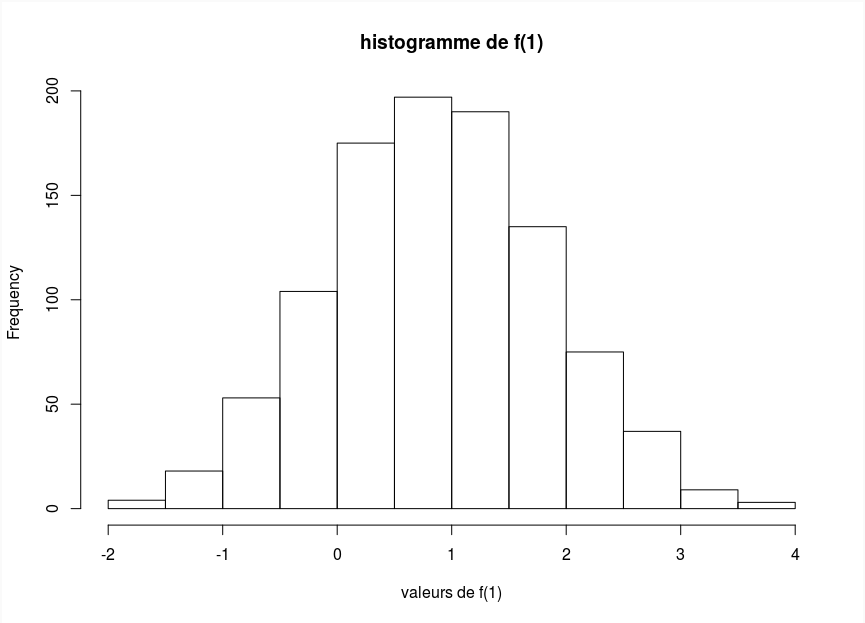


Boîte noire stochastique

Homoscédasticité : La variance est la même pour chaque observation



Homoscédasticité



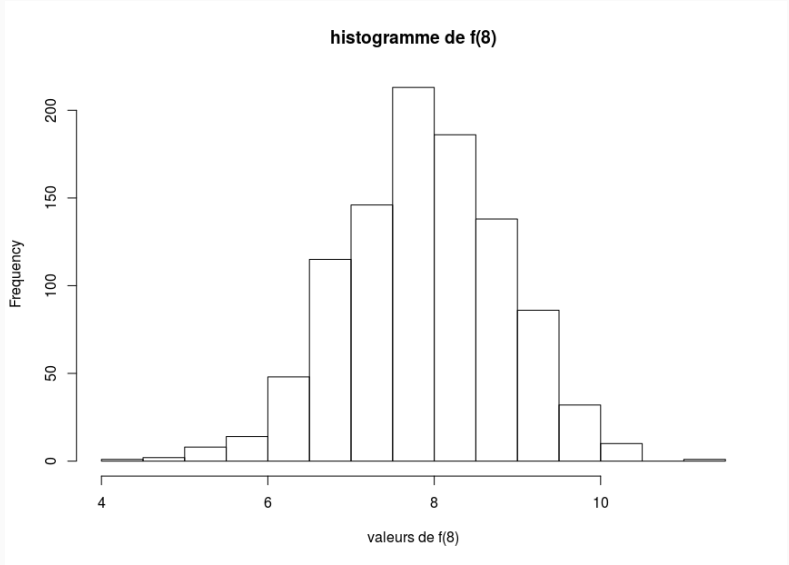
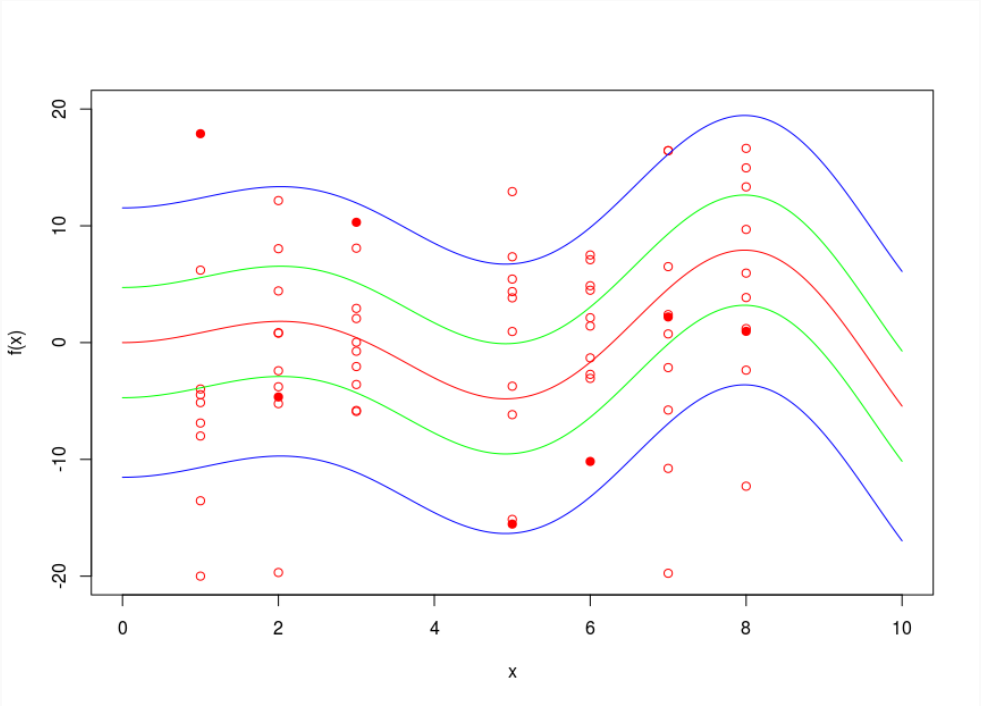


Figure 3: Histogramme de la sortie f en $x = 1$ et $x = 8$.

Boîte noire stochastique

Homoscédasticité



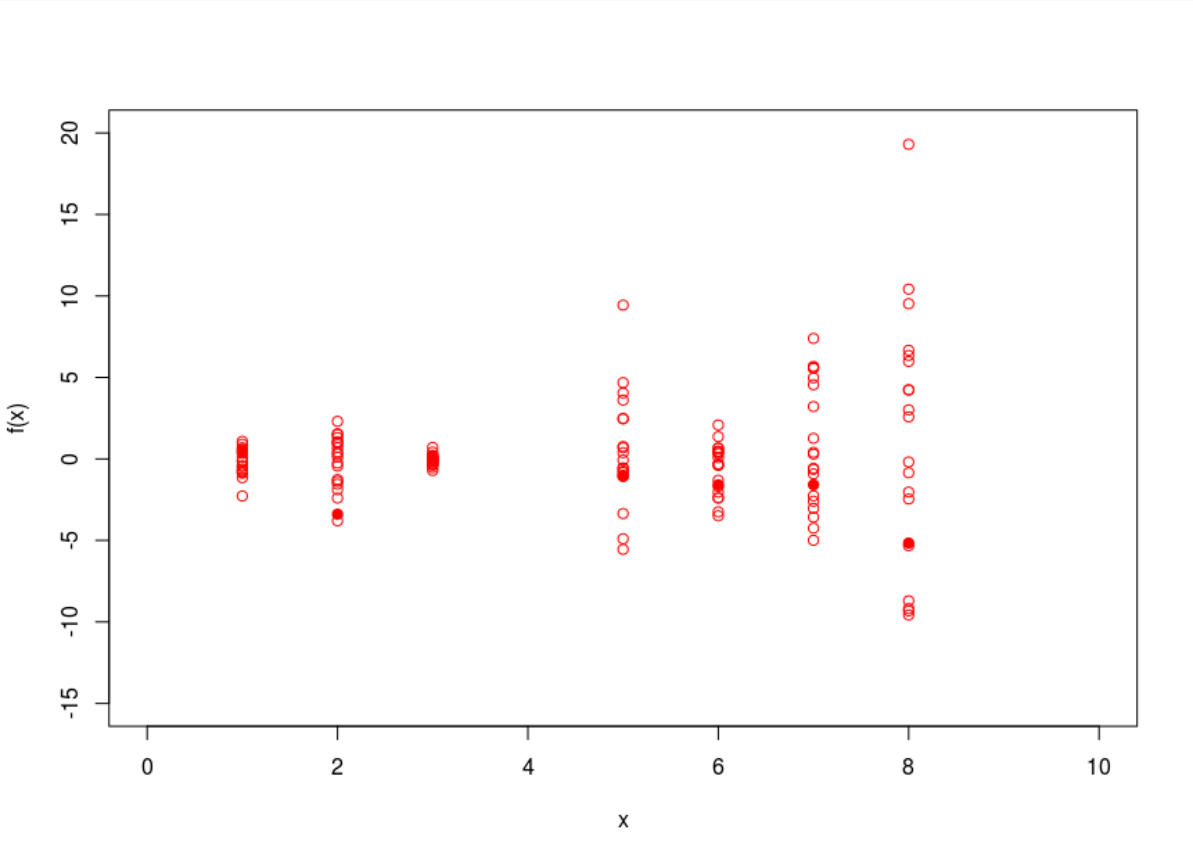
Si on connaît :

- La moyenne
- La variance du bruit en un seul point

→ **On peut reconstruire la distribution de probabilité en tout point**

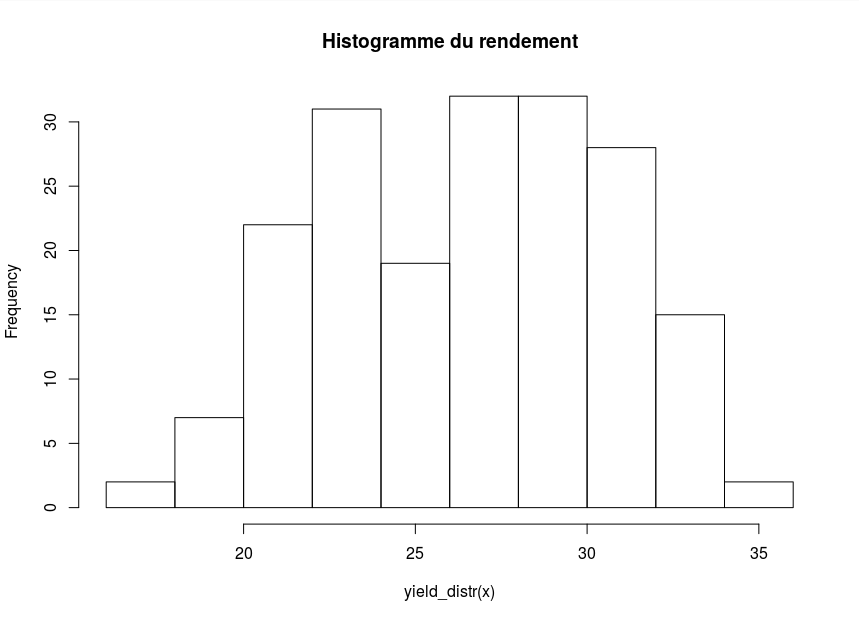
Boîte noire stochastique

Hétéroscédasticité : La variance est différente pour chaque observation



Boîte noire stochastique

Sunflo est Hétéroscédastique !



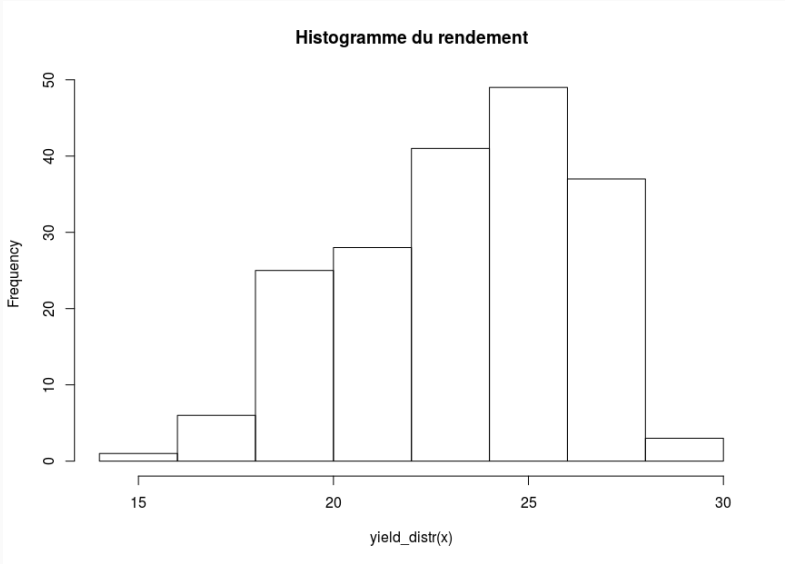


Figure 4: Deux distribution associées à deux phénotypes différents

Estimation des quantiles

Estimation des quantiles

Nous supposons que le code est structuré en espace.

- Les méthodes à noyau permettent de travailler avec cette hypothèse. [Steinwart and Christmann, 2008]

- Un noyau $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ peut être vu comme une mesure de dépendance entre deux éléments de l'espace $\mathbb{X}$.

MIA TOULOUSE

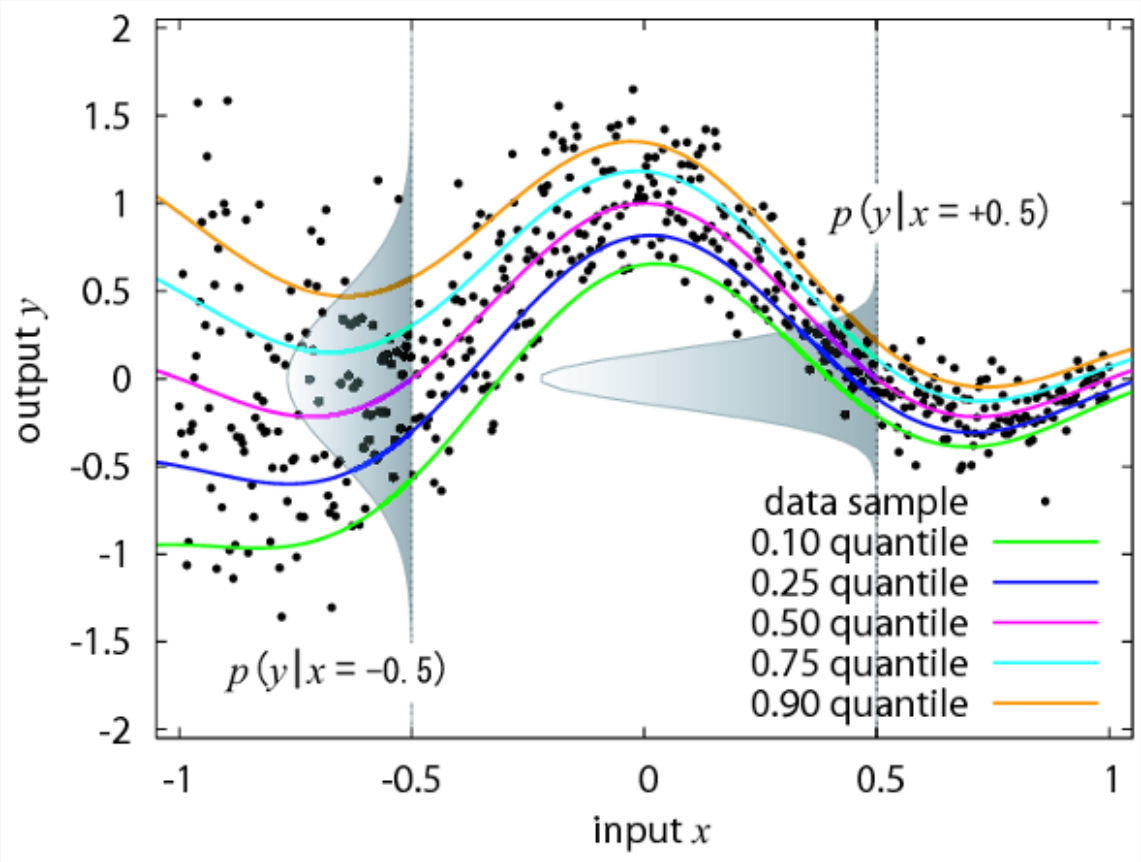
Estimation des quantiles

Nous allons privilégier deux approches non paramétriques :

- Une approche bayésienne utilisant les processus gaussiens.[Plumlee and Tuo, 2014][Rasmussen, 2006]

- Une approche fréquentiste de type machine à support de vecteur (SVM). [Takeuchi et al., 2006]

Boîte noire déterministe



Estimation des quantiles

Définition

Soit L une fonction de perte, le risque associé à L s'écrit :

$$\mathcal{R}[f] := \mathbb{E}_{p(x,y)}[L(y - f(x))]$$

Le risque empirique associé est :

$$\mathcal{R}_{emp}[f] := \frac{1}{m} \sum_{i=1}^{m} L(y_i - f(x_i))$$

- Typiquement la perte quadratique est utilisée pour estimer la moyenne conditionnelle.
- Il nous faut une autre fonction de perte pour estimer un quantile.

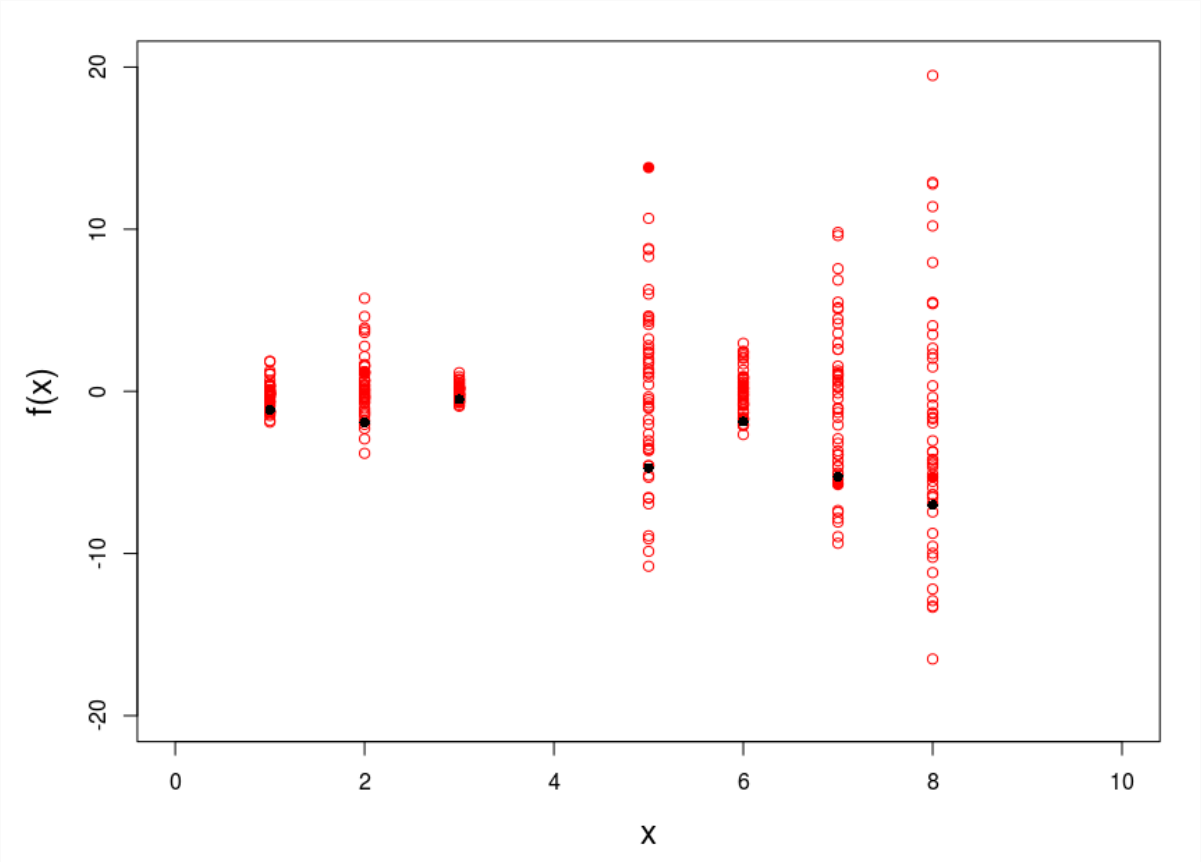
Approche Bayésienne

On suppose que chaque quantile suit un processus gaussien, de moyenne $m_\alpha(.)$ et de variance $k_\alpha(.,.)$. Nous noterons :

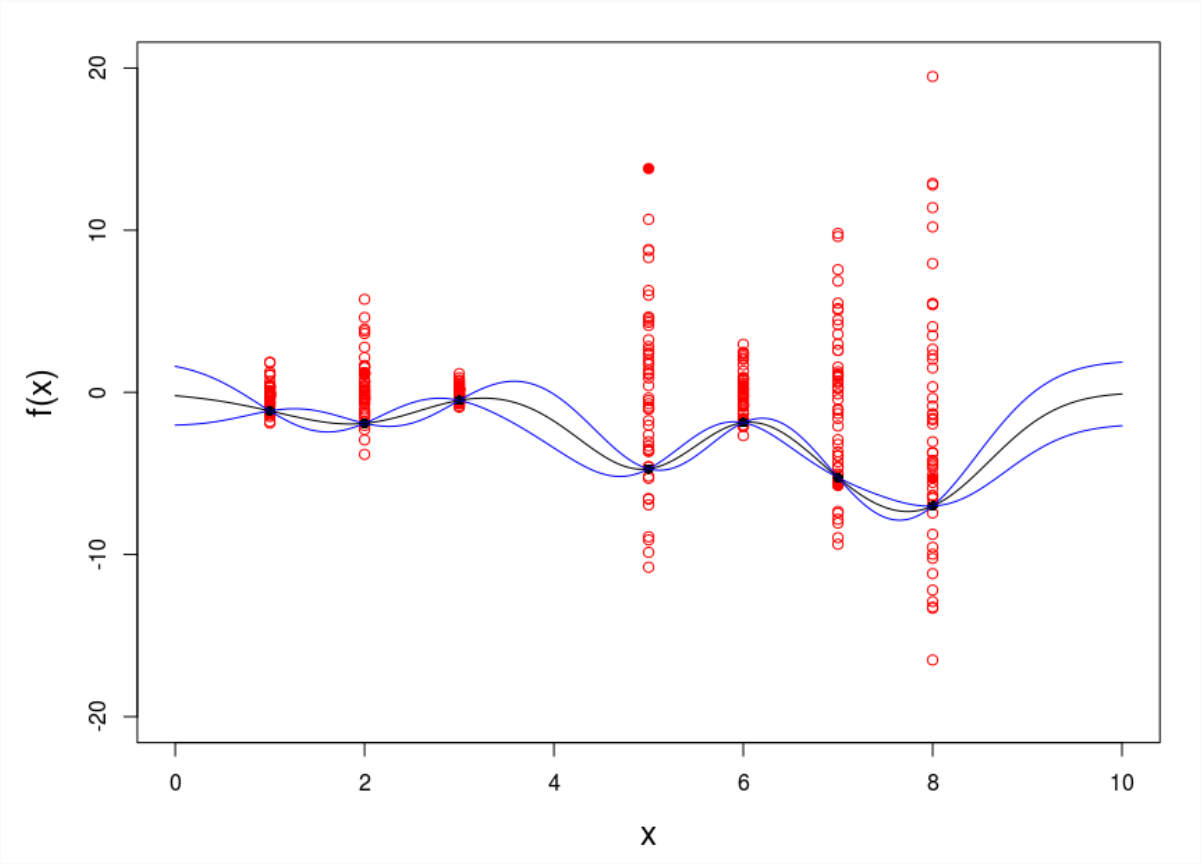
$$q_\alpha(x) \sim \mathbb{GP}(m_\alpha(x), k_\alpha(x, x'))$$

Comment faire passer le processus gaussien au bon endroit ?

Approche Bayésienne



Approche Bayésienne

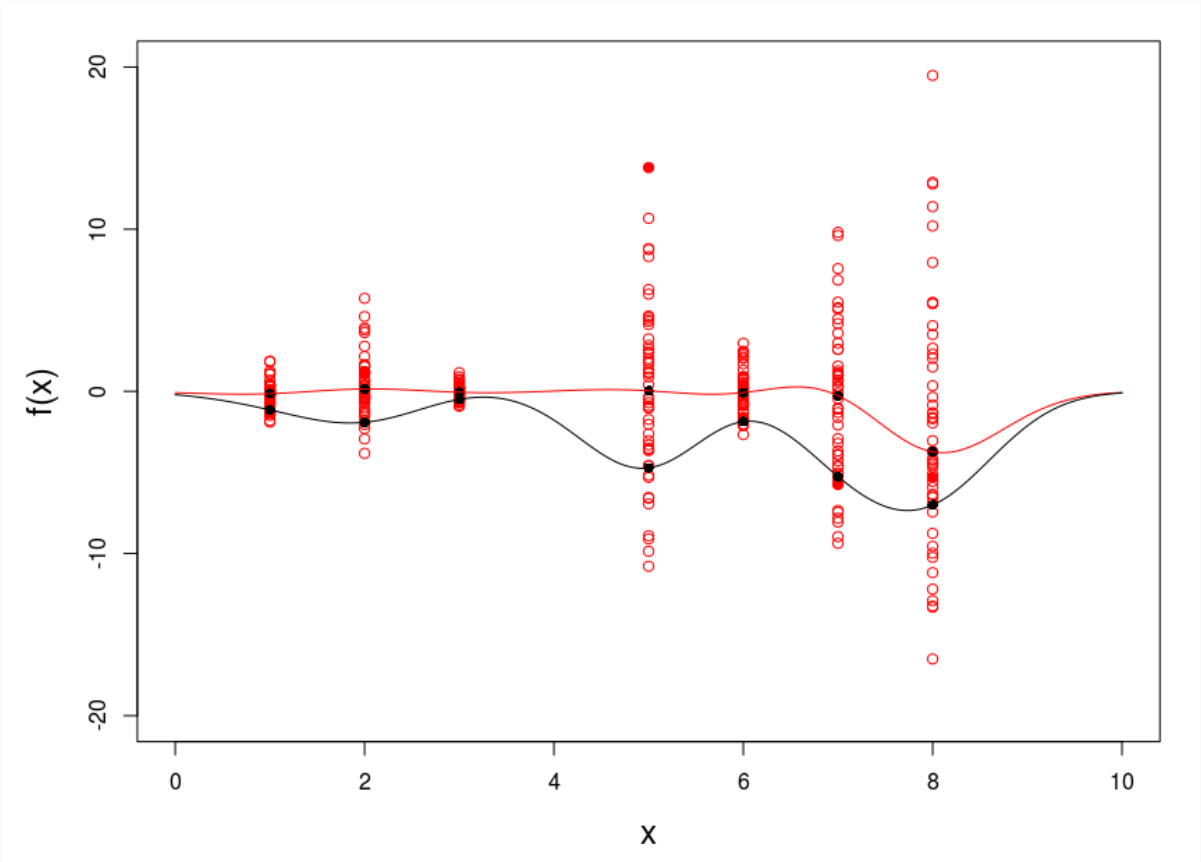


Définition de l'estimateur

$$m_{q_\alpha}(x^*) = k(x^*, x)^T (K + \sigma_n^2 I)^{-1} Y_\alpha$$

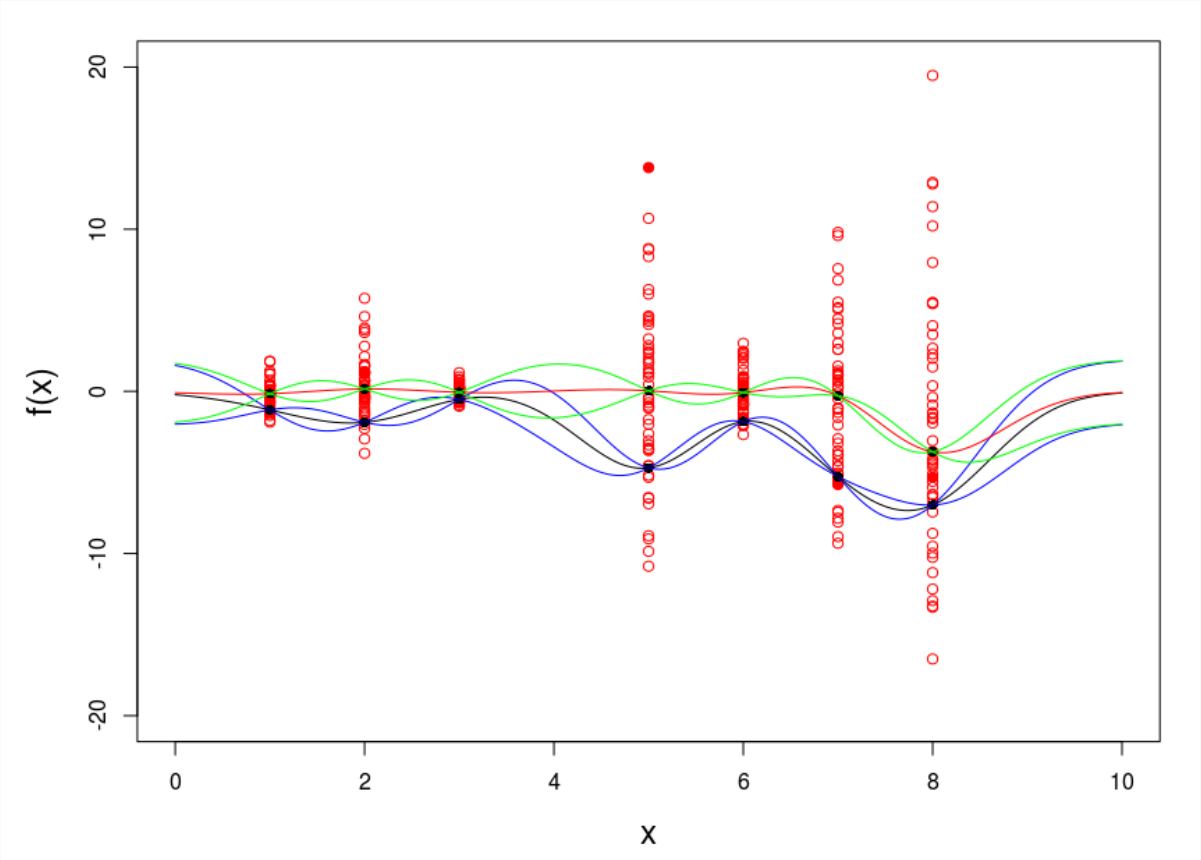
$$\mathbb{V}_{q_\alpha}(x^*) = k(x^*, x^*) - k^T(x^*, x)(K + \sigma_n^2 I)^{-1} k(x^*, x)$$

Approche Bayésienne



→ Nous pouvons itérer cette méthode

Approche Bayésienne



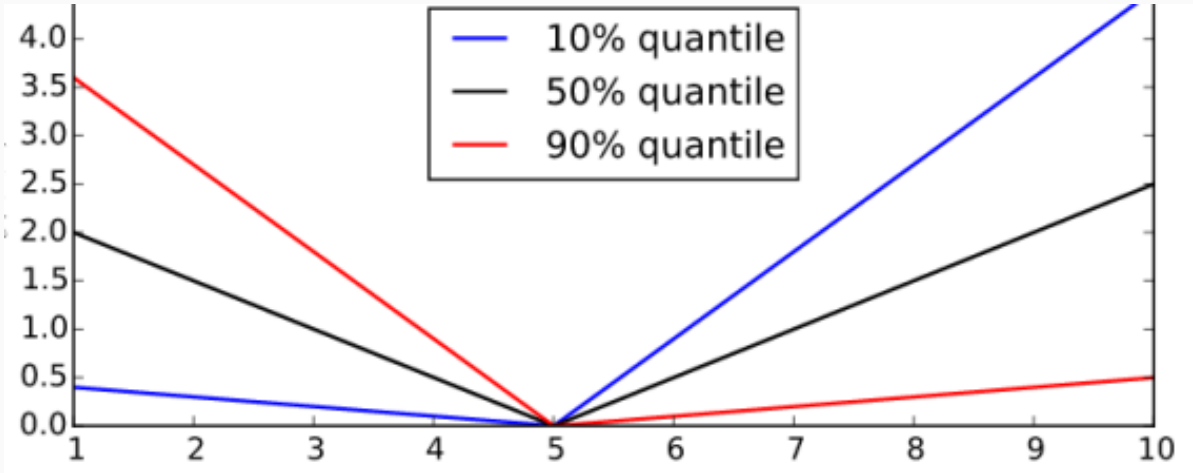
Les intervalles de confiance se croisent !

Approche fréquentiste

Définition

Soit $\xi \in \mathbb{R}$ et $\alpha \in (0,1)$, la fonction pinball est définie comme :

$$l_\alpha(\xi) = (\alpha - \mathbb{1}_{(\xi<0)})\xi$$



Approche fréquentiste

Proposition

Soient Y une variable aléatoire de fonction de répartition F et $\alpha \in (0, 1)$ alors,

$$\arg\min_{q \in \mathbb{R}} \mathbb{E}[l_\alpha(Y - q)]$$

est le quantile d'ordre α.

Approche fréquentiste

Définition

Le risque régularisé que nous souhaitons minimiser s'écrit :

$$\mathcal{R}_{reg}[f] := \mathbb{E}_{p(x,y)}[l_\alpha(y - f(x))] + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}}$$

Et le critère empirique associé à notre échantillon :

$$\mathcal{R}_{reg,emp}[f] := \frac{1}{m} \sum_{i=1}^{m} l_\alpha(y_i - f(x_i)) + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}}$$

Théroème [Takeuchi et al., 2006]

Le minimiseur de $\mathcal{R}_{reg,emp}$ est un estimateur de q_α.

MIA TOULOUSE

Approche fréquentiste

Minimisation du risque empirique :

$$\mathcal{R}_{reg,emp}[f] := \frac{1}{n}\sum_{i=1}^{n} l_\alpha(y_i - f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

Le minimiseur existe et est unique.

Il s'écrit :

$$f(x) = \sum_{i=1}^{n} \gamma_i k(x_i, x)$$

Approche fréquentiste

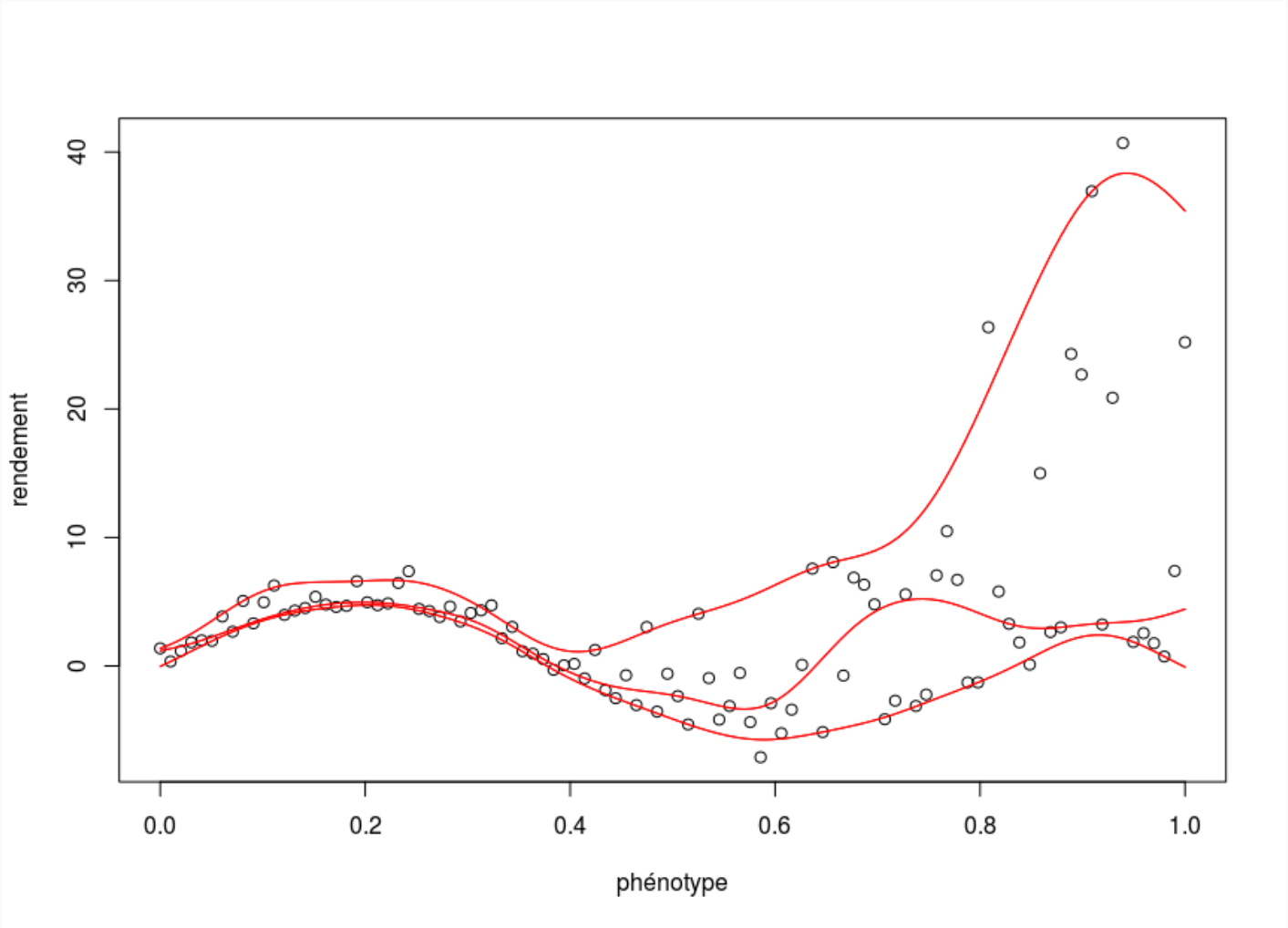
- Avantage : On se place dans un espace où f est combinaison linéaire de $k(x_i, .)$
- Le problème se réécrit :

$$\underset{\gamma}{\text{minimize}} \frac{1}{2}\gamma^T K\gamma - \gamma^T y \quad \textit{avec } K \in \mathbb{M}_n, \quad \gamma \in \mathbb{R}^n$$

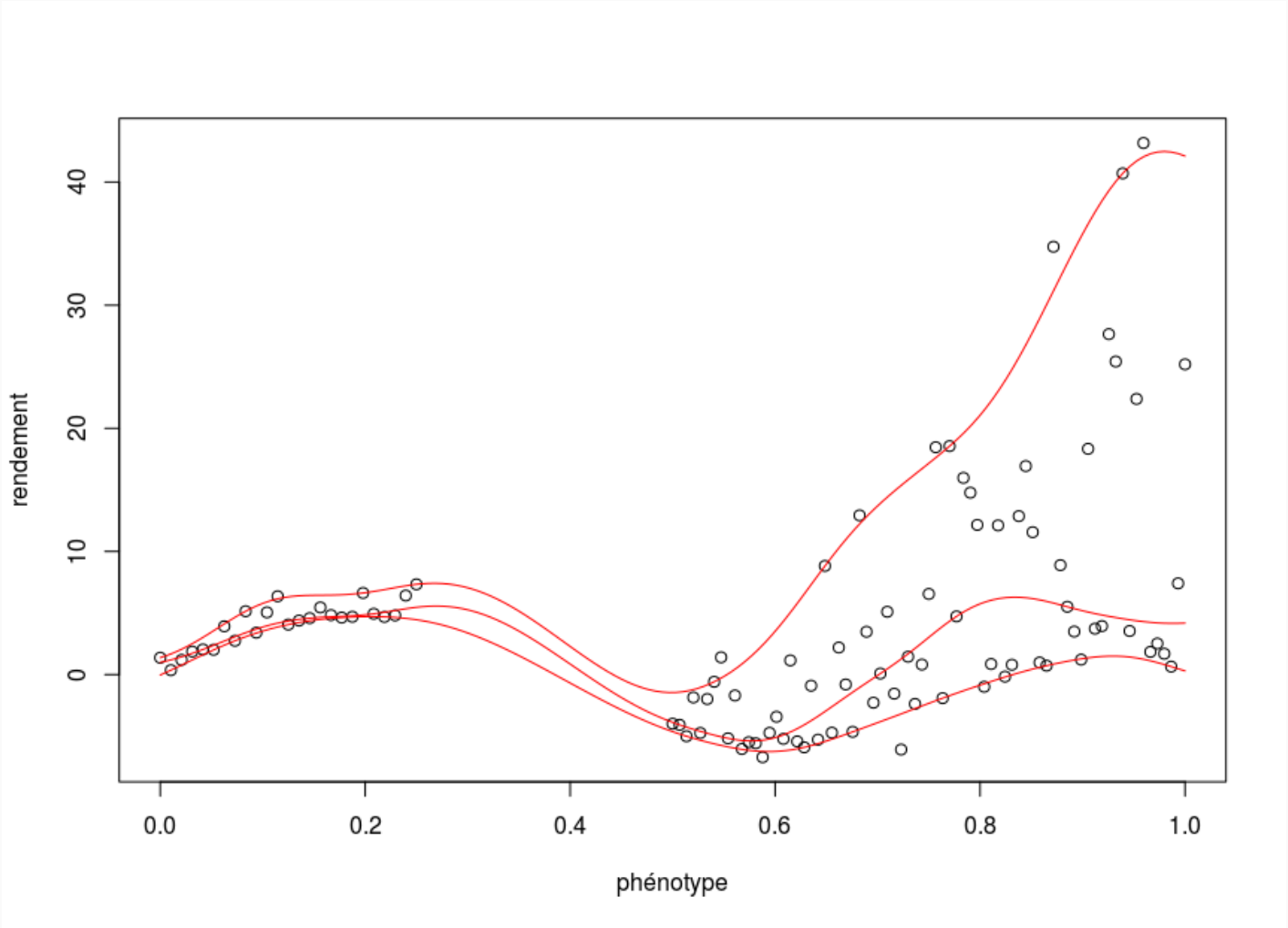
Sous les contraintes

$$\frac{1}{\lambda n}(\alpha - 1) \leq \gamma_i \leq \frac{1}{\lambda n}\alpha \;\; \forall \;\; 1 \leq i \leq n \;\; \textit{et} \;\; \sum_{i=1}^{n} \gamma_i = 0$$

Approche fréquentiste



Approche fréquentiste

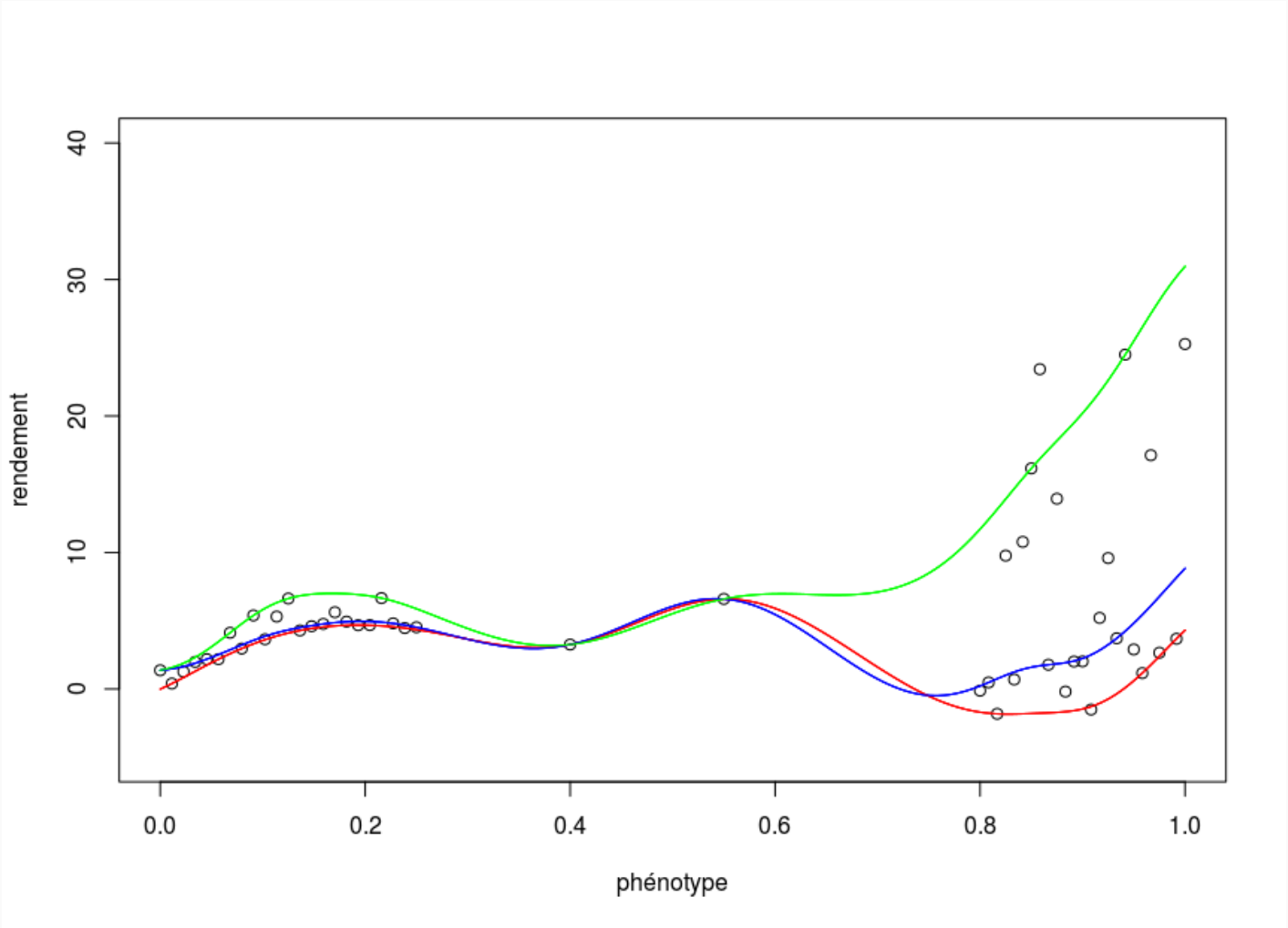


Approche fréquentiste

- Avantage principal : Nous n'avons pas besoin de concentrer l'information.

- Inconvénient principal : Nous n'avons pas d'indications sur l'incertitude locale.

Approche fréquentiste

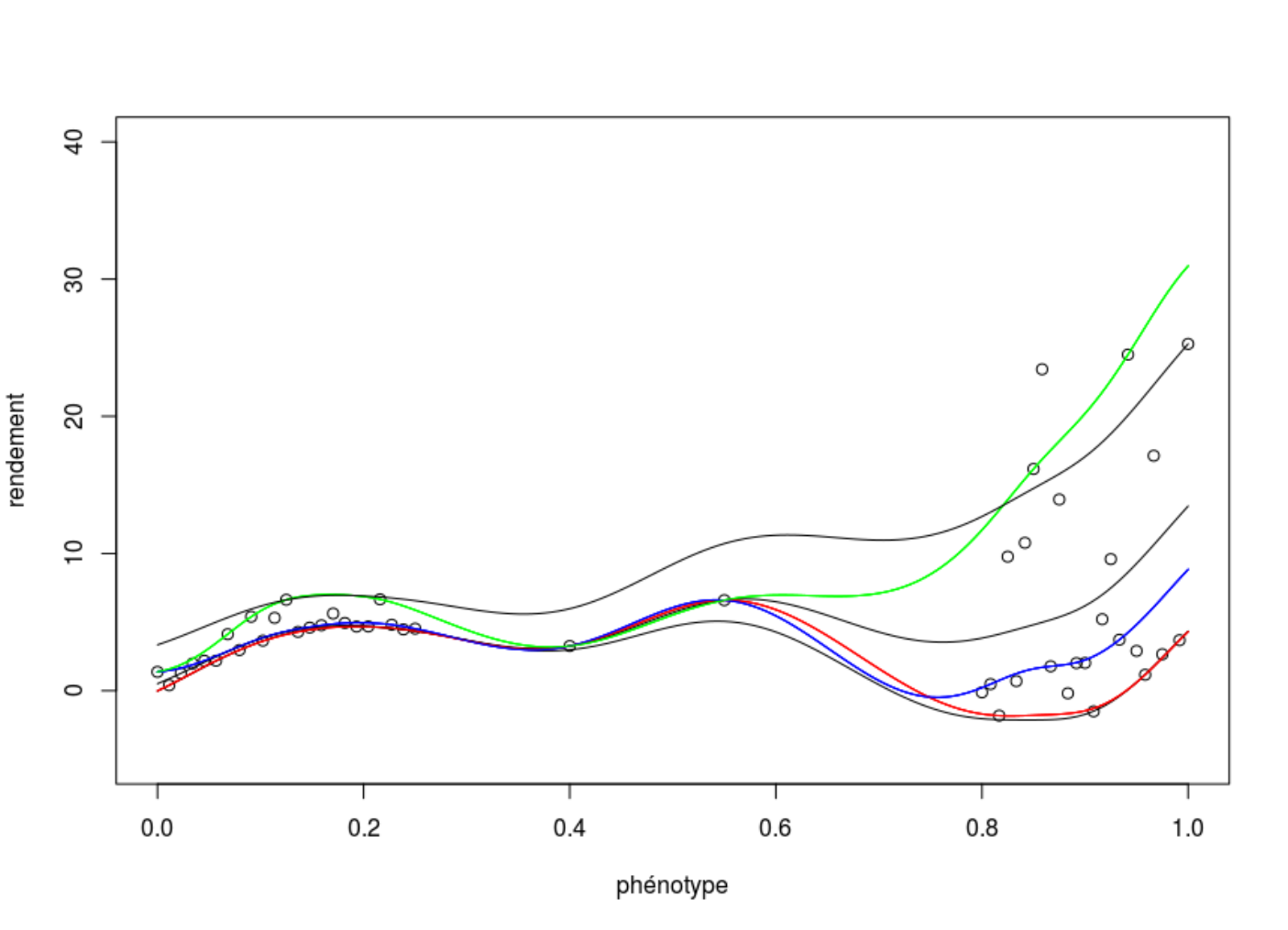


Approche fréquentiste

- Méthode tirée de [Sangnier et al., 2016][Alvarez et al., 2012] pour l'estimation multitaches.
- $$f : \mathbb{X} \to \mathbb{R}^D \quad \text{et non} \quad g : \mathbb{X} \to \mathbb{R}$$

  avec D le nombre de quantiles à estimer.

Approche fréquentiste



Différentes pistes de recherche

- Donner du sens aux intervalles de confiance dans le cas d'une régression multiquantile gaussienne.
- Trouver une manière de quantifier l'incertitude locale pour les méthodes de types SVM.
- Reconstruire les distributions à l'aide des quantiles pour avoir un autre point de comparaison.

MIA TOULOUSE

References I


Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. (2012).
Kernels for vector-valued functions: A review.
Foundations and Trends® in Machine Learning, 4(3):195–266.


Debaeke, P., Casadebaig, P., Mestries, E., Palleau, J.-P., Salvi, F., Bertoux, V., and Uyttewaal, V. (2011).
Simulation dynamique des interactions génotype x environnement x conduite de culture: application à l'évaluation variétale en tournesol.
http://www. inra. fr/ciag/revue.


Plumlee, M. and Tuo, R. (2014).
Building accurate emulators for stochastic simulations via quantile kriging.
Technometrics, 56(4):466–473.

MIA TOULOUSE

References II


Rasmussen, C. E. (2006).
Gaussian processes for machine learning.


Sangnier, M., Fercoq, O., and d'Alché Buc, F. (2016).
Joint quantile regression in vector-valued rkhss.
In *Advances in Neural Information Processing Systems*, pages 3693–3701.


Steinwart, I. and Christmann, A. (2008).
Support vector machines.
Springer Science & Business Media.


Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006).
Nonparametric quantile estimation.
Journal of Machine Learning Research, 7(Jul):1231–1264.

MIA TOULOUSE

Questions?

MIA TOULOUSE