Introduction à la statistique non paramétrique

Sandra Plancade

INRA (Institut National de Recherche en Agronomie)

ENSIIE - 2016

Introduction à la statistique non paramétrique

Les statistiques mathématiques

- On dispose d'un échantillon d'observations issu d'une population
- Statistique inférentielle : on veut estimer une fonction / quantité relative à cette population à partir de l'échantillon, en particulier
  - $(X_1, \ldots, X_n)$ i.i.d. de densité f : estimer f
  - $((X_1, Y_1), \ldots, (X_n, Y_n))$ i.i.d. et $b(x) = \mathbb{E}[Y_i | X_i = x]$: estimer b (fonction de régression)

Les statistiques mathématiques

- On dispose d'un échantillon d'observations issu d'une population
- Statistique inférentielle : on veut estimer une fonction / quantité relative à cette population à partir de l'échantillon, en particulier
  - $(X_1, \ldots, X_n)$ i.i.d. de densité f : estimer f
  - $((X_1, Y_1), \ldots, (X_n, Y_n))$ i.i.d. et $b(x) = \mathbb{E}[Y_i | X_i = x]$: estimer b (fonction de régression)

- Rappel : pour l'estimation d'une fonction f relative à une population, on appelle procédure d'estimation une application qui à un échantillon $\mathcal{E} \in \mathcal{X}^n$ associe une fonction $\hat{f}_n$:

$$\begin{aligned} \mathcal{X}^n &\rightarrow \mathcal{F} \\ \mathcal{E} &\rightarrow \hat{f}_n \end{aligned}$$

  - L'estimateur $\hat{f}_n$ ne doit pas dépendre de f, seulement de l'échantillon $\mathcal{E}$.
  - On appelle souvent "estimateur" la procédure d'estimation.

- Propriété des estimateurs :
  - Convergence : $\widehat{f}_n \rightarrow f$ quand la taille n d'échantillon augmente
  - Consistence (= absence de biais) : $\mathbb{E}[\widehat{f}_n] = f$
  - ...

- Tests

- Rappel : pour l'estimation d'une fonction f relative à une population, on appelle procédure d'estimation une application qui à un échantillon $\mathcal{E} \in \mathcal{X}^n$ associe une fonction $\hat{f}_n$:

$$\begin{aligned} \mathcal{X}^n &\rightarrow \mathcal{F} \\ \mathcal{E} &\rightarrow \hat{f}_n \end{aligned}$$

  - L'estimateur $\hat{f}_n$ ne doit pas dépendre de f, seulement de l'échantillon $\mathcal{E}$.
  - On appelle souvent "estimateur" la procédure d'estimation.
- Propriété des estimateurs :
  - Convergence : $\widehat{f}_n \rightarrow f$ quand la taille n d'échantillon augmente
  - Consistence (= absence de biais) : $\mathbb{E}[\widehat{f}_n] = f$
  - ...
- Tests

- Rappel : pour l'estimation d'une fonction f relative à une population, on appelle procédure d'estimation une application qui à un échantillon $\mathcal{E} \in \mathcal{X}^n$ associe une fonction $\hat{f}_n$:

$$\begin{aligned} \mathcal{X}^n &\rightarrow \mathcal{F} \\ \mathcal{E} &\rightarrow \hat{f}_n \end{aligned}$$

  - L'estimateur $\hat{f}_n$ ne doit pas dépendre de f, seulement de l'échantillon $\mathcal{E}$.
  - On appelle souvent "estimateur" la procédure d'estimation.
- Propriété des estimateurs :
  - Convergence : $\widehat{f}_n \rightarrow f$ quand la taille n d'échantillon augmente
  - Consistence (= absence de biais) : $\mathbb{E}[\widehat{f}_n] = f$
  - ...
- Tests

- Rappel : pour l'estimation d'une fonction f relative à une population, on appelle procédure d'estimation une application qui à un échantillon $\mathcal{E} \in \mathcal{X}^n$ associe une fonction $\hat{f}_n$:

$$\begin{aligned} \mathcal{X}^n &\rightarrow \mathcal{F} \\ \mathcal{E} &\rightarrow \hat{f}_n \end{aligned}$$

  - L'estimateur $\hat{f}_n$ ne doit pas dépendre de f, seulement de l'échantillon $\mathcal{E}$.
  - On appelle souvent "estimateur" la procédure d'estimation.
- Propriété des estimateurs :
  - Convergence : $\widehat{f}_n \rightarrow f$ quand la taille n d'échantillon augmente
  - Consistence (= absence de biais) : $\mathbb{E}[\widehat{f}_n] = f$
  - ...
- Tests

Statistique paramétrique : c'est quoi ?

- Statistiques "classiques"
- On suppose que la fonction à estimer est connue à un vecteur de paramètres près.
  Exemples :
  (1) Soit $(X_1, \ldots, X_n)$ échantillon i.i.d de distribution $\mathcal{N}(\mu, \sigma^2)$: estimer μ et σ.
  (2) Soit $((X_1, Y_1), \ldots, (X_n, Y_n))$ échantillon i.i.d. tel que $Y_i = f_\beta(X_i) + \varepsilon_i$ avec $\mathbb{E}[\varepsilon_i] = 0$ et $f_\beta(x) = \beta x$: estimer β.
- Remarque : souvent on sait que la fonction n'a pas **exactement** la forme paramétrique considérée mais on considère que cette approximation est raisonnable.
- Il est nécessaire de vérifier a postériori l'adéquation des données au modèle paramétrique.

Statistique paramétrique : c'est quoi ?

- Statistiques "classiques"
- On suppose que la fonction à estimer est connue à un vecteur de paramètres près.
  Exemples :
  (1) Soit $(X_1, \ldots, X_n)$ échantillon i.i.d de distribution $\mathcal{N}(\mu, \sigma^2)$: estimer μ et σ.
  (2) Soit $((X_1, Y_1), \ldots, (X_n, Y_n))$ échantillon i.i.d. tel que $Y_i = f_\beta(X_i) + \varepsilon_i$ avec $\mathbb{E}[\varepsilon_i] = 0$ et $f_\beta(x) = \beta x$: estimer β.
- Remarque : souvent on sait que la fonction n'a pas **exactement** la forme paramétrique considérée mais on considère que cette approximation est raisonnable.
- Il est nécessaire de vérifier a postériori l'adéquation des données au modèle paramétrique.

Statistique paramétrique : c'est quoi ?

- Statistiques "classiques"
- On suppose que la fonction à estimer est connue à un vecteur de paramètres près.
  Exemples :
  (1) Soit $(X_1, \ldots, X_n)$ échantillon i.i.d de distribution $\mathcal{N}(\mu, \sigma^2)$: estimer μ et σ.
  (2) Soit $((X_1, Y_1), \ldots, (X_n, Y_n))$ échantillon i.i.d. tel que $Y_i = f_\beta(X_i) + \varepsilon_i$ avec $\mathbb{E}[\varepsilon_i] = 0$ et $f_\beta(x) = \beta x$: estimer β.
- Remarque : souvent on sait que la fonction n'a pas **exactement** la forme paramétrique considérée mais on considère que cette approximation est raisonnable.
- Il est nécessaire de vérifier a postériori l'adéquation des données au modèle paramétrique.

Statistique non paramétrique : c'est quoi ?

- On ne suppose pas de forme paramétrique pour la fonction à estimer.
- On s'autorise toutes les formes a priori
  ou on se restreint à un espace de fonctions de dimension infini.
  Exple : $\mathcal{F} = \{f : [0,1] \to \mathbb{R}, f \text{ croissante } \}$.
- Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $b(x) = \mathbb{E}[Y|X = x]$

Histogramme
On se ramène à l'estimation
d'un nombre fini de paramètres
mais le nombre de paramètres
n'est pas fixé à l'avance

Statistique non paramétrique : c'est quoi ?

- On ne suppose pas de forme paramétrique pour la fonction à estimer.
- On s'autorise toutes les formes a priori
  ou on se restreint à un espace de fonctions de dimension infini.
  Exple : $\mathcal{F} = \{f : [0,1] \to \mathbb{R}, f \text{ croissante } \}$.
- Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $b(x) = \mathbb{E}[Y|X = x]$

Histogramme
On se ramène à l'estimation d'un nombre fini de paramètres **mais** le nombre de paramètres n'est pas fixé à l'avance

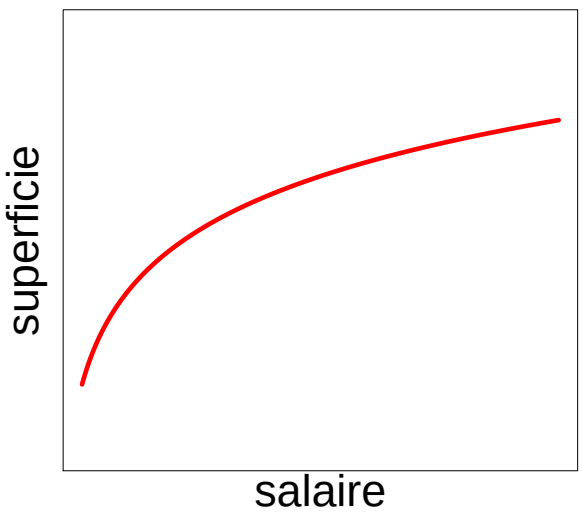
Statistique non paramétrique : c'est quoi ?

- On ne suppose pas de forme paramétrique pour la fonction à estimer.
- On s'autorise toutes les formes a priori
  ou on se restreint à un espace de fonctions de dimension infini.
  Exple : $\mathcal{F} = \{f : [0,1] \to \mathbb{R}, f \text{ croissante } \}$.
- Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $b(x) = \mathbb{E}[Y|X = x]$

Histogramme
On se ramène à l'estimation d'un nombre fini de paramètres **mais** le nombre de paramètres n'est pas fixé à l'avance

Statistique non paramétrique : c'est quoi ?

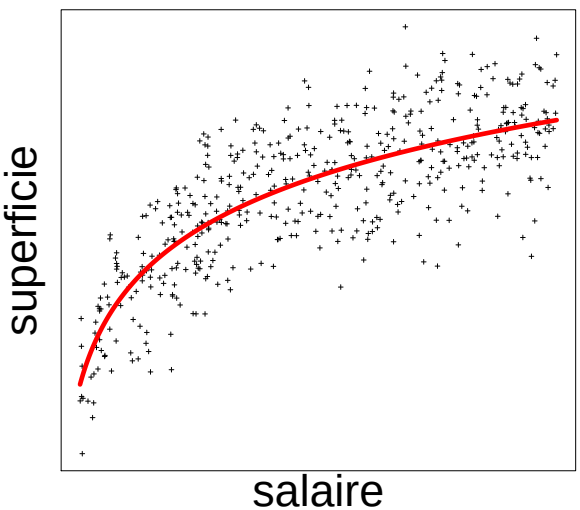
- On ne suppose pas de forme paramétrique pour la fonction à estimer.
- On s'autorise toutes les formes a priori
  ou on se restreint à un espace de fonctions de dimension infini.
  Exple : $\mathcal{F} = \{f : [0,1] \to \mathbb{R}, f \text{ croissante } \}$.
- Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $b(x) = \mathbb{E}[Y|X = x]$



Histogramme
On se ramène à l'estimation d'un nombre fini de paramètres **mais** le nombre de paramètres n'est pas fixé à l'avance

Statistique non paramétrique : c'est quoi ?

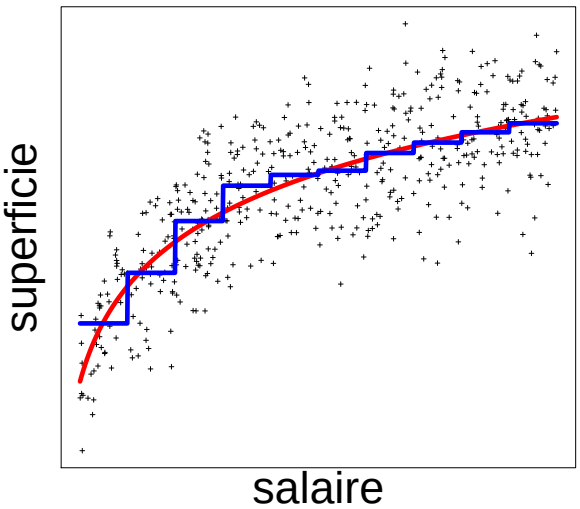
- On ne suppose pas de forme paramétrique pour la fonction à estimer.
- On s'autorise toutes les formes a priori
  ou on se restreint à un espace de fonctions de dimension infini.
  Exple : $\mathcal{F} = \{f : [0,1] \to \mathbb{R}, f \text{ croissante } \}$.
- Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $b(x) = \mathbb{E}[Y|X = x]$



Histogramme
On se ramène à l'estimation d'un nombre fini de paramètres **mais** le nombre de paramètres n'est pas fixé à l'avance

Statistique non paramétrique : c'est quoi ?

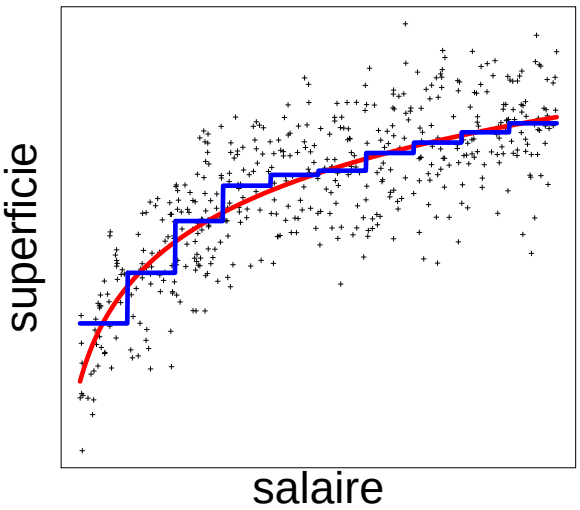
- On ne suppose pas de forme paramétrique pour la fonction à estimer.
- On s'autorise toutes les formes a priori
  ou on se restreint à un espace de fonctions de dimension infini.
  Exple : $\mathcal{F} = \{f : [0,1] \to \mathbb{R}, f \text{ croissante } \}$.
- Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $b(x) = \mathbb{E}[Y|X = x]$



Histogramme
On se ramène à l'estimation d'un nombre fini de paramètres **mais** le nombre de paramètres n'est pas fixé à l'avance

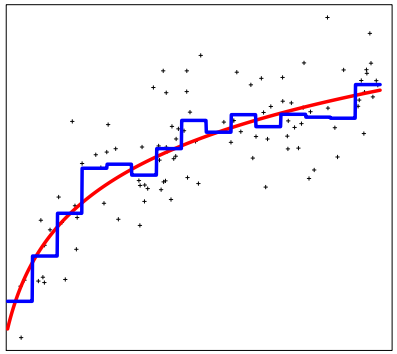
Statistique non paramétrique : c'est quoi ?

- On ne suppose pas de forme paramétrique pour la fonction à estimer.
- On s'autorise toutes les formes a priori
  ou on se restreint à un espace de fonctions de dimension infini.
  Exple : $\mathcal{F} = \{f : [0,1] \to \mathbb{R}, f \text{ croissante } \}$.
- Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $b(x) = \mathbb{E}[Y|X = x]$

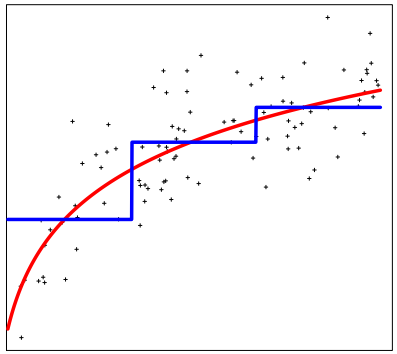


Histogramme
On se ramène à l'estimation d'un nombre fini de paramètres **mais** le nombre de paramètres n'est pas fixé à l'avance

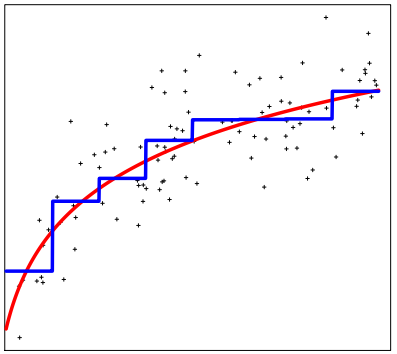
Choix du nombre de paramètres de l'histogramme



Trop de paramètres



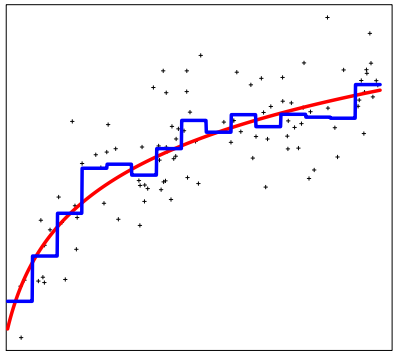
Pas assez de paramètres



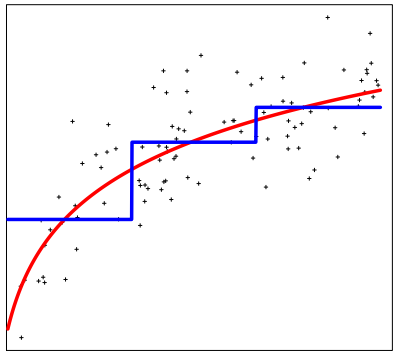
Compromis

- Trop de paramètres : pas assez de données pour estimer chaque paramètre.
  $\hookrightarrow$ Grande variance
- Pas assez de paramètres : modèle trop imprécis
  $\hookrightarrow$ Grand biais

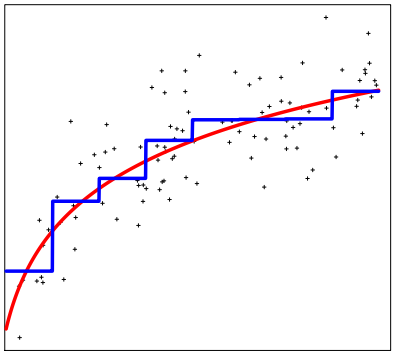
Choix du nombre de paramètres de l'histogramme



Trop de paramètres



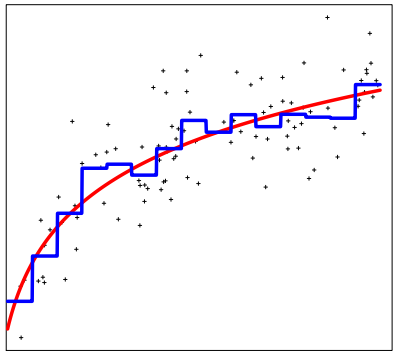
Pas assez de paramètres



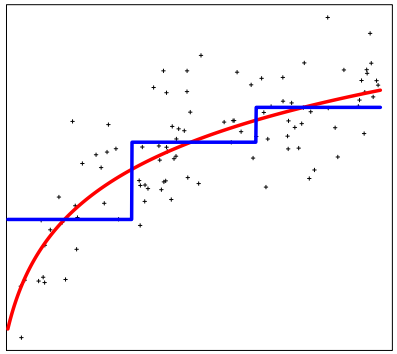
Compromis

- Trop de paramètres : pas assez de données pour estimer chaque paramètre.
  $\hookrightarrow$ Grande variance
- Pas assez de paramètres : modèle trop imprécis
  $\hookrightarrow$ Grand biais

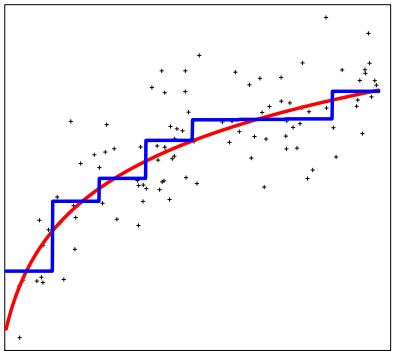
Choix du nombre de paramètres de l'histogramme



Trop de paramètres



Pas assez de paramètres



Compromis

- Trop de paramètres : pas assez de données pour estimer chaque paramètre.
  $\hookrightarrow$ Grande variance
- Pas assez de paramètres : modèle trop imprécis
  $\hookrightarrow$ Grand biais

Biais et variance

Contexte

- $\mathcal{E}$: échantillon d'observation de taille n tiré dans une population de distribution $\mathcal{F}$ (inconnue)

  Exple : $\mathcal{E} = (X_i, Y_i)_{i=1,\dots,n}$

- On veut estimer une fonction f

  Exple : $f(x) = \mathbb{E}[Y_i | X_i = x]$

- On dispose d'une procédure d'estimation qui à un échantillon $\mathcal{E}$ associe un estimateur $\hat{f}_n$.

  Exple : régression linéaire, histogramme à 4 morceaux, etc

  Rq : si on tire un nouvel échantillon, l'estimateur sera différent.

- On considère une distance fonctionnelle $d(f, g)$, l'erreur d'estimation est alors quantifiée par $d(f, \hat{f}_n)$.

  Exple : $(f(x) - g(x))^2$, $\|f - g\|_{L^2}$

Biais et variance

Contexte

- $\mathcal{E}$: échantillon d'observation de taille n tiré dans une population de distribution $\mathcal{F}$ (inconnue)

  Exple : $\mathcal{E} = (X_i, Y_i)_{i=1,\dots,n}$

- On veut estimer une fonction f

  Exple : $f(x) = \mathbb{E}[Y_i | X_i = x]$

- On dispose d'une procédure d'estimation qui à un échantillon $\mathcal{E}$ associe un estimateur $\hat{f}_n$.

  Exple : régression linéaire, histogramme à 4 morceaux, etc

  Rq : si on tire un nouvel échantillon, l'estimateur sera différent.

- On considère une distance fonctionnelle $d(f, g)$, l'erreur d'estimation est alors quantifiée par $d(f, \hat{f}_n)$.

  Exple : $(f(x) - g(x))^2$, $\|f - g\|_{L^2}$

Biais et variance

Contexte

- $\mathcal{E}$: échantillon d'observation de taille n tiré dans une population de distribution $\mathcal{F}$ (inconnue)

  Exple : $\mathcal{E} = (X_i, Y_i)_{i=1,\dots,n}$

- On veut estimer une fonction f

  Exple : $f(x) = \mathbb{E}[Y_i | X_i = x]$

- On dispose d'une procédure d'estimation qui à un échantillon $\mathcal{E}$ associe un estimateur $\hat{f}_n$.

  Exple : régression linéaire, histogramme à 4 morceaux, etc

  Rq : si on tire un nouvel échantillon, l'estimateur sera différent.

- On considère une distance fonctionnelle $d(f, g)$, l'erreur d'estimation est alors quantifiée par $d(f, \hat{f}_n)$.

  Exple : $(f(x) - g(x))^2$, $\|f - g\|_{L^2}$

Biais et variance

Contexte

- $\mathcal{E}$: échantillon d'observation de taille n tiré dans une population de distribution $\mathcal{F}$ (inconnue)

  Exple : $\mathcal{E} = (X_i, Y_i)_{i=1,\dots,n}$
- On veut estimer une fonction f

  Exple : $f(x) = \mathbb{E}[Y_i | X_i = x]$
- On dispose d'une procédure d'estimation qui à un échantillon $\mathcal{E}$ associe un estimateur $\hat{f}_n$.

  Exple : régression linéaire, histogramme à 4 morceaux, etc

  Rq : si on tire un nouvel échantillon, l'estimateur sera différent.
- On considère une distance fonctionnelle $d(f, g)$, l'erreur d'estimation est alors quantifiée par $d(f, \hat{f}_n)$.

  Exple : $(f(x) - g(x))^2$, $\|f - g\|_{L^2}$

Biais et variance (2)

- Le Biais $d(f, \mathbb{E}[\hat{f}_n])$ quantifie la distance entre la fonction à estimer et l'estimateur moyen dans le modèle choisi.

  Le modèle considéré est-il suffisamment correct pour approcher f ?

- La Variance $\mathbb{E}\left[d(\hat{f}_n, \mathbb{E}[\hat{f}_n])\right]$ quantifie la variabilité de l'estimateur par rapport au tirage d'un échantillon.

  Le modèle considéré peut-il être correctement estimé à partir d'un échantillon de taille n tiré dans la population ?

- Le biais et la variance ne dépendent pas de l'échantillon d'observation particulier.
- Usuellement, on cherchera un compromis entre le biais et la variance

Biais et variance (2)

- Le Biais $d(f, \mathbb{E}[\hat{f}_n])$ quantifie la distance entre la fonction à estimer et l'estimateur moyen dans le modèle choisi.

  Le modèle considéré est-il suffisamment correct pour approcher f ?

- La Variance $\mathbb{E}\left[d(\hat{f}_n, \mathbb{E}[\hat{f}_n])\right]$ quantifie la variabilité de l'estimateur par rapport au tirage d'un échantillon.

  Le modèle considéré peut-il être correctement estimé à partir d'un échantillon de taille n tiré dans la population ?

- Le biais et la variance ne dépendent pas de l'échantillon d'observation particulier.
- Usuellement, on cherchera un compromis entre le biais et la variance

Biais et variance (2)

- Le Biais $d(f, \mathbb{E}[\hat{f}_n])$ quantifie la distance entre la fonction à estimer et l'estimateur moyen dans le modèle choisi.

  Le modèle considéré est-il suffisamment correct pour approcher f ?

- La Variance $\mathbb{E}\left[d(\hat{f}_n, \mathbb{E}[\hat{f}_n])\right]$ quantifie la variabilité de l'estimateur par rapport au tirage d'un échantillon.

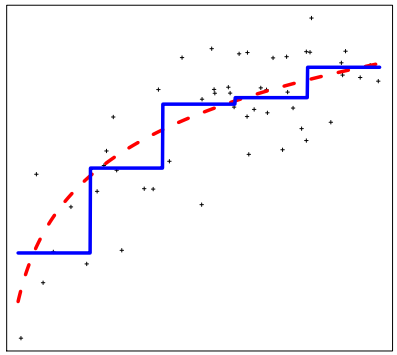
  Le modèle considéré peut-il être correctement estimé à partir d'un échantillon de taille n tiré dans la population ?

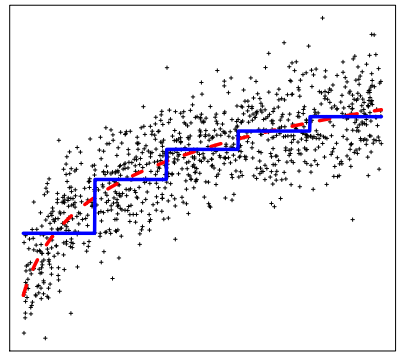
- Le biais et la variance ne dépendent pas de l'échantillon d'observation particulier.
- Usuellement, on cherchera un compromis entre le biais et la variance

Compromis biais-variance

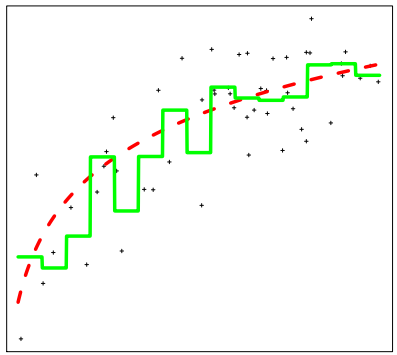
$n = 50$ observations $n = 1000$ observations

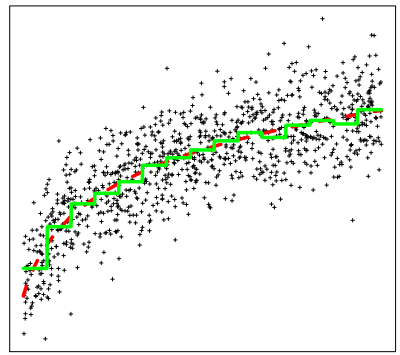
5 intervalles





20 intervalles



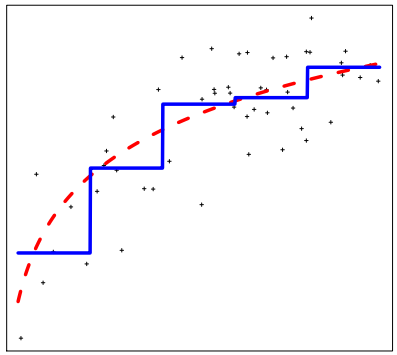


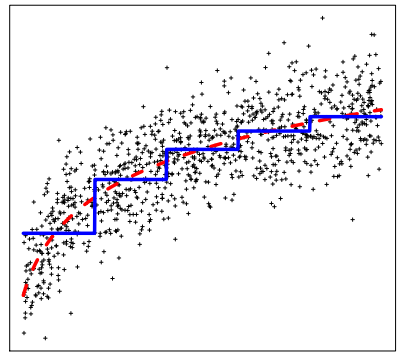
- La taille d'histogramme optimale dépend de la taille d'échantillon.

Compromis biais-variance

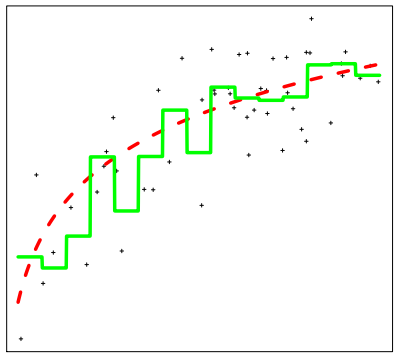
$n = 50$ observations $n = 1000$ observations

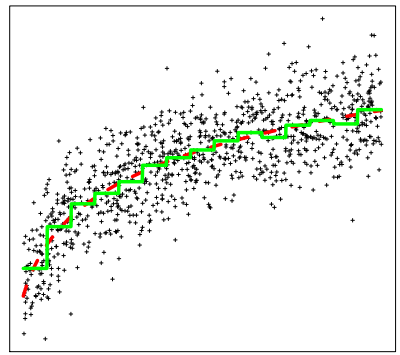
5 intervalles





20 intervalles



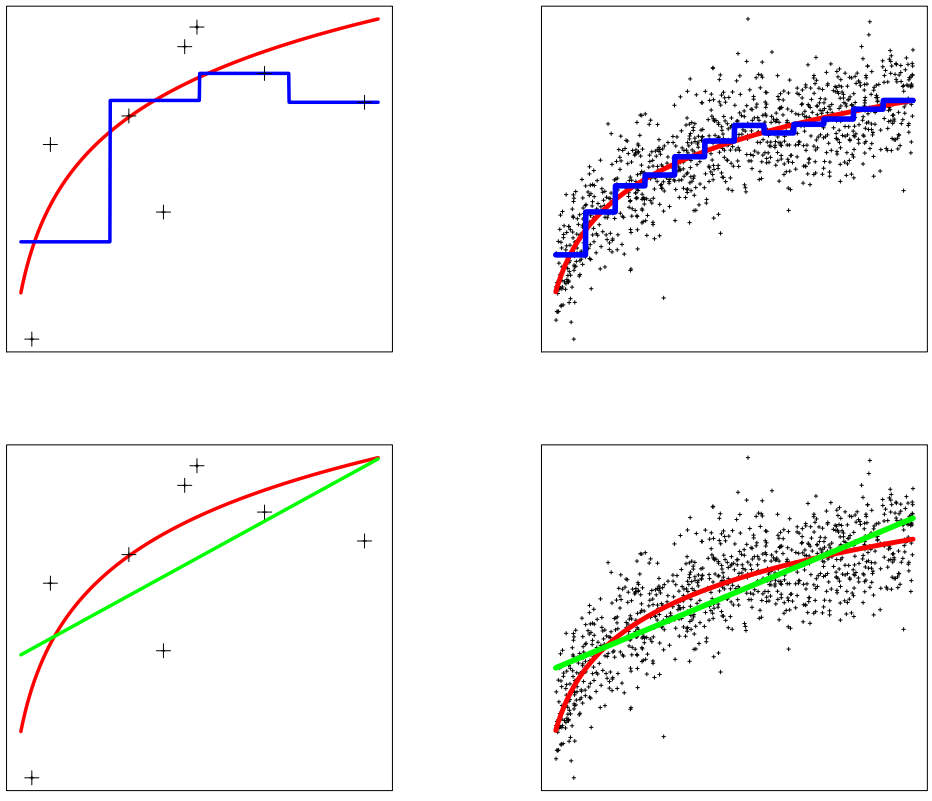


- La taille d'histogramme optimale dépend de la taille d'échantillon.

Paramétrique ou non paramétrique (1) ?

- Comparaison entre histogramme et modèle linéaire

$n = 8$ $n = 1000$

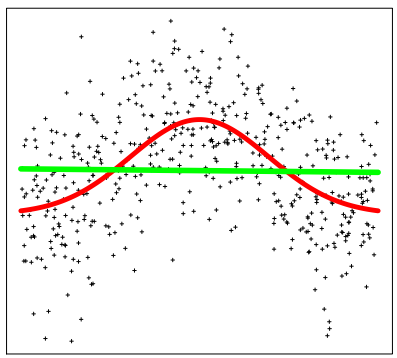


- Peu d'observations : modèle non paramétrique mal estimé
- Beaucoup d'observations : modèle non-paramétrique plus précis

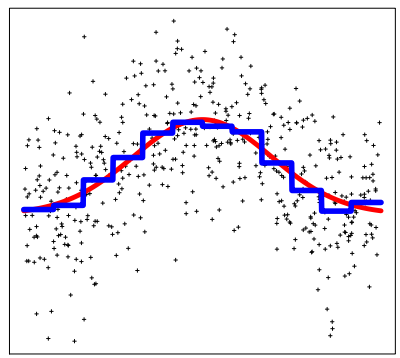
Paramétrique ou non paramétrique (2) ?

- Supposons que la fonction à estimer soit uni-modale (mais qu'on l'ignore !), et qu'on considère un modèle paramétrique linéaire.
- Comparaison entre histogramme et modèle linéaire

Modèle linéaire



Histogramme

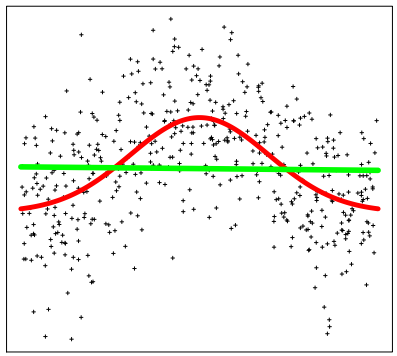


- Modèle paramétrique "très faux" : impossible d'obtenir une bonne estimation, quelle que soit la taille d'échantillon.

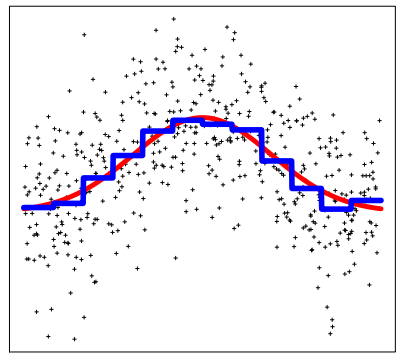
Paramétrique ou non paramétrique (2) ?

- Supposons que la fonction à estimer soit uni-modale (mais qu'on l'ignore !), et qu'on considère un modèle paramétrique linéaire.
- Comparaison entre histogramme et modèle linéaire

Modèle linéaire



Histogramme



- Modèle paramétrique "très faux" : impossible d'obtenir une bonne estimation, quelle que soit la taille d'échantillon.

Résumé

- Quand utiliser des méthodes d'estimation non-paramétriques ?
  - Quand les modèles paramétriques envisagés n'ajuste pas bien les données
  - Quand la taille d'échantilllon est suffisante.
- Avantage du non-paramétrique : moins d'a priori sur les données
- Inconvénient du non-paramétrique : vitesse de convergence plus lente i.e. il faut plus d'observations pour obtenir une même qualité d'approximation.

Résumé

- Quand utiliser des méthodes d'estimation non-paramétriques ?
  - Quand les modèles paramétriques envisagés n'ajuste pas bien les données
  - Quand la taille d'échantilllon est suffisante.
- Avantage du non-paramétrique : moins d'a priori sur les données
- Inconvénient du non-paramétrique : vitesse de convergence plus lente i.e. il faut plus d'observations pour obtenir une même qualité d'approximation.

Résumé

- Quand utiliser des méthodes d'estimation non-paramétriques ?
  - Quand les modèles paramétriques envisagés n'ajuste pas bien les données
  - Quand la taille d'échantilllon est suffisante.
- Avantage du non-paramétrique : moins d'a priori sur les données
- Inconvénient du non-paramétrique : vitesse de convergence plus lente i.e. il faut plus d'observations pour obtenir une même qualité d'approximation.

Résumé

- Quand utiliser des méthodes d'estimation non-paramétriques ?
  - Quand les modèles paramétriques envisagés n'ajuste pas bien les données
  - Quand la taille d'échantilllon est suffisante.
- Avantage du non-paramétrique : moins d'a priori sur les données
- Inconvénient du non-paramétrique : vitesse de convergence plus lente i.e. il faut plus d'observations pour obtenir une même qualité d'approximation.

Fonctions de répartition

Fonction de répartition (*cumulative distribution function*)

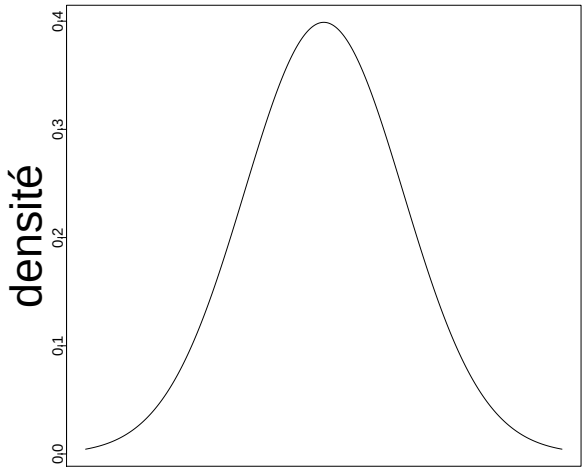
La fonction de répartition F d'une v.a. X réelle est définie par :

$$\begin{array}{rccc} F: & \mathbb{R} & \to & [0,1] \\ & x & \to & \mathbb{P}[X \leq x] \end{array}$$

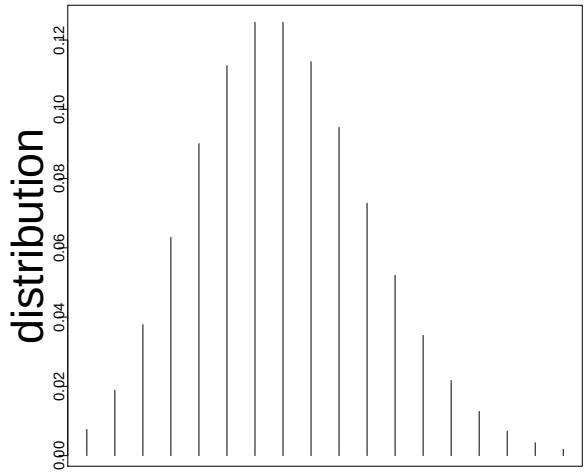
- F est croissante
- Si X est une v.a. continue et possède une densité f, alors F est dérivable et $F' = f$.
- F est définie pour toute variable aléatoire réelle, elle est càd-lag (continue à droite, dérivable à gauche).

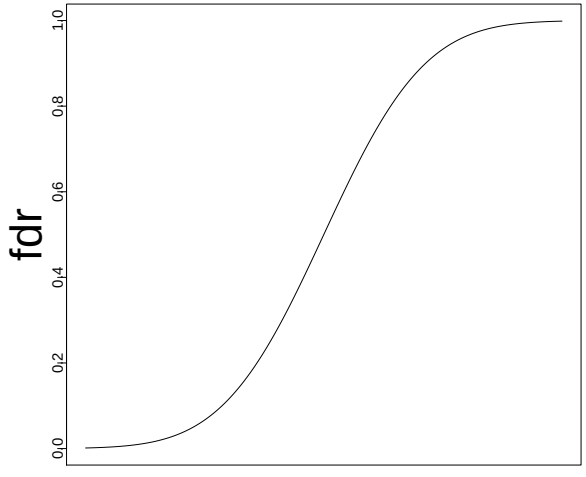
Exemples

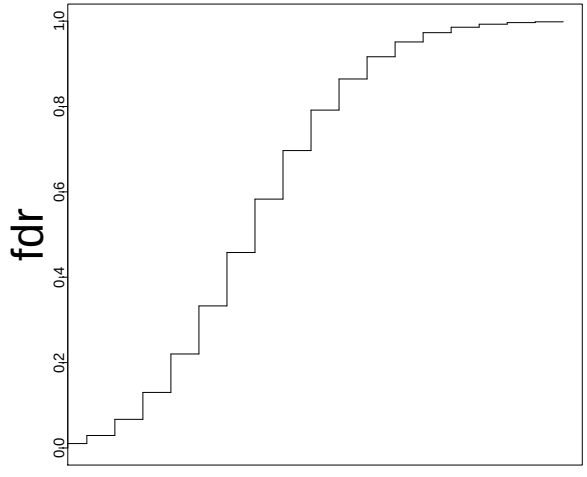
v.a. continue



v.a. discrète





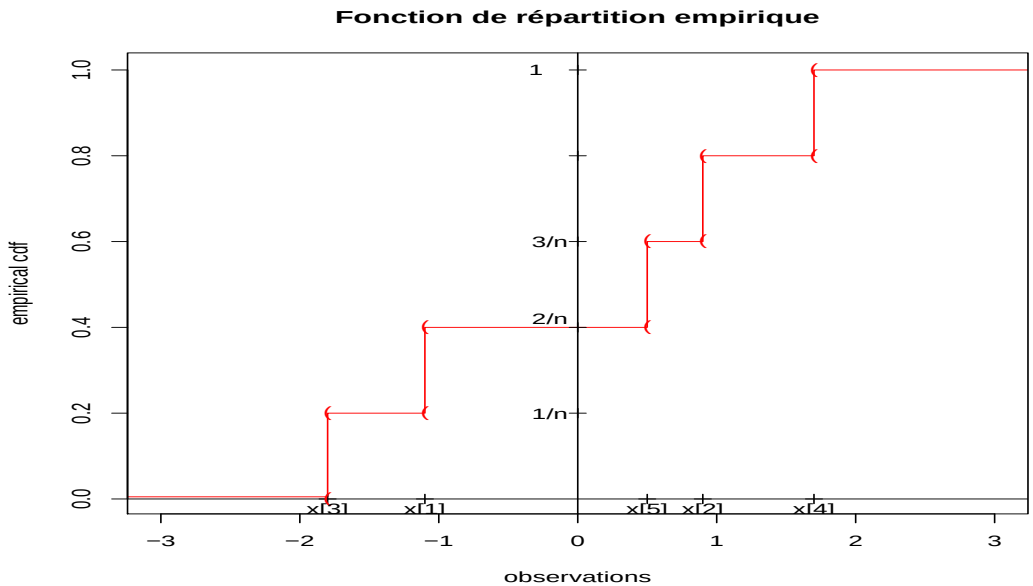


Estimer une fonction de répartition

On observe $X_1, \ldots, X_n$ variables aléatoires (v.a.) réelles, i.i.d. de fonction de répartition F. L'estimateur naturel de la fdr F est la fdr empirique $\hat{F}_n$ définie par

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x} = \frac{1}{n} \#\{i, X_i \leq x\}$$

C'est un estimateur non paramétrique de la fdr F.



→ Qualité de cet estimateur ?

Propriétés ponctuelles de $\hat{F}_n(x)$ (*i.e.* pour x fixé)

- $\hat{F}_n$ est un estimateur sans biais de F :

$$\mathbb{E}[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{1}_{X_i \leq x}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X_i \leq x) = F(x)$$

- Conséquence : erreur en moyenne quadratique

$$\begin{aligned}
\mathbb{E}[(\hat{F}_n(x) - F(x))^2] &= \mathbb{E}[(\hat{F}_n(x) - \mathbb{E}[\hat{F}_n(x)])^2] \\
&= \mathbb{V}\mathrm{ar}(\hat{F}_n(x)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(1_{X_i \leq x}) \\
&= \frac{1}{n}\mathbb{V}\mathrm{ar}(1_{X_1 \leq x}) \\
&= \frac{F(x)(1-F(x))}{n} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

- La vitesse de convergence de $\hat{F}_n$ pour le risque quadratique en un point fixé est $1/n$

Propriétés ponctuelles de $\hat{F}_n(x)$ (*i.e.* pour x fixé)

- $\hat{F}_n$ est un estimateur sans biais de F :

$$\mathbb{E}[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{1}_{X_i \leq x}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X_i \leq x) = F(x)$$

- Conséquence : erreur en moyenne quadratique

$$\begin{aligned}
\mathbb{E}[(\hat{F}_n(x) - F(x))^2] &= \mathbb{E}[(\hat{F}_n(x) - \mathbb{E}[\hat{F}_n(x)])^2] \\
&= \mathbb{V}\mathrm{ar}(\hat{F}_n(x)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(1_{X_i \leq x}) \\
&= \frac{1}{n}\mathbb{V}\mathrm{ar}(1_{X_1 \leq x}) \\
&= \frac{F(x)(1-F(x))}{n} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

- La vitesse de convergence de $\hat{F}_n$ pour le risque quadratique en un point fixé est $1/n$

Propriétés ponctuelles de $\hat{F}_n(x)$ (*i.e.* pour x fixé)

- $\hat{F}_n$ est un estimateur sans biais de F :

$$\mathbb{E}[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{1}_{X_i \leq x}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X_i \leq x) = F(x)$$

- Conséquence : erreur en moyenne quadratique

$$\begin{aligned}
\mathbb{E}[(\hat{F}_n(x) - F(x))^2] &= \mathbb{E}[(\hat{F}_n(x) - \mathbb{E}[\hat{F}_n(x)])^2] \\
&= \mathbb{V}\mathrm{ar}(\hat{F}_n(x)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(1_{X_i \leq x}) \\
&= \frac{1}{n}\mathbb{V}\mathrm{ar}(1_{X_1 \leq x}) \\
&= \frac{F(x)(1-F(x))}{n} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

- La vitesse de convergence de $\hat{F}_n$ pour le risque quadratique en un point fixé est $1/n$

Propriétés ponctuelles de $\hat{F}_n(x)$ (*i.e.* pour x fixé)

- $\hat{F}_n$ est un estimateur sans biais de F :

$$\mathbb{E}[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{1}_{X_i \leq x}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X_i \leq x) = F(x)$$

- Conséquence : erreur en moyenne quadratique

$$\begin{aligned}
\mathbb{E}[(\hat{F}_n(x) - F(x))^2] &= \mathbb{E}[(\hat{F}_n(x) - \mathbb{E}[\hat{F}_n(x)])^2] \\
&= \mathbb{V}\mathrm{ar}(\hat{F}_n(x)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(1_{X_i \leq x}) \\
&= \frac{1}{n}\mathbb{V}\mathrm{ar}(1_{X_1 \leq x}) \\
&= \frac{F(x)(1-F(x))}{n} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

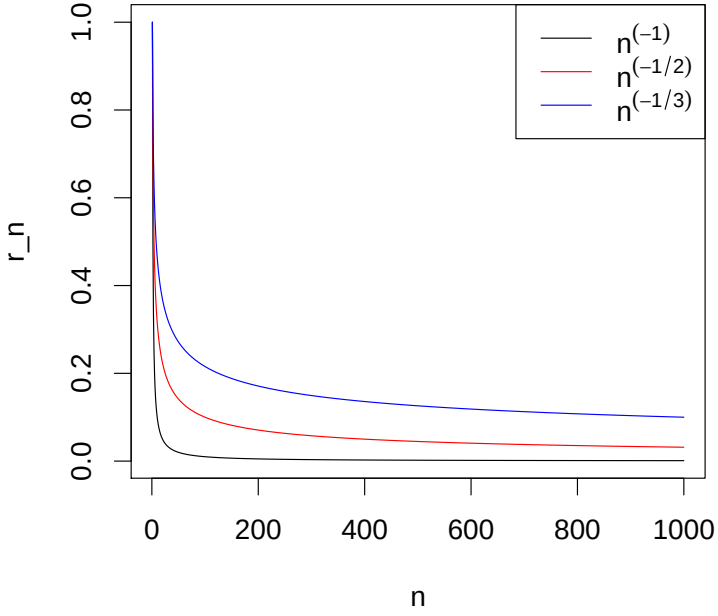
- La vitesse de convergence de $\hat{F}_n$ pour le risque quadratique en un point fixé est $1/n$

Vitesse de convergence

- $\hat{f}_n$ un estimateur d'une fonction f calculé à partir d'un échantillon de taille n
- d une distance fonctionnelle,
- $(r_n)_{n\in\mathbb{N}}$ une suite positive décroissante

On dit que $\hat{f}_n$ converge vers f à la vitesse r_n pour la distance d si il existe une constante C telle que

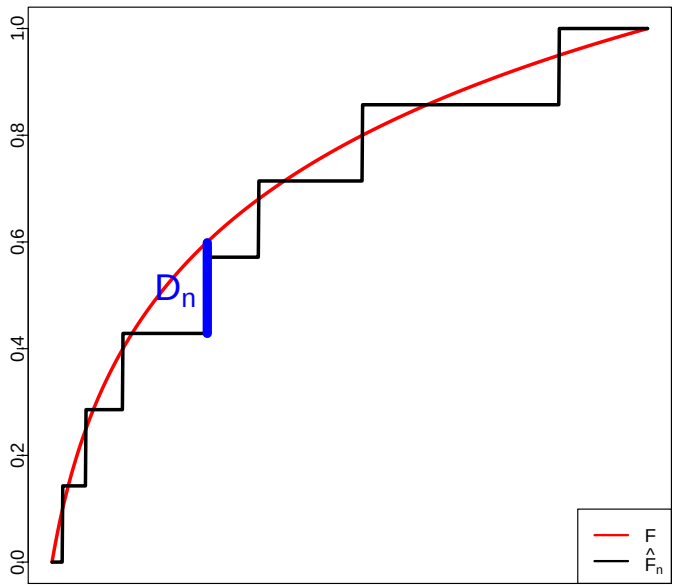
$$\mathbb{E}\left[d(\hat{f}_n, f)\right] \leq C r_n$$



- $1/n =$ vitesse de convergence des estimateurs paramétriques.
- En non paramétrique, la vitesse est usuellement moins bonne, sauf pour la fdr.

Propriétés uniformes de $\hat{F}_n$ (i.e. pour le "sup sur x")

- Soit $D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$.

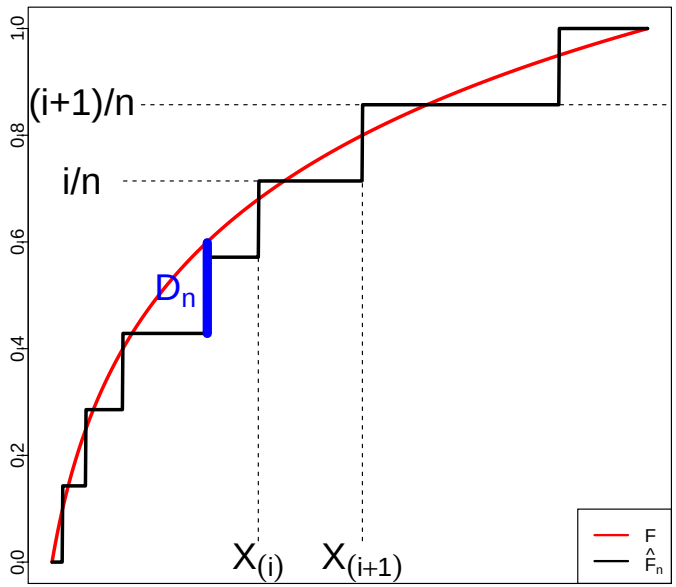


- D_n est calculable car le sup est nécessairement atteint en un point X_i.

$$D_n = \max_{1 \leq i \leq n} \left\{ \left| F(X_{(i)}) - \frac{i}{n} \right|, \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right\}$$

Propriétés uniformes de $\hat{F}_n$ (i.e. pour le "sup sur x")

- Soit $D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$.



- D_n est calculable car le sup est nécessairement atteint en un point X_i.

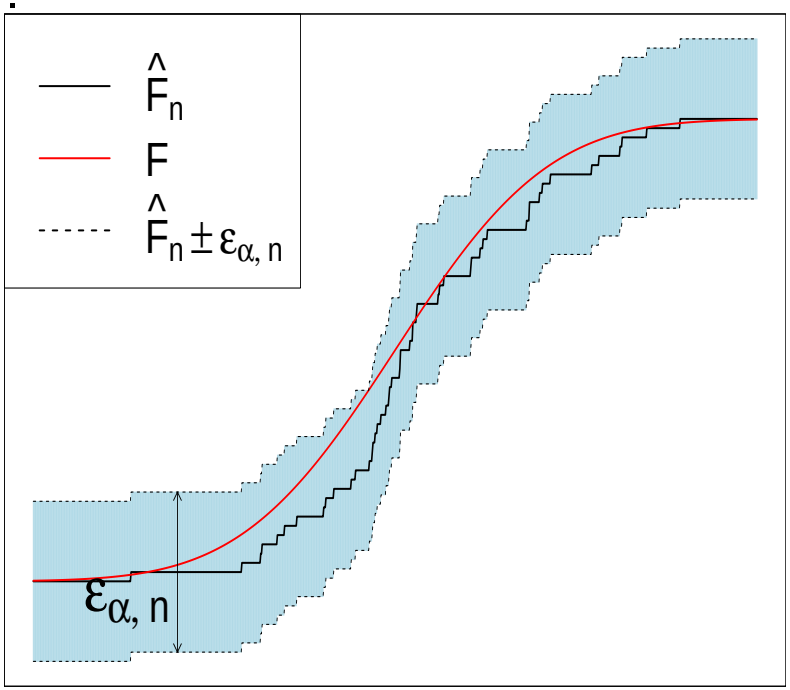
$$D_n = \max_{1 \leq i \leq n} \left\{ \left| F(X_{(i)}) - \frac{i}{n} \right|, \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right\}$$

Construction de *bandes de confiance*

Def : $\left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right]$ est une bande de confiance de niveau α pour F si :

$$\mathbb{P}\left[D_n > \varepsilon_{\alpha,n}\right] = \mathbb{P}\left[F(x) \in \left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right], \forall x \in \mathbb{R}\right] \geq 1-\alpha$$

.



Construction de bandes de confiance

- Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW)

$$\forall n \in \mathbb{N}, \forall \varepsilon > 0, \quad \mathbb{P}(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

- Conséquence :

$$\begin{aligned} \mathbb{P}(F(x) \in [\hat{F}_n(x) - \varepsilon; \hat{F}_n(x) + \varepsilon]) &= 1 - \mathbb{P}(|\hat{F}_n(x) - F(x)| > \varepsilon) \\ &\geq 1 - \mathbb{P}(\sup_x |\hat{F}_n(x) - F(x)| > \varepsilon) \\ &\geq 1 - 2e^{-2n\varepsilon^2}. \end{aligned}$$

- Pour un niveau de seuil $\alpha > 0$, soit $\varepsilon_{\alpha,n}$ tel que $2e^{-2n\varepsilon^2} = \alpha$, *i.e.* $\varepsilon = \sqrt{\log(2/\alpha)/(2n)}$. Alors $\left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right]$ est une bande de confiance de niveau α pour $F(x)$.

Construction de bandes de confiance

- Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW)

$$\forall n \in \mathbb{N}, \forall \varepsilon > 0, \quad \mathbb{P}(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

- Conséquence :

$$\begin{aligned} \mathbb{P}(F(x) \in [\hat{F}_n(x) - \varepsilon; \hat{F}_n(x) + \varepsilon]) &= 1 - \mathbb{P}(|\hat{F}_n(x) - F(x)| > \varepsilon) \\ &\geq 1 - \mathbb{P}(\sup_{x} |\hat{F}_n(x) - F(x)| > \varepsilon) \\ &\geq 1 - 2e^{-2n\varepsilon^2}. \end{aligned}$$

- Pour un niveau de seuil $\alpha > 0$, soit $\varepsilon_{\alpha,n}$ tel que $2e^{-2n\varepsilon^2} = \alpha$, *i.e.* $\varepsilon = \sqrt{\log(2/\alpha)/(2n)}$. Alors $\left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right]$ est une bande de confiance de niveau α pour $F(x)$.

Construction de bandes de confiance

- Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW)

$$\forall n \in \mathbb{N}, \forall \varepsilon > 0, \quad \mathbb{P}(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

- Conséquence :

$$\begin{aligned} \mathbb{P}(F(x) \in [\hat{F}_n(x) - \varepsilon; \hat{F}_n(x) + \varepsilon]) &= 1 - \mathbb{P}(|\hat{F}_n(x) - F(x)| > \varepsilon) \\ &\geq 1 - \mathbb{P}(\sup_x |\hat{F}_n(x) - F(x)| > \varepsilon) \\ &\geq 1 - 2e^{-2n\varepsilon^2}. \end{aligned}$$

- Pour un niveau de seuil $\alpha > 0$, soit $\varepsilon_{\alpha,n}$ tel que $2e^{-2n\varepsilon^2} = \alpha$, *i.e.* $\varepsilon = \sqrt{\log(2/\alpha)/(2n)}$. Alors $\left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right]$ est une bande de confiance de niveau α pour $F(x)$.

Convergence en distribution

- Convergence vers la distribution de Kolmogorov : Quel que soit F, D_n converge en loi vers la distribution de Kolmogorov α_K indépendante de F

$$\mathbb{P}\left[D_n > \frac{c}{\sqrt{n}}\right] \xrightarrow[n\to\infty]{} \alpha_K(c) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2c^2} \quad (1)$$

- Application : test d'adéquation à une distribution donnée
  - On dispose d'un échantillon $X_1, \ldots, X_n$ de fdr F et on veut savoir si $F = F_0$ avec F_0 la fdr d'une distribution de référence.
  - $H_0 : F = F_0$
  - Test : on calcule $\hat{F}_n$, et sous H_0, $\hat{F}_n - F_0$ suit approximativement la distribution (1).
- A partir de (1), on ne peut pas construire une bande de confiance exacte mais seulement une bande de confiance *asymptotique* (qui peut être bien meilleure que la bande exacte !)

Convergence en distribution

- Convergence vers la distribution de Kolmogorov : Quel que soit F, D_n converge en loi vers la distribution de Kolmogorov α_K indépendante de F

$$\mathbb{P}\left[D_n > \frac{c}{\sqrt{n}}\right] \xrightarrow[n\to\infty]{} \alpha_K(c) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2c^2} \quad (1)$$

- Application : test d'adéquation à une distribution donnée
  - On dispose d'un échantillon $X_1, \ldots, X_n$ de fdr F et on veut savoir si $F = F_0$ avec F_0 la fdr d'une distribution de référence.
  - $H_0 : F = F_0$
  - Test : on calcule $\hat{F}_n$, et sous H_0, $\hat{F}_n - F_0$ suit approximativement la distribution (1).
- A partir de (1), on ne peut pas construire une bande de confiance exacte mais seulement une bande de confiance *asymptotique* (qui peut être bien meilleure que la bande exacte !)

Convergence en distribution

- Convergence vers la distribution de Kolmogorov : Quel que soit F, D_n converge en loi vers la distribution de Kolmogorov α_K indépendante de F

$$\mathbb{P}\left[D_n > \frac{c}{\sqrt{n}}\right] \xrightarrow[n\to\infty]{} \alpha_K(c) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2c^2} \quad (1)$$

- Application : test d'adéquation à une distribution donnée
  - On dispose d'un échantillon $X_1, \ldots, X_n$ de fdr F et on veut savoir si $F = F_0$ avec F_0 la fdr d'une distribution de référence.
  - $H_0 : F = F_0$
  - Test : on calcule $\hat{F}_n$, et sous H_0, $\hat{F}_n - F_0$ suit approximativement la distribution (1).
- A partir de (1), on ne peut pas construire une bande de confiance exacte mais seulement une bande de confiance *asymptotique* (qui peut être bien meilleure que la bande exacte !)

Mise en oeuvre d'un test statistique

- Test = démarche consistant à accepter ou rejeter une hypothèse nulle à partir d'un échantillon d'observation.
- Définition d'une hypothèse nulle H_0 : hypothèse nulle qu'on cherche à rejeter. Il s'agit souvent d'une hypothèse d'égalité ("A=B").
- Définition d'une hypothèse alternative H_1. Deux cas typiques :
  - différence ("A$\neq$B") : test bilatère
  - inégalité avec a priori sur le sens ("A$>$B") : test unilatère
- Choix d'une statistique de test S (fonction des observations) et détermination de la distribution de S sous H_0
  $\hookrightarrow$ Ce choix dépend d'hypothèses a priori sur la distribution des données : hyp paramétriques (ex : distribution gaussienne) ou non paramétriques (ex : distribution symétrique).
- Tests non paramétriques : la distribution des données sous H_0 et/ou sous H_1 n'est pas spécifiée.

Principe des tests

Contexte

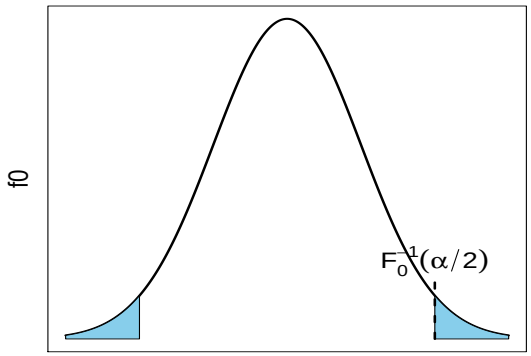
- Echantillon d'observation $\mathcal{E}$
- Hypothèse nulle H_0
- Statistique de test $S^{\text{obs}} = t(\mathcal{E})$
- Distribution de S sous l'hypothèse nulle : $S \sim_{H_0} \mathcal{F}_0$ (fdr F_0, densité f_0).

  $\hookrightarrow$ Remarque Parfois, on a simplement

  $$S \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{F}_0$$

  On parle alors de test asymptotique.
- Hypothèse alternative H_1, qui définit un test unilatère ou bilatère.

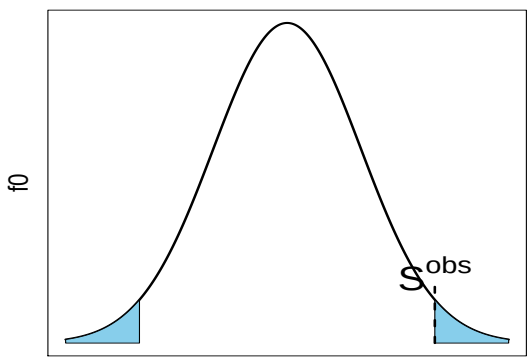
- Test de niveau α : rejet de H_0 si



unilatère : $S^{\text{obs}} \geq F_0^{-1}(1-\alpha)$

bilatère avec f_0 symétrique : $|S^{\text{obs}}| \geq F_0^{-1}(1-\alpha/2)$

- Degré de significativité ou p-value : plus petit niveau γ tel que H_0 est rejetée.



unilatère : $\gamma = \mathbb{P}_{T\sim\mathcal{F}_0}[T \geq S^{\text{obs}}] = 1 - F_0(S^{\text{obs}})$

bilatère avec f_0 symétrique :
$\gamma = \mathbb{P}_{T\sim\mathcal{F}_0}[|T| \geq |S^{\text{obs}}|] = 2(1 - F_0(S^{\text{obs}}))$

Commentaires

- Lorsqu'on ne rejette pas H_0, cela ne signifie pas qu'on accepte H_0.
  Ex : on veut tester si deux échantillons $(X_1, \ldots, X_n)$ i.i.d. et $(Y_1, \ldots, Y_n)$ i.i.d. sont issus de la même distribution. Si H_0 n'est pas rejetée, cela peut signifier que
  - H_0 est vraie

    OU
  - Les deux échantillons sont issus de distributions différentes mais cette différence n'est pas assez importante pour être détectée par le test utilisé avec la taille d'échantillon n.
- Pour une même hypothèse nulle, plusieurs tests sont parfois disponibles, et fournissent généralement des p-values différentes car les tests ont des puissances différentes.
- La puissance d'un test est la probabilité de rejeter l'hypothèse nulle pour une hypothèse alternative donnée.
- Les tests paramétriques sont souvent plus puissants que les tests non paramétriques.

Commentaires

- Lorsqu'on ne rejette pas H_0, cela ne signifie pas qu'on accepte H_0.
  Ex : on veut tester si deux échantillons $(X_1, \ldots, X_n)$ i.i.d. et $(Y_1, \ldots, Y_n)$ i.i.d. sont issus de la même distribution. Si H_0 n'est pas rejetée, cela peut signifier que
  - H_0 est vraie

    OU
  - Les deux échantillons sont issus de distributions différentes mais cette différence n'est pas assez importante pour être détectée par le test utilisé avec la taille d'échantillon n.
- Pour une même hypothèse nulle, plusieurs tests sont parfois disponibles, et fournissent généralement des p-values différentes car les tests ont des puissances différentes.
- La puissance d'un test est la probabilité de rejeter l'hypothèse nulle pour une hypothèse alternative donnée.
- Les tests paramétriques sont souvent plus puissants que les tests non paramétriques.

Commentaires

- Lorsqu'on ne rejette pas H_0, cela ne signifie pas qu'on accepte H_0.
  Ex : on veut tester si deux échantillons $(X_1, \ldots, X_n)$ i.i.d. et $(Y_1, \ldots, Y_n)$ i.i.d. sont issus de la même distribution. Si H_0 n'est pas rejetée, cela peut signifier que
  - H_0 est vraie

    OU
  - Les deux échantillons sont issus de distributions différentes mais cette différence n'est pas assez importante pour être détectée par le test utilisé avec la taille d'échantillon n.
- Pour une même hypothèse nulle, plusieurs tests sont parfois disponibles, et fournissent généralement des p-values différentes car les tests ont des puissances différentes.
- La puissance d'un test est la probabilité de rejeter l'hypothèse nulle pour une hypothèse alternative donnée.
- Les tests paramétriques sont souvent plus puissants que les tests non paramétriques.

Contexte de ce chapitre

Dans la suite, on considerera les situations suivantes.

- On observe un échantillon $(X_1, \ldots, X_n)$ de v.a. réelles i.i.d. (indépendantes identiquement distribuées)
  - Test d'adéquation : la distribution de X_i est-elle égale à une distribution donnée, ou appartient-elle à un ensemble de distributions donné ?
  - Test de médiane : la médiane de la distribution de X_i est-elle nulle ?
- On observe deux échantillons de v.a. réelles $(X_1, \ldots, X_n)$ i.i.d. et $(Y_1, \ldots, Y_m)$ i.i.d.
  - Test d'homogéneité : les distributions de X_i et Y_i sont-elles égales ?
  - Test de corrélation : les variables X_i et Y_i sont-elles corrélées ?

Exemples

- Tests d'adéquation à une loi : H_0 : "X suit la loi F_0" contre H_1 : "X ne suit pas la loi F_0".

  Exple : On connait la distribution de l'âge de décés pour l'ensemble de la population française. On mesure l'âge de décés de 100 bretons, et on veut tester si la distribution est identique au reste de la population

- Tests d'adéquation à une famille de lois : H_0 : "X est gaussienne" (paramètres non spécifiés) contre H_1 : "X n'est pas gaussienne".

- Tests de comparaison (ou homogénéité) : H_0 :"X et Y ont la même loi" contre H_1 : "Xet Y n'ont pas la même loi".

  Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 100 femmes.

- Tests d'indépendance : $H_0 : \{X_i\} \amalg \{Y_i\}$ contre H_1 : "$\{X_i\}$ ne sont pas indépendants des $\{Y_i\}$".

  Exple : on veut tester si l'espérance de vie X dépend du salaire moyen Y à partir de l'observation d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$.

Exemples

- Tests d'adéquation à une loi : H_0 : "X suit la loi F_0" contre H_1 : "X ne suit pas la loi F_0".

  Exple : On connait la distribution de l'âge de décés pour l'ensemble de la population française. On mesure l'âge de décés de 100 bretons, et on veut tester si la distribution est identique au reste de la population

- Tests d'adéquation à une famille de lois : H_0 : "X est gaussienne" (paramètres non spécifiés) contre H_1 : "X n'est pas gaussienne".

- Tests de comparaison (ou homogénéité) : H_0 :"X et Y ont la même loi" contre H_1 : "Xet Y n'ont pas la même loi".

  Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 100 femmes.

- Tests d'indépendance : H_0 : $\{X_i\} \amalg \{Y_i\}$ contre H_1 : "$\{X_i\}$ ne sont pas indépendants des $\{Y_i\}$".

  Exple : on veut tester si l'espérance de vie X dépend du salaire moyen Y à partir de l'observation d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$.

Exemples

- Tests d'adéquation à une loi : H_0 : "X suit la loi F_0" contre H_1 : "X ne suit pas la loi F_0".

  Exple : On connait la distribution de l'âge de décés pour l'ensemble de la population française. On mesure l'âge de décés de 100 bretons, et on veut tester si la distribution est identique au reste de la population

- Tests d'adéquation à une famille de lois : H_0 : "X est gaussienne" (paramètres non spécifiés) contre H_1 : "X n'est pas gaussienne".

- Tests de comparaison (ou homogénéité) : H_0 :"X et Y ont la même loi" contre H_1 : "Xet Y n'ont pas la même loi".

  Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 100 femmes.

- Tests d'indépendance : $H_0 : \{X_i\} \amalg \{Y_i\}$ contre H_1 : "$\{X_i\}$ ne sont pas indépendants des $\{Y_i\}$".

  Exple : on veut tester si l'espérance de vie X dépend du salaire moyen Y à partir de l'observation d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$.

Exemples

- Tests d'adéquation à une loi : H_0 : "X suit la loi F_0" contre H_1 : "X ne suit pas la loi F_0".

  Exple : On connait la distribution de l'âge de décés pour l'ensemble de la population française. On mesure l'âge de décés de 100 bretons, et on veut tester si la distribution est identique au reste de la population

- Tests d'adéquation à une famille de lois : H_0 : "X est gaussienne" (paramètres non spécifiés) contre H_1 : "X n'est pas gaussienne".

- Tests de comparaison (ou homogénéité) : H_0 :"X et Y ont la même loi" contre H_1 : "Xet Y n'ont pas la même loi".

  Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 100 femmes.

- Tests d'indépendance : H_0 : $\{X_i\} \amalg \{Y_i\}$ contre H_1 : "$\{X_i\}$ ne sont pas indépendants des $\{Y_i\}$".

  Exple : on veut tester si l'espérance de vie X dépend du salaire moyen Y à partir de l'observation d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$.

Exemples

- Tests d'adéquation à une loi : H_0 : "X suit la loi F_0" contre H_1 : "X ne suit pas la loi F_0".

  Exple : On connait la distribution de l'âge de décés pour l'ensemble de la population française. On mesure l'âge de décés de 100 bretons, et on veut tester si la distribution est identique au reste de la population

- Tests d'adéquation à une famille de lois : H_0 : "X est gaussienne" (paramètres non spécifiés) contre H_1 : "X n'est pas gaussienne".

- Tests de comparaison (ou homogénéité) : H_0 :"X et Y ont la même loi" contre H_1 : "Xet Y n'ont pas la même loi".

  Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 100 femmes.

- Tests d'indépendance : H_0 : $\{X_i\} \amalg \{Y_i\}$ contre H_1 : "$\{X_i\}$ ne sont pas indépendants des $\{Y_i\}$".

  Exple : on veut tester si l'espérance de vie X dépend du salaire moyen Y à partir de l'observation d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$.

Cas discret fini : test d'adéquation du χ^2 de Pearson

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. discrètes à r modalités de distribution $p \in \mathbb{R}^r$, et soit une distribution $p_0 \in \mathbb{R}^r$ fixée. On teste $H_0 : "p = p_0"$ contre $H_1 : "p \neq p_0"$

  Ex : on veut tester si les opinions politiques des habitants d'une ville diffèrent de la population nationale.

- Statistique de Pearson

$$T_n = n \sum_{k=1}^{r} \frac{(\hat{p}_k - p_0(k))^2}{p_0(k)}, \tag{1}$$

  où $\hat{p}_k = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = k\}}/n$

- Résultat : Sous H_0 :

$$T_n \underset{n \to \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \chi^2(r-1).$$

  et on rejette H_0 pour les grandes valeurs de T_n.

- Remarque : C'est en fait un test paramétrique ! puisque la loi discrète des X_i dépend d'un nombre fini de paramètres.

Cas discret fini : test d'adéquation du χ^2 de Pearson

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. discrètes à r modalités de distribution $p \in \mathbb{R}^r$, et soit une distribution $p_0 \in \mathbb{R}^r$ fixée. On teste H_0 : "$p = p_0$" contre H_1 : "$p \neq p_0$"

  Ex : on veut tester si les opinions politiques des habitants d'une ville diffèrent de la population nationale.

- Statistique de Pearson

$$T_n = n \sum_{k=1}^{r} \frac{(\hat{p}_k - p_0(k))^2}{p_0(k)}, \tag{1}$$

  où $\hat{p}_k = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = k\}}/n$

- Résultat : Sous H_0 :

$$T_n \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \chi^2(r-1).$$

  et on rejette H_0 pour les grandes valeurs de T_n.

- Remarque : C'est en fait un test paramétrique ! puisque la loi discrète des X_i dépend d'un nombre fini de paramètres.

Cas discret fini : test d'adéquation du χ^2 de Pearson

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. discrètes à r modalités de distribution $p \in \mathbb{R}^r$, et soit une distribution $p_0 \in \mathbb{R}^r$ fixée. On teste H_0 : "$p = p_0$" contre H_1 : "$p \neq p_0$"

  Ex : on veut tester si les opinions politiques des habitants d'une ville diffèrent de la population nationale.

- Statistique de Pearson

$$T_n = n \sum_{k=1}^{r} \frac{(\hat{p}_k - p_0(k))^2}{p_0(k)}, \tag{1}$$

  où $\hat{p}_k = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = k\}}/n$

- Résultat : Sous H_0 :

$$T_n \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \chi^2(r-1).$$

  et on rejette H_0 pour les grandes valeurs de T_n.

- Remarque : C'est en fait un test paramétrique ! puisque la loi discrète des X_i dépend d'un nombre fini de paramètres.

Cas discret fini : test d'adéquation du χ^2 de Pearson

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. discrètes à r modalités de distribution $p \in \mathbb{R}^r$, et soit une distribution $p_0 \in \mathbb{R}^r$ fixée. On teste H_0 : "$p = p_0$" contre H_1 : "$p \neq p_0$"

  Ex : on veut tester si les opinions politiques des habitants d'une ville diffèrent de la population nationale.

- Statistique de Pearson

$$T_n = n \sum_{k=1}^{r} \frac{(\hat{p}_k - p_0(k))^2}{p_0(k)}, \tag{1}$$

  où $\hat{p}_k = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = k\}}/n$

- Résultat : Sous H_0 :

$$T_n \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \chi^2(r-1).$$

  et on rejette H_0 pour les grandes valeurs de T_n.

- Remarque : C'est en fait un test paramétrique ! puisque la loi discrète des X_i dépend d'un nombre fini de paramètres.

Cas discret fini : test d'adéquation du χ^2 de Pearson

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. discrètes à r modalités de distribution $p \in \mathbb{R}^r$, et soit une distribution $p_0 \in \mathbb{R}^r$ fixée. On teste H_0 : "$p = p_0$" contre H_1 : "$p \neq p_0$"

  Ex : on veut tester si les opinions politiques des habitants d'une ville diffèrent de la population nationale.

- Statistique de Pearson

$$T_n = n \sum_{k=1}^{r} \frac{(\hat{p}_k - p_0(k))^2}{p_0(k)}, \tag{1}$$

  où $\hat{p}_k = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = k\}}/n$

- Résultat : Sous H_0 :

$$T_n \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \chi^2(r-1).$$

  et on rejette H_0 pour les grandes valeurs de T_n.

- Remarque : C'est en fait un test paramétrique ! puisque la loi discrète des X_i dépend d'un nombre fini de paramètres.

Cas continu : test de Kolmogorov Smirnov (KS)

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. continues de fdr F,et soit F_0 une fdr fixée. On teste $H_0 : "F = F_0"$ contre $H_1 : "F \neq F_0"$
- Statistique de KS :

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

$\hookrightarrow$ *On rejette H_0 pour les grandes valeurs de D_n.*

- La statistique D_n est asymptotiquement libre en loi sous H_0.
- La loi asymptotique est tabulée pour les petites valeurs de n, et une formule approchée est disponible pour les grandes valeurs de n.

Cas continu : test de Kolmogorov Smirnov (KS)

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. continues de fdr F,et soit F_0 une fdr fixée. On teste $H_0 : "F = F_0"$ contre $H_1 : "F \neq F_0"$
- Statistique de KS :

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

$\hookrightarrow$ *On rejette H_0 pour les grandes valeurs de D_n.*

- La statistique D_n est asymptotiquement libre en loi sous H_0.
- La loi asymptotique est tabulée pour les petites valeurs de n, et une formule approchée est disponible pour les grandes valeurs de n.

Cas continu : test de Kolmogorov Smirnov (KS)

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. continues de fdr F,et soit F_0 une fdr fixée. On teste $H_0 : "F = F_0"$ contre $H_1 : "F \neq F_0"$
- Statistique de KS :

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

$\hookrightarrow$ *On rejette H_0 pour les grandes valeurs de D_n.*

- La statistique D_n est asymptotiquement libre en loi sous H_0.
- La loi asymptotique est tabulée pour les petites valeurs de n, et une formule approchée est disponible pour les grandes valeurs de n.

Cas continu : test de Kolmogorov Smirnov (KS)

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. continues de fdr F,et soit F_0 une fdr fixée. On teste $H_0 : "F = F_0"$ contre $H_1 : "F \neq F_0"$
- Statistique de KS :

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

$\hookrightarrow$ *On rejette H_0 pour les grandes valeurs de D_n.*

- La statistique D_n est asymptotiquement libre en loi sous H_0.
- La loi asymptotique est tabulée pour les petites valeurs de n, et une formule approchée est disponible pour les grandes valeurs de n.

Cas continu : test de Kolmogorov Smirnov (KS)

- Contexte. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de v.a. continues de fdr F,et soit F_0 une fdr fixée. On teste $H_0 : "F = F_0"$ contre $H_1 : "F \neq F_0"$
- Statistique de KS :

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

$\hookrightarrow$ *On rejette H_0 pour les grandes valeurs de D_n.*

- La statistique D_n est asymptotiquement libre en loi sous H_0.
- La loi asymptotique est tabulée pour les petites valeurs de n, et une formule approchée est disponible pour les grandes valeurs de n.

Tests d'adéquation à une famille de lois

- Contexte On veut tester l'hypothèse nulle $H_0' : F \in \mathcal{F}$ où $\mathcal{F}$ est une famille de distribution.
- Cas discret : χ^2 se généralise au cas de $H_0' : F = F_0(\theta_1, \ldots, \theta_s)$ où $\theta_1, \ldots, \theta_s$ inconnus.
- Cas continu : KS ne s'applique pas directement à ce cadre. En effet, leurs statistiques ne sont plus libres en loi (même asymptotiquement) sous H_0'. Il faut adapter ces tests pour chaque famille de loi considérée.

  $\hookrightarrow$ Le test de Lillilefors est une adaptation de KS pour H_0' : $F \in \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$.

  $\hookrightarrow$ D'autres tests existent comme le test de Shapiro-Wilk.

Tests d'adéquation à une famille de lois

- Contexte On veut tester l'hypothèse nulle $H_0' : F \in \mathcal{F}$ où $\mathcal{F}$ est une famille de distribution.
- Cas discret : χ^2 se généralise au cas de $H_0' : F = F_0(\theta_1, \ldots, \theta_s)$ où $\theta_1, \ldots, \theta_s$ inconnus.
- Cas continu : KS ne s'applique pas directement à ce cadre. En effet, leurs statistiques ne sont plus libres en loi (même asymptotiquement) sous H_0'. Il faut adapter ces tests pour chaque famille de loi considérée.

  $\hookrightarrow$ Le test de Lillilefors est une adaptation de KS pour H_0' : $F \in \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$.

  $\hookrightarrow$ D'autres tests existent comme le test de Shapiro-Wilk.

Tests d'adéquation à une famille de lois

- Contexte On veut tester l'hypothèse nulle $H_0' : F \in \mathcal{F}$ où $\mathcal{F}$ est une famille de distribution.
- Cas discret : χ^2 se généralise au cas de $H_0' : F = F_0(\theta_1, \ldots, \theta_s)$ où $\theta_1, \ldots, \theta_s$ inconnus.
- Cas continu : KS ne s'applique pas directement à ce cadre. En effet, leurs statistiques ne sont plus libres en loi (même asymptotiquement) sous H_0'. Il faut adapter ces tests pour chaque famille de loi considérée.

  $\hookrightarrow$ Le test de Lillilefors est une adaptation de KS pour $H_0' : F \in \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$.

  $\hookrightarrow$ D'autres tests existent comme le test de Shapiro-Wilk.

Tests d'adéquation à une famille de lois

- Contexte On veut tester l'hypothèse nulle $H_0' : F \in \mathcal{F}$ où $\mathcal{F}$ est une famille de distribution.
- Cas discret : χ^2 se généralise au cas de $H_0' : F = F_0(\theta_1, \ldots, \theta_s)$ où $\theta_1, \ldots, \theta_s$ inconnus.
- Cas continu : KS ne s'applique pas directement à ce cadre. En effet, leurs statistiques ne sont plus libres en loi (même asymptotiquement) sous H_0'. Il faut adapter ces tests pour chaque famille de loi considérée.

  $\hookrightarrow$ Le test de Lillilefors est une adaptation de KS pour $H_0' : F \in \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$.

  $\hookrightarrow$ D'autres tests existent comme le test de Shapiro-Wilk.

Tests de médiane (ou de symétrie)

- Objectif général. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On veut tester si la distribution de X est symétrique par rapport à 0.
- Objectif restreint. Pour rejeter l'hypothèse de symétrie, une condition suffisante est que la médiane soit différente de zéro. On va donc considérer le test induit "la médiane de X est-elle nulle ?".
- *Exple : on veut tester si le fait d'emménager en couple a une influence sur le poids. Pour cela, on mesure la différence de poids avant et un an après l'emménagement pour 100 individus.*

Tests de médiane (ou de symétrie)

- Objectif général. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On veut tester si la distribution de X est symétrique par rapport à 0.
- Objectif restreint. Pour rejeter l'hypothèse de symétrie, une condition suffisante est que la médiane soit différente de zéro. On va donc considérer le test induit "la médiane de X est-elle nulle ?".
- *Exple : on veut tester si le fait d'emménager en couple a une influence sur le poids. Pour cela, on mesure la différence de poids avant et un an après l'emménagement pour 100 individus.*

Tests de médiane (ou de symétrie)

- Objectif général. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On veut tester si la distribution de X est symétrique par rapport à 0.
- Objectif restreint. Pour rejeter l'hypothèse de symétrie, une condition suffisante est que la médiane soit différente de zéro. On va donc considérer le test induit "la médiane de X est-elle nulle ?".
- *Exple : on veut tester si le fait d'emménager en couple a une influence sur le poids. Pour cela, on mesure la différence de poids avant et un an après l'emménagement pour 100 individus.*

Tests de signe

- Contexte. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On teste

$$H_0 : \mathbb{P}(X \leq 0) = 1/2 \quad \text{contre} \quad H_1 : \mathbb{P}(X \leq 0) > 1/2$$
$$\text{ou} \qquad H_1' : \mathbb{P}(X \leq 0) < 1/2$$

- Statistique de signe

$$S_n = \sum_{i=1}^{n} 1\{X_i \leq 0\} \sim \mathcal{B}(n, m), \quad \text{où} \quad m = \mathbb{P}(X \leq 0)$$

Sous H_0, on a $S_n \sim \mathcal{B}(n, 1/2)$ et sous $H_1 : m > 1/2$, S_n est stochastiquement plus grande que sous H_0. On rejette donc H_0 pour les grandes valeurs de S_n.

- Remarque. C'est en fait un test paramétrique !
- Test très général, mais qui utilise très peu d'information sur les variables, donc peu puissant.
  $\hookrightarrow$ test de signe et rang : plus puissant.

Tests de signe

- Contexte. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On teste

$$H_0 : \mathbb{P}(X \leq 0) = 1/2 \quad \text{contre} \quad H_1 : \mathbb{P}(X \leq 0) > 1/2$$
$$\text{ou} \quad H_1' : \mathbb{P}(X \leq 0) < 1/2$$

- Statistique de signe

$$S_n = \sum_{i=1}^{n} 1\{X_i \leq 0\} \sim \mathcal{B}(n, m), \quad \text{où} \quad m = \mathbb{P}(X \leq 0)$$

Sous H_0, on a $S_n \sim \mathcal{B}(n, 1/2)$ et sous $H_1 : m > 1/2$, S_n est stochastiquement plus grande que sous H_0. On rejette donc H_0 pour les grandes valeurs de S_n.

- Remarque. C'est en fait un test paramétrique !
- Test très général, mais qui utilise très peu d'information sur les variables, donc peu puissant.
  $\hookrightarrow$ test de signe et rang : plus puissant.

Tests de signe

- Contexte. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On teste

$$H_0 : \mathbb{P}(X \leq 0) = 1/2 \quad \text{contre} \quad H_1 : \mathbb{P}(X \leq 0) > 1/2$$
$$\text{ou} \quad H_1' : \mathbb{P}(X \leq 0) < 1/2$$

- Statistique de signe

$$S_n = \sum_{i=1}^{n} 1\{X_i \leq 0\} \sim \mathcal{B}(n, m), \quad \text{où} \quad m = \mathbb{P}(X \leq 0)$$

Sous H_0, on a $S_n \sim \mathcal{B}(n, 1/2)$ et sous $H_1 : m > 1/2$, S_n est stochastiquement plus grande que sous H_0. On rejette donc H_0 pour les grandes valeurs de S_n.

- Remarque. C'est en fait un test paramétrique !
- Test très général, mais qui utilise très peu d'information sur les variables, donc peu puissant.
  $\hookrightarrow$ test de signe et rang : plus puissant.

Tests de signe

- Contexte. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On teste

$$H_0 : \mathbb{P}(X \leq 0) = 1/2 \quad \text{contre} \quad H_1 : \mathbb{P}(X \leq 0) > 1/2$$
$$\text{ou} \quad H_1' : \mathbb{P}(X \leq 0) < 1/2$$

- Statistique de signe

$$S_n = \sum_{i=1}^{n} 1\{X_i \leq 0\} \sim \mathcal{B}(n, m), \quad \text{où} \quad m = \mathbb{P}(X \leq 0)$$

Sous H_0, on a $S_n \sim \mathcal{B}(n, 1/2)$ et sous $H_1 : m > 1/2$, S_n est stochastiquement plus grande que sous H_0. On rejette donc H_0 pour les grandes valeurs de S_n.

- Remarque. C'est en fait un test paramétrique !
- Test très général, mais qui utilise très peu d'information sur les variables, donc peu puissant.
  $\hookrightarrow$ test de signe et rang : plus puissant.

Tests de signe

- Contexte. Soient $X_1, \ldots, X_n$ v.a. réelles i.i.d. On teste

$$H_0 : \mathbb{P}(X \leq 0) = 1/2 \quad \text{contre} \quad H_1 : \mathbb{P}(X \leq 0) > 1/2$$
$$\text{ou} \quad H_1' : \mathbb{P}(X \leq 0) < 1/2$$

- Statistique de signe

$$S_n = \sum_{i=1}^{n} 1\{X_i \leq 0\} \sim \mathcal{B}(n, m), \quad \text{où} \quad m = \mathbb{P}(X \leq 0)$$

Sous H_0, on a $S_n \sim \mathcal{B}(n, 1/2)$ et sous $H_1 : m > 1/2$, S_n est stochastiquement plus grande que sous H_0. On rejette donc H_0 pour les grandes valeurs de S_n.

- Remarque. C'est en fait un test paramétrique !
- Test très général, mais qui utilise très peu d'information sur les variables, donc peu puissant.
  $\hookrightarrow$ test de signe et rang : plus puissant.

Test de signe et de rang : idée

- *Exple : on veut tester si le fait d'emménager en couple a une influence sur le poids. Pour cela, on mesure la différence de poids avant et un an après l'emménagement pour 100 individus.*
- Avec le test de signe, une personne qui a pris 5 kg a autant d'effet sur la statistique de test qu'une personne qui a pris 1kg (seul le signe de X_i est pris en compte).
- Avec le test de signe et rang, la valeur absolue de la variable $|X_i|$ est prise en compte.

Test de signe et de rang : idée

- *Exple : on veut tester si le fait d'emménager en couple a une influence sur le poids. Pour cela, on mesure la différence de poids avant et un an après l'emménagement pour 100 individus.*
- Avec le test de signe, une personne qui a pris 5 kg a autant d'effet sur la statistique de test qu'une personne qui a pris 1kg (seul le signe de X_i est pris en compte).
- Avec le test de signe et rang, la valeur absolue de la variable $|X_i|$ est prise en compte.

Test de signe et de rang : idée

- *Exple : on veut tester si le fait d'emménager en couple a une influence sur le poids. Pour cela, on mesure la différence de poids avant et un an après l'emménagement pour 100 individus.*
- Avec le test de signe, une personne qui a pris 5 kg a autant d'effet sur la statistique de test qu'une personne qui a pris 1kg (seul le signe de X_i est pris en compte).
- Avec le test de signe et rang, la valeur absolue de la variable $|X_i|$ est prise en compte.

Statistique de rang

- Statistique de rang. Soit $X_1, \dots X_n$ échantillon de v.a. réelles, et $X^\star = (X_{(1)}, \dots, X_{(n)})$ la statistique d'ordre associée telle que $X_{(1)} \leq \cdots \leq X_{(n)}$. Le vecteur R_X des rangs de X une permutation de $\{1, \dots, n\}$ telle que $X_i = X_{(R_X(i))}$.
- S'il y a des ex-aequos, le vecteur R_X n'est pas unique. Si la distribution de X est diffuse (i.e. $\mathbb{P}[X = x] = 0$ pour tout x), il n'y a pas d'ex-aequo.
- Exemple

$$\begin{cases} X = (5, 2, 6, 3, 8, 1) \\ R_X = (4, 2, 5, 3, 6, 1) \end{cases} \qquad \begin{cases} X_1 = 5 = X_{(4)} \\ R_X(1) = 4 \end{cases}$$

Statistique de rang

- Statistique de rang. Soit $X_1, \dots X_n$ échantillon de v.a. réelles, et $X^\star = (X_{(1)}, \dots, X_{(n)})$ la statistique d'ordre associée telle que $X_{(1)} \leq \cdots \leq X_{(n)}$. Le vecteur R_X des rangs de X une permutation de $\{1, \dots, n\}$ telle que $X_i = X_{(R_X(i))}$.
- S'il y a des ex-aequos, le vecteur R_X n'est pas unique. Si la distribution de X est diffuse (i.e. $\mathbb{P}[X = x] = 0$ pour tout x), il n'y a pas d'ex-aequo.
- Exemple

$$\begin{cases} X = (5, 2, 6, 3, 8, 1) \\ R_X = (4, 2, 5, 3, 6, 1) \end{cases} \qquad \begin{cases} X_1 = 5 = X_{(4)} \\ R_X(1) = 4 \end{cases}$$

Statistique de rang

- Statistique de rang. Soit $X_1, \dots X_n$ échantillon de v.a. réelles, et $X^\star = (X_{(1)}, \dots, X_{(n)})$ la statistique d'ordre associée telle que $X_{(1)} \leq \cdots \leq X_{(n)}$. Le vecteur R_X des rangs de X une permutation de $\{1, \dots, n\}$ telle que $X_i = X_{(R_X(i))}$.
- S'il y a des ex-aequos, le vecteur R_X n'est pas unique. Si la distribution de X est diffuse (i.e. $\mathbb{P}[X = x] = 0$ pour tout x), il n'y a pas d'ex-aequo.
- Exemple

$$\begin{cases} X = (5, 2, 6, 3, 8, 1) \\ R_X = (4, 2, 5, 3, 6, 1) \end{cases} \qquad \begin{cases} X_1 = 5 = X_{(4)} \\ R_X(1) = 4 \end{cases}$$

Statistique de rang

- Statistique de rang. Soit $X_1, \dots X_n$ échantillon de v.a. réelles, et $X^\star = (X_{(1)}, \dots, X_{(n)})$ la statistique d'ordre associée telle que $X_{(1)} \leq \cdots \leq X_{(n)}$. Le vecteur R_X des rangs de X une permutation de $\{1, \dots, n\}$ telle que $X_i = X_{(R_X(i))}$.
- S'il y a des ex-aequos, le vecteur R_X n'est pas unique. Si la distribution de X est diffuse (i.e. $\mathbb{P}[X = x] = 0$ pour tout x), il n'y a pas d'ex-aequo.
- Exemple

$$\begin{cases} X = (5, 2, 6, 3, 8, 1) \\ R_X = (4, 2, 5, 3, 6, 1) \end{cases} \qquad \begin{cases} X_1 = 5 = X_{(4)} \\ R_X(1) = 4 \end{cases}$$

Tests de signe et rang (Wilcoxon signed rank test) (1)

- Contexte. Soit $X_1, \dots X_n$ un échantillon de v.a. réelles de loi supposée diffuse. On veut tester :
  H_0 : "la loi de X est symétrique (par rapport à 0)" contre
  H_1 : "la loi de X n'est pas symétrique".

- Statistique de Wilcoxon

$$W_n^+ = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i > 0\}},$$

  où $R_{|X|}$ le vecteur des rangs associé à $(|X_1|, \dots, |X_n|)$

- Soit $W_n^- = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i < 0\}}$, alors

$$W_n^+ + W_n^- = \frac{n(n+1)}{2} \quad \text{p.s.}$$

- Il existe des variantes prenant en compte des ex-aequos.

Tests de signe et rang (Wilcoxon signed rank test) (1)

- Contexte. Soit $X_1, \dots X_n$ un échantillon de v.a. réelles de loi supposée diffuse. On veut tester :
  H_0 : "la loi de X est symétrique (par rapport à 0)" contre
  H_1 : "la loi de X n'est pas symétrique".

- Statistique de Wilcoxon

$$W_n^+ = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i > 0\}},$$

où $R_{|X|}$ le vecteur des rangs associé à $(|X_1|, \dots, |X_n|)$

- Soit $W_n^- = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i < 0\}}$, alors

$$W_n^+ + W_n^- = \frac{n(n+1)}{2} \quad \text{p.s.}$$

- Il existe des variantes prenant en compte des ex-aequos.

Tests de signe et rang (Wilcoxon signed rank test) (1)

- Contexte. Soit $X_1, \dots X_n$ un échantillon de v.a. réelles de loi supposée diffuse. On veut tester :
  H_0 : "la loi de X est symétrique (par rapport à 0)" contre
  H_1 : "la loi de X n'est pas symétrique".

- Statistique de Wilcoxon

$$W_n^+ = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i > 0\}},$$

  où $R_{|X|}$ le vecteur des rangs associé à $(|X_1|, \dots, |X_n|)$

- Soit $W_n^- = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i < 0\}}$, alors

$$W_n^+ + W_n^- = \frac{n(n+1)}{2} \quad \text{p.s.}$$

- Il existe des variantes prenant en compte des ex-aequos.

Tests de signe et rang (Wilcoxon signed rank test) (1)

- Contexte. Soit $X_1, \dots X_n$ un échantillon de v.a. réelles de loi supposée diffuse. On veut tester :
  H_0 : "la loi de X est symétrique (par rapport à 0)" contre
  H_1 : "la loi de X n'est pas symétrique".

- Statistique de Wilcoxon

$$W_n^+ = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i > 0\}},$$

  où $R_{|X|}$ le vecteur des rangs associé à $(|X_1|, \dots, |X_n|)$

- Soit $W_n^- = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i < 0\}}$, alors

$$W_n^+ + W_n^- = \frac{n(n+1)}{2} \quad \text{p.s.}$$

- Il existe des variantes prenant en compte des ex-aequos.

Tests de signe et rang (Wilcoxon signed rank test) (1)

- Contexte. Soit $X_1, \dots X_n$ un échantillon de v.a. réelles de loi supposée diffuse. On veut tester :
  H_0 : "la loi de X est symétrique (par rapport à 0)" contre
  H_1 : "la loi de X n'est pas symétrique".

- Statistique de Wilcoxon

$$W_n^+ = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i > 0\}},$$

  où $R_{|X|}$ le vecteur des rangs associé à $(|X_1|, \dots, |X_n|)$

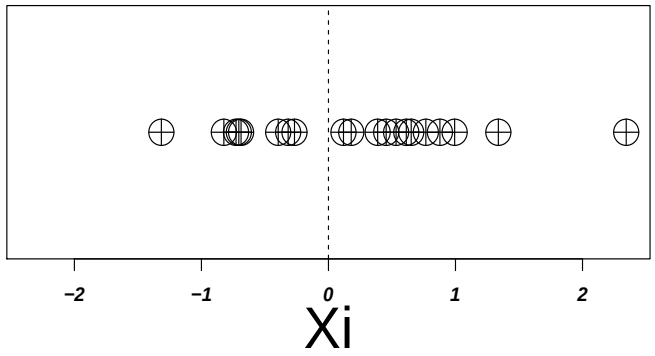
- Soit $W_n^- = \sum_{i=1}^{n} R_{|X|}(i) 1_{\{X_i < 0\}}$, alors

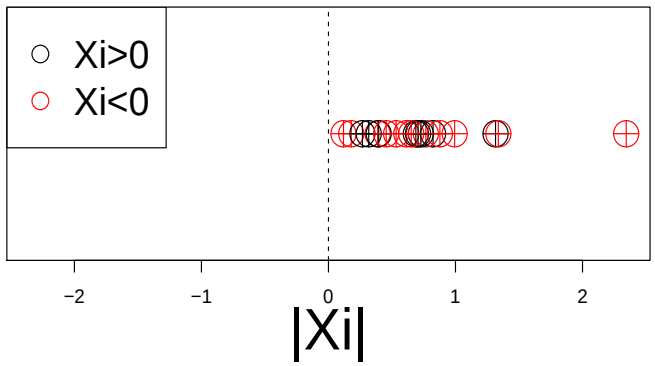
$$W_n^+ + W_n^- = \frac{n(n+1)}{2} \quad \text{p.s.}$$

- Il existe des variantes prenant en compte des ex-aequos.

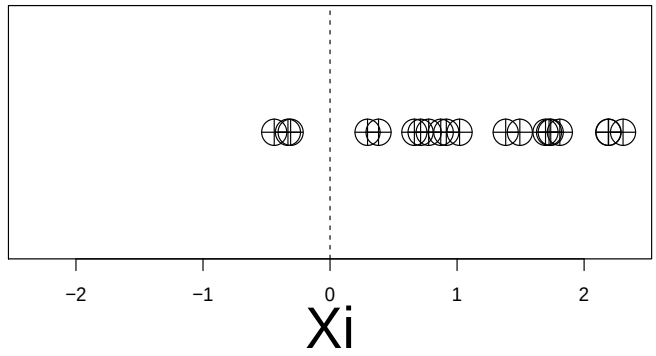
Exemple

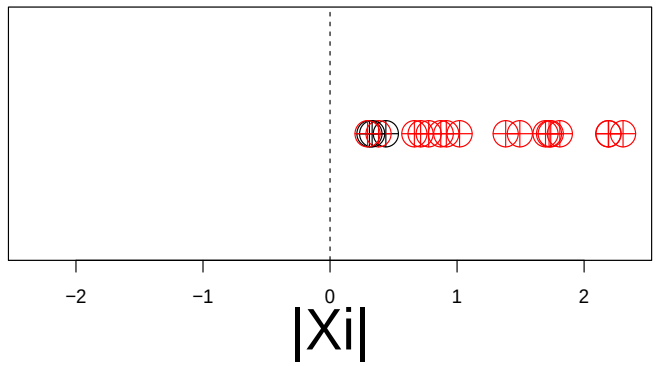
median $= 0$





median $\neq 0$





Lorsque la mediane de X est différente de 0, les rangs des X_i positifs ne sont pas distribués uniformément sur $\{1, \ldots, n\}$.

Tests de signe et rang (Wilcoxon signed rank test) (2)

- Théorème. Sous H_0 : "la loi de X est symétrique par rapport à 0", W_n^+ est libre en loi. De plus,

$$\mathbb{E}_{H_0}(W_n^+) = \frac{n(n+1)}{4}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_n^+) = \frac{n(n+1)(2n+1)}{24}$$

$$\text{et} \quad \frac{W_n^+ - \mathbb{E}_{H_0}(W_n^+)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_n^+)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

- Test de H_0. On rejette H_0 pour les grandes valeurs de $|W_n^+ - n(n+1)/4|$.
- La loi de W_n^+ sous H_0 est tabulée usuellement pour $n \leq 20$, et une approximation est calculée pour $n > 20$.
  $\hookrightarrow$ Ce n'est pas un test asymptotique car W_n^+ est libre en loi strictement, pas asymptotiquement.

Tests de signe et rang (Wilcoxon signed rank test) (2)

- Théorème. Sous H_0 : "la loi de X est symétrique par rapport à 0", W_n^+ est libre en loi. De plus,

$$\mathbb{E}_{H_0}(W_n^+) = \frac{n(n+1)}{4}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_n^+) = \frac{n(n+1)(2n+1)}{24}$$

$$\text{et} \quad \frac{W_n^+ - \mathbb{E}_{H_0}(W_n^+)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_n^+)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

- Test de H_0. On rejette H_0 pour les grandes valeurs de $|W_n^+ - n(n+1)/4|$.
- La loi de W_n^+ sous H_0 est tabulée usuellement pour $n \leq 20$, et une approximation est calculée pour $n > 20$.
  $\hookrightarrow$ Ce n'est pas un test asymptotique car W_n^+ est libre en loi strictement, pas asymptotiquement.

Tests de signe et rang (Wilcoxon signed rank test) (2)

- Théorème. Sous H_0 : "la loi de X est symétrique par rapport à 0", W_n^+ est libre en loi. De plus,

$$\mathbb{E}_{H_0}(W_n^+) = \frac{n(n+1)}{4}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_n^+) = \frac{n(n+1)(2n+1)}{24}$$

$$\text{et} \quad \frac{W_n^+ - \mathbb{E}_{H_0}(W_n^+)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_n^+)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

- Test de H_0. On rejette H_0 pour les grandes valeurs de $|W_n^+ - n(n+1)/4|$.
- La loi de W_n^+ sous H_0 est tabulée usuellement pour $n \leq 20$, et une approximation est calculée pour $n > 20$.
  $\hookrightarrow$ Ce n'est pas un test asymptotique car W_n^+ est libre en loi strictement, pas asymptotiquement.

Echantillon appariés/non appariés

Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons i.i.d. tirés dans deux populations de loi respectives F et G.

- Les échantillons sont dits appariés si X_i et Y_i sont associés dans le design de l'expérience :
  - X_i et Y_i sont des mesures d'une même quantité pour un même individu, par exemple à deux temps différents ou sous deux traitements différents
  - X_i et Y_i sont deux quantités différentes mesurées sur un même individu.
- Si les deux échantillons sont tirés de façon indépendante dans deux populations, ils sont non appariées.
- Si les échantillons sont appariés, ils sont nécessairement de même taille $(n = m)$.

Echantillon appariés/non appariés

Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons i.i.d. tirés dans deux populations de loi respectives F et G.

- Les échantillons sont dits appariés si X_i et Y_i sont associés dans le design de l'expérience :
  - X_i et Y_i sont des mesures d'une même quantité pour un même individu, par exemple à deux temps différents ou sous deux traitements différents
  - X_i et Y_i sont deux quantités différentes mesurées sur un même individu.
- Si les deux échantillons sont tirés de façon indépendante dans deux populations, ils sont non appariées.
- Si les échantillons sont appariés, ils sont nécessairement de même taille ($n = m$).

Echantillon appariés/non appariés

Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons i.i.d. tirés dans deux populations de loi respectives F et G.

- Les échantillons sont dits appariés si X_i et Y_i sont associés dans le design de l'expérience :
  - X_i et Y_i sont des mesures d'une même quantité pour un même individu, par exemple à deux temps différents ou sous deux traitements différents
  - X_i et Y_i sont deux quantités différentes mesurées sur un même individu.
- Si les deux échantillons sont tirés de façon indépendante dans deux populations, ils sont non appariées.
- Si les échantillons sont appariés, ils sont nécessairement de même taille ($n = m$).

Echantillon appariés/non appariés

Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons i.i.d. tirés dans deux populations de loi respectives F et G.

- Les échantillons sont dits appariés si X_i et Y_i sont associés dans le design de l'expérience :
  - X_i et Y_i sont des mesures d'une même quantité pour un même individu, par exemple à deux temps différents ou sous deux traitements différents
  - X_i et Y_i sont deux quantités différentes mesurées sur un même individu.
- Si les deux échantillons sont tirés de façon indépendante dans deux populations, ils sont non appariées.
- Si les échantillons sont appariés, ils sont nécessairement de même taille ($n = m$).

- Contexte $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons indépendants de v.a. réelles diffuses de lois respectives F et G. On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".

Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 150 femmes.

Pour des échantillons appariés

- $n = m$ et on se ramène à un unique échantillon $(Z_1, \ldots, Z_n)$ avec $Z_i = X_i - Y_i$.
- Sous H_0, la loi de Z est symétrique. D'où le test induit H_0' : "La loi de Z est symétrique" contre H_1' :"La loi de Z n'est pas symétrique".
- On applique le test de signe et rang de Wilcoxon.

Pour des échantillons non-appariés

- Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou Mann-Whitney).

- Contexte $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons indépendants de v.a. réelles diffuses de lois respectives F et G. On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".

Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 150 femmes.

Pour des échantillons appariés

- $n = m$ et on se ramène à un unique échantillon $(Z_1, \ldots, Z_n)$ avec $Z_i = X_i - Y_i$.
- Sous H_0, la loi de Z est symétrique. D'où le test induit H_0' : "La loi de Z est symétrique" contre H_1' :"La loi de Z n'est pas symétrique".
- On applique le test de signe et rang de Wilcoxon.

Pour des échantillons non-appariés

- Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou Mann-Whitney).

- Contexte $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons indépendants de v.a. réelles diffuses de lois respectives F et G. On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".

Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 150 femmes.

Pour des échantillons appariés

- $n = m$ et on se ramène à un unique échantillon $(Z_1, \ldots, Z_n)$ avec $Z_i = X_i - Y_i$.
- Sous H_0, la loi de Z est symétrique. D'où le test induit H_0' : "La loi de Z est symétrique" contre H_1' :"La loi de Z n'est pas symétrique".
- On applique le test de signe et rang de Wilcoxon.

Pour des échantillons non-appariés

- Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou Mann-Whitney).

- Contexte $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons indépendants de v.a. réelles diffuses de lois respectives F et G. On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".

Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 150 femmes.

Pour des échantillons appariés

- $n = m$ et on se ramène à un unique échantillon $(Z_1, \ldots, Z_n)$ avec $Z_i = X_i - Y_i$.
- Sous H_0, la loi de Z est symétrique. D'où le test induit H_0' : "La loi de Z est symétrique" contre H_1' :"La loi de Z n'est pas symétrique".
- On applique le test de signe et rang de Wilcoxon.

Pour des échantillons non-appariés

- Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou Mann-Whitney).

- Contexte $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons indépendants de v.a. réelles diffuses de lois respectives F et G. On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".

Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 150 femmes.

Pour des échantillons appariés

- $n = m$ et on se ramène à un unique échantillon $(Z_1, \ldots, Z_n)$ avec $Z_i = X_i - Y_i$.
- Sous H_0, la loi de Z est symétrique. D'où le test induit H_0' : "La loi de Z est symétrique" contre H_1' :"La loi de Z n'est pas symétrique".
- On applique le test de signe et rang de Wilcoxon.

Pour des échantillons non-appariés

- Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou Mann-Whitney).

- Contexte $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons indépendants de v.a. réelles diffuses de lois respectives F et G. On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".

Exple : on veut tester si la distribution des salaires dans une entreprise dépend du sexe, à partir de l'observation du salaire de 100 hommes et de 150 femmes.

Pour des échantillons appariés

- $n = m$ et on se ramène à un unique échantillon $(Z_1, \ldots, Z_n)$ avec $Z_i = X_i - Y_i$.
- Sous H_0, la loi de Z est symétrique. D'où le test induit H_0' : "La loi de Z est symétrique" contre H_1' :"La loi de Z n'est pas symétrique".
- On applique le test de signe et rang de Wilcoxon.

Pour des échantillons non-appariés

- Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou Mann-Whitney).

Test de Kolmogorov Smirnov de comparaison de 2 échantillons

Statistique de KS pour deux échantillons

$$D'_{n,m} = \sqrt{\frac{n+m}{nm}} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)|.$$

où $\hat{F}_n$ et $\hat{G}_m$ sont les fdr empiriques des échantillons $(X_i)_{i=1,\dots,n}$ and $(Y_i)_{i=1,\dots,m}$ On rejette H_0 pour les grandes valeurs de $D_{n,m}$.

Propriétés

- La statistique $D'_{n,m}$ est libre en loi sous H_0. Cette loi est tabulée.

Test de Mann-Whitney (1)

- Procédure. On classe les variables $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ par leur rang global et on note $R_1, R_2, \ldots, R_n \in \{1, \ldots, n+m\}$ les rangs associés à l'échantillon X.
- Exemple. $X = (3, 5, 2)$; $Y = (1, 4)$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon X sont $R_1 = 3, R_2 = 5, R_3 = 2$.
- Idée. Sous H_0, $R_1, \ldots, R_n$ sont uniformément répartis sur $\{1, \ldots, n+m\}$.
- Statistique. Soit

$$\Sigma_1 = R_1 + \cdots + R_n \quad \text{et} \quad W_{YX} = \Sigma_1 - \frac{n(n+1)}{2}$$

- Propriété : W_{YX} est égal au nombre de paires (X_i, Y_j) (parmi les nm paires possibles) telles que $X_i \geq Y_j$.

Test de Mann-Whitney (1)

- Procédure. On classe les variables $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ par leur rang global et on note $R_1, R_2, \ldots, R_n \in \{1, \ldots, n+m\}$ les rangs associés à l'échantillon X.
- Exemple. $X = (3, 5, 2)\,;\ Y = (1, 4)$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon X sont $R_1 = 3, R_2 = 5, R_3 = 2$.
- Idée. Sous H_0, $R_1, \ldots, R_n$ sont uniformément répartis sur $\{1, \ldots, n+m\}$.
- Statistique. Soit

$$\Sigma_1 = R_1 + \cdots + R_n \quad \text{et} \quad W_{YX} = \Sigma_1 - \frac{n(n+1)}{2}$$

- Propriété : W_{YX} est égal au nombre de paires (X_i, Y_j) (parmi les nm paires possibles) telles que $X_i \geq Y_j$.

Test de Mann-Whitney (1)

- Procédure. On classe les variables $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ par leur rang global et on note $R_1, R_2, \ldots, R_n \in \{1, \ldots, n+m\}$ les rangs associés à l'échantillon X.
- Exemple. $X = (3, 5, 2)$; $Y = (1, 4)$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon X sont $R_1 = 3, R_2 = 5, R_3 = 2$.
- Idée. Sous H_0, $R_1, \ldots, R_n$ sont uniformément répartis sur $\{1, \ldots, n+m\}$.
- Statistique. Soit

$$\Sigma_1 = R_1 + \cdots + R_n \quad \text{et} \quad W_{YX} = \Sigma_1 - \frac{n(n+1)}{2}$$

- Propriété : W_{YX} est égal au nombre de paires (X_i, Y_j) (parmi les nm paires possibles) telles que $X_i \geq Y_j$.

Test de Mann-Whitney (1)

- Procédure. On classe les variables $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ par leur rang global et on note $R_1, R_2, \ldots, R_n \in \{1, \ldots, n+m\}$ les rangs associés à l'échantillon X.
- Exemple. $X = (3, 5, 2)$; $Y = (1, 4)$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon X sont $R_1 = 3, R_2 = 5, R_3 = 2$.
- Idée. Sous H_0, $R_1, \ldots, R_n$ sont uniformément répartis sur $\{1, \ldots, n+m\}$.
- Statistique. Soit

$$\Sigma_1 = R_1 + \cdots + R_n \quad \text{et} \quad W_{YX} = \Sigma_1 - \frac{n(n+1)}{2}$$

- Propriété : W_{YX} est égal au nombre de paires (X_i, Y_j) (parmi les nm paires possibles) telles que $X_i \geq Y_j$.

Test de Mann-Whitney (1)

- Procédure. On classe les variables $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ par leur rang global et on note $R_1, R_2, \ldots, R_n \in \{1, \ldots, n+m\}$ les rangs associés à l'échantillon X.
- Exemple. $X = (3, 5, 2)$; $Y = (1, 4)$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon X sont $R_1 = 3, R_2 = 5, R_3 = 2$.
- Idée. Sous H_0, $R_1, \ldots, R_n$ sont uniformément répartis sur $\{1, \ldots, n+m\}$.
- Statistique. Soit

$$\Sigma_1 = R_1 + \cdots + R_n \quad \text{et} \quad W_{YX} = \Sigma_1 - \frac{n(n+1)}{2}$$

- Propriété : W_{YX} est égal au nombre de paires (X_i, Y_j) (parmi les nm paires possibles) telles que $X_i \geq Y_j$.

Test de Mann-Whitney (1)

- Procédure. On classe les variables $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ par leur rang global et on note $R_1, R_2, \ldots, R_n \in \{1, \ldots, n+m\}$ les rangs associés à l'échantillon X.
- Exemple. $X = (3, 5, 2)$; $Y = (1, 4)$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon X sont $R_1 = 3, R_2 = 5, R_3 = 2$.
- Idée. Sous H_0, $R_1, \ldots, R_n$ sont uniformément répartis sur $\{1, \ldots, n+m\}$.
- Statistique. Soit

$$\Sigma_1 = R_1 + \cdots + R_n \quad \text{et} \quad W_{YX} = \Sigma_1 - \frac{n(n+1)}{2}$$

- Propriété : W_{YX} est égal au nombre de paires (X_i, Y_j) (parmi les nm paires possibles) telles que $X_i \geq Y_j$.

Preuve de la propriété

$$\#\{(i,j), X_i \geq Y_j\} = \sum_{i=1}^{n} \#\{j, X_i \geq Y_j\}$$

Or, par définition de R,

$$R_i = \#\{j, Y_j \leq X_i\} + \#\{i', X_{i'} \leq X_i\}$$

D'où

$$\begin{aligned}
\#\{(i,j), X_i \geq Y_j\} &= \sum_{i=1}^{n} (R_i - \#\{i', X_i \geq X_{i'}\}) \\
&= \sum_{i=1}^{n} R_i - \sum_{i=1}^{n} \#\{i', X_i \geq X_{i'}\} \\
&= \Sigma_1 - \sum_{i=1}^{n} i = W_{YX}
\end{aligned}$$

Test de Mann-Whitney (2)

- Théorème. Sous $H_0 : F = G$, la loi de W_{XY} est libre et symétrique par rapport à $nm/2$. De plus,

$$\mathbb{E}_{H_0}(W_{XY}) = \frac{nm}{2}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_{XY}) = \frac{nm(n+m+1)}{12}$$

$$\text{et} \quad \frac{W_{XY} - \mathbb{E}_{H_0}(W_{XY})}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_{XY})}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

- Test. On rejette H_0 pour les grandes valeurs de $|W_{YX} - nm/2|$.
  - Loi tabulée pour les petites valeurs de n et m (≤ 10).
  - Pour les grandes valeurs, on utilise l'approximation gaussienne.

Test de Mann-Whitney (2)

- Théorème. Sous $H_0 : F = G$, la loi de W_{XY} est libre et symétrique par rapport à $nm/2$. De plus,

$$\mathbb{E}_{H_0}(W_{XY}) = \frac{nm}{2}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_{XY}) = \frac{nm(n+m+1)}{12}$$

$$\text{et} \quad \frac{W_{XY} - \mathbb{E}_{H_0}(W_{XY})}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_{XY})}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

- Test. On rejette H_0 pour les grandes valeurs de $|W_{YX} - nm/2|$.
  - Loi tabulée pour les petites valeurs de n et m (≤ 10).
  - Pour les grandes valeurs, on utilise l'approximation gaussienne.

Test de Mann-Whitney (2)

- Théorème. Sous $H_0 : F = G$, la loi de W_{XY} est libre et symétrique par rapport à $nm/2$. De plus,

$$\mathbb{E}_{H_0}(W_{XY}) = \frac{nm}{2}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_{XY}) = \frac{nm(n+m+1)}{12}$$

$$\text{et} \quad \frac{W_{XY} - \mathbb{E}_{H_0}(W_{XY})}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_{XY})}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

- Test. On rejette H_0 pour les grandes valeurs de $|W_{YX} - nm/2|$.
  - Loi tabulée pour les petites valeurs de n et m (≤ 10).
  - Pour les grandes valeurs, on utilise l'approximation gaussienne.

Test de Mann-Whitney (2)

- Théorème. Sous $H_0 : F = G$, la loi de W_{XY} est libre et symétrique par rapport à $nm/2$. De plus,

$$\mathbb{E}_{H_0}(W_{XY}) = \frac{nm}{2}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_{XY}) = \frac{nm(n+m+1)}{12}$$

$$\text{et} \quad \frac{W_{XY} - \mathbb{E}_{H_0}(W_{XY})}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_{XY})}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

- Test. On rejette H_0 pour les grandes valeurs de $|W_{YX} - nm/2|$.
  - Loi tabulée pour les petites valeurs de n et m (≤ 10).
  - Pour les grandes valeurs, on utilise l'approximation gaussienne.

Tester l'indépendance entre deux variables

Deux v.a. U et V sont dites indépendantes si pour tout $E \subset \mathbb{R}$, $\mathbb{P}[U \in E | V = v]$ ne dépend pas de v.

- Deux variables discrètes finies
  - *Ex : lien entre la couleur des yeux et des cheveux*
  - Tests sur les tables de contingence.
- Une variable discrète finie et une variable continue
  - *Ex : lien entre genre et salaire dans une entreprise*
  - On se ramène à un test de comparaison de populations.
- Deux variables continues : plus complexe
  - *Ex : lien entre PIB et taux d'alphabétisation*
  - première solution : hypothèses paramétriques
  - deuxième solution : tester un type de dépendance particulière : la corrélation (aussi appelée correlation linéaire).

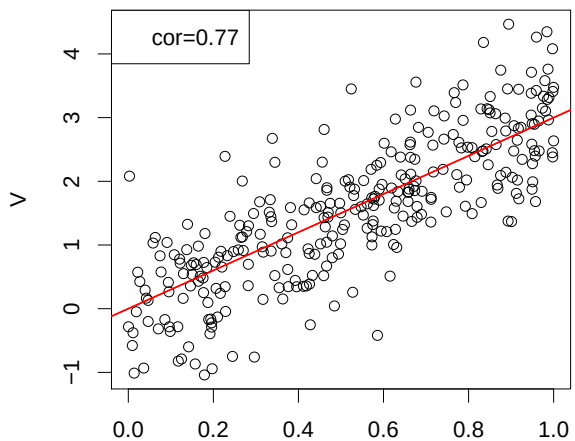
Correlation (linéaire)

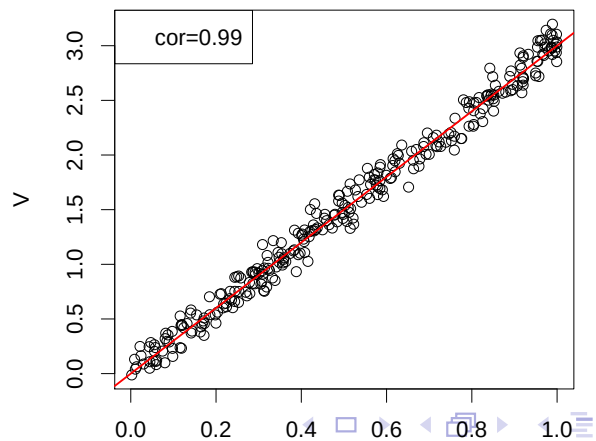
- Covariance et corrélation entre deux v.a. réelles U et V :

$$\begin{cases} \mathsf{cov}(U,V) = \mathbb{E}\left[(U-\mathbb{E}[U])(V-\mathbb{E}[V])\right] = \mathbb{E}[UV] - \mathbb{E}[U]\mathbb{E}[V] \\ \mathsf{cor}(U,V) = \dfrac{\mathsf{cov}(U,V)}{\sqrt{\mathsf{var}(U)\mathsf{var}(V)}} \end{cases}$$

- La corrélation mesure l'intensité du lien linéaire entre deux variables.
- En particulier, si $V = \alpha U + \varepsilon$ avec $\varepsilon \perp U$, alors

$$\mathsf{cor}(U,V) = \sqrt{\frac{\alpha^2 \mathsf{var}(U)}{\alpha^2 \mathsf{var}(U) + \mathsf{var}(\varepsilon)}}$$





Correlation et indépendance

- Indépendance $\Rightarrow$ corrélation nulle

  En effet, si U et V sont indépendantes, $\mathbb{E}[U\,|\,V] = \mathbb{E}[U]$ donc

$$\mathbb{E}[UV] = \mathbb{E}\left[\mathbb{E}[\mathbb{U}|\,V]\,V\right] = \mathbb{E}\left[\mathbb{E}[\mathbb{U}]\,V\right] = \mathbb{E}[\mathbb{U}]\mathbb{E}[V]$$

- Mais corrélation nulle $\not\Rightarrow$ indépendance.

  Exple : soit U et V deux v.a. avec U de distribution uniforme sur $[-1, 1]$ et $V = U^2$.

  - U et V sont dépendantes.
  - U et V ne sont pas corrélées, en effet :

$$\mathbb{E}[U] = 0, \quad \mathbb{E}[UV] = \mathbb{E}[U^3] = 0 \quad \Rightarrow \quad \mathrm{cor}(U, V) = 0$$

Correlation et indépendance

- Indépendance $\Rightarrow$ corrélation nulle

  En effet, si U et V sont indépendantes, $\mathbb{E}[U|V] = \mathbb{E}[U]$ donc

$$\mathbb{E}[UV] = \mathbb{E}\left[\mathbb{E}[\mathbb{U}|V]\,V\right] = \mathbb{E}\left[\mathbb{E}[\mathbb{U}]\,V\right] = \mathbb{E}[\mathbb{U}]\mathbb{E}[V]$$

- Mais corrélation nulle $\not\Rightarrow$ indépendance.

  Exple : soit U et V deux v.a. avec U de distribution uniforme sur $[-1, 1]$ et $V = U^2$.

  - U et V sont dépendantes.
  - U et V ne sont pas corrélées, en effet :

$$\mathbb{E}[U] = 0, \quad \mathbb{E}[UV] = \mathbb{E}[U^3] = 0 \quad \Rightarrow \quad \text{cor}(U, V) = 0$$

Correlation et indépendance

- Indépendance $\Rightarrow$ corrélation nulle

  En effet, si U et V sont indépendantes, $\mathbb{E}[U|V] = \mathbb{E}[U]$ donc

$$\mathbb{E}[UV] = \mathbb{E}\left[\mathbb{E}[\mathbb{U}|V]\,V\right] = \mathbb{E}\left[\mathbb{E}[\mathbb{U}]\,V\right] = \mathbb{E}[\mathbb{U}]\mathbb{E}[V]$$

- Mais corrélation nulle $\not\Rightarrow$ indépendance.

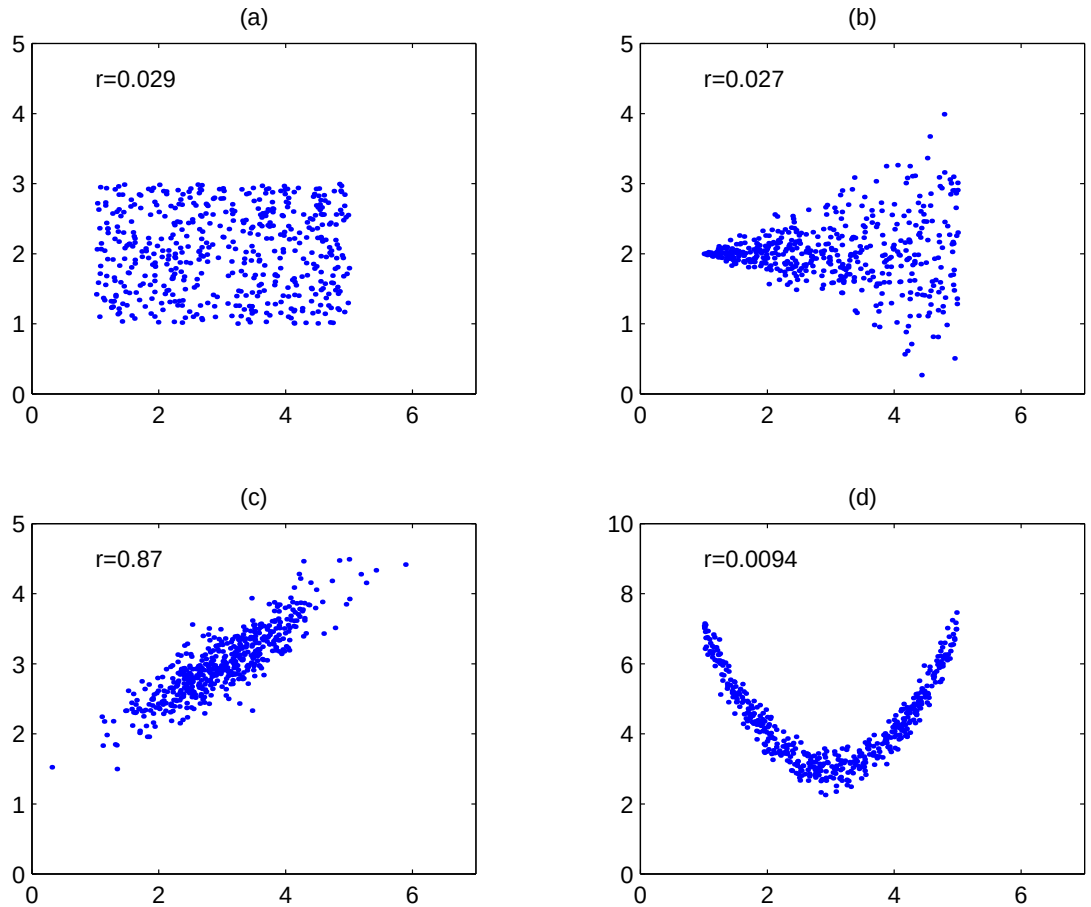
  Exple : soit U et V deux v.a. avec U de distribution uniforme sur $[-1, 1]$ et $V = U^2$.

  - U et V sont dépendantes.
  - U et V ne sont pas corrélées, en effet :

$$\mathbb{E}[U] = 0, \quad \mathbb{E}[UV] = \mathbb{E}[U^3] = 0 \quad \Rightarrow \quad \text{cor}(U, V) = 0$$

Exemples de corrélation (ou non)

Soient $(X_1, \ldots, X_n)$ i.i.d. et $(Y_1, \ldots, Y_n)$ i.i.d. deux échantillons appariés. Les graphiques suivants représentent les points (X_i, Y_i).



Tests de corrélation : contexte

- On dispose de deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ de v.a. réelles et appariées. On notera X (resp Y) la v.a. de même distribution que X_i (resp. Y_i).
  $\hookrightarrow$ Ex : deux quantités X et Y mesurées sur un ensemble d'individus.
- Heuristiquement, on veut savoir si il y a un "lien" entre X et Y, mais l'indépendance est difficile à tester, on considère donc le test induit : "X et Y sont-ils corrélés ?"
- On teste H_0 : "X et Y sont non corrélées" contre H_1 : "X et Y sont corrélées".

Tests de corrélation : contexte

- On dispose de deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ de v.a. réelles et appariées. On notera X (resp Y) la v.a. de même distribution que X_i (resp. Y_i).
  $\hookrightarrow$ Ex : deux quantités X et Y mesurées sur un ensemble d'individus.
- Heuristiquement, on veut savoir si il y a un "lien" entre X et Y, mais l'indépendance est difficile à tester, on considère donc le test induit : "X et Y sont-ils corrélés ?"
- On teste H_0 : "X et Y sont non corrélées" contre H_1 : "X et Y sont corrélées".

Tests de corrélation : contexte

- On dispose de deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ de v.a. réelles et appariées. On notera X (resp Y) la v.a. de même distribution que X_i (resp. Y_i).
  $\hookrightarrow$ Ex : deux quantités X et Y mesurées sur un ensemble d'individus.
- Heuristiquement, on veut savoir si il y a un "lien" entre X et Y, mais l'indépendance est difficile à tester, on considère donc le test induit : "X et Y sont-ils corrélés ?"
- On teste H_0 : "X et Y sont non corrélées" contre H_1 : "X et Y sont corrélées".

Tests de corrélation : contexte

- On dispose de deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ de v.a. réelles et appariées. On notera X (resp Y) la v.a. de même distribution que X_i (resp. Y_i).
  $\hookrightarrow$ Ex : deux quantités X et Y mesurées sur un ensemble d'individus.
- Heuristiquement, on veut savoir si il y a un "lien" entre X et Y, mais l'indépendance est difficile à tester, on considère donc le test induit : "X et Y sont-ils corrélés ?"
- On teste H_0 : "X et Y sont non corrélées" contre H_1 : "X et Y sont corrélées".

Corrélation de Pearson

Coefficient de corrélation de Pearson :
$\text{cor}(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$ est estimée par :

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}.$$

- $-1 \leq r \leq 1$ avec égalité lorsque la relation entre $X = \alpha Y$.
- La distribution exacte de r sous H_0 dépend des distributions de X et Y mais r suit asymptotiquement une loi gaussienne centrée.
- La fonction `cor.test` de R avec `'pearson'` implémente le test de Pearson sous approximation gaussienne (attention aux petites tailles d'échantillons !)

Corrélation de Pearson

Coefficient de corrélation de Pearson :

$\text{cor}(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$ est estimée par :

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}.$$

- $-1 \leq r \leq 1$ avec égalité lorsque la relation entre $X = \alpha Y$.
- La distribution exacte de r sous H_0 dépend des distributions de X et Y mais r suit asymptotiquement une loi gaussienne centrée.
- La fonction `cor.test` de R avec `'pearson'` implémente le test de Pearson sous approximation gaussienne (attention aux petites tailles d'échantillons !)

Corrélation de Pearson

Coefficient de corrélation de Pearson :

$\text{cor}(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$ est estimée par :

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}.$$

- $-1 \leq r \leq 1$ avec égalité lorsque la relation entre $X = \alpha Y$.
- La distribution exacte de r sous H_0 dépend des distributions de X et Y mais r suit asymptotiquement une loi gaussienne centrée.
- La fonction `cor.test` de R avec `'pearson'` implémente le test de Pearson sous approximation gaussienne (attention aux petites tailles d'échantillons !)

Corrélation de Pearson

Coefficient de corrélation de Pearson :

$\text{cor}(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$ est estimée par :

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}.$$

- $-1 \leq r \leq 1$ avec égalité lorsque la relation entre $X = \alpha Y$.
- La distribution exacte de r sous H_0 dépend des distributions de X et Y mais r suit asymptotiquement une loi gaussienne centrée.
- La fonction `cor.test` de R avec `'pearson'` implémente le test de Pearson sous approximation gaussienne (attention aux petites tailles d'échantillons !)

Corrélation de Pearson

Coefficient de corrélation de Pearson :

$\text{cor}(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$ est estimée par :

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}.$$

- $-1 \leq r \leq 1$ avec égalité lorsque la relation entre $X = \alpha Y$.
- La distribution exacte de r sous H_0 dépend des distributions de X et Y mais r suit asymptotiquement une loi gaussienne centrée.
- La fonction `cor.test` de R avec `'pearson'` implémente le test de Pearson sous approximation gaussienne (attention aux petites tailles d'échantillons !)

Lien entre corrélation de Pearson et régression linéaire.

- Modèle linéaire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (\varepsilon_i)_{i=1,\dots,n} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2)$$

$$\text{et} \quad \left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Résultat

$$\hat{\beta}_1 = r \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

- Le test de student implémenté dans `lm` est équivalent au test de Pearson `cor.test`.

- On constate bien que le test de corrélation de Pearson est paramétrique. Pour obtenir une statistique libre en loi sous H_0, il faut se "débarrasser" des valeurs prises par les variables. Une possibilité est de passer par les rangs des variables.

Lien entre corrélation de Pearson et régression linéaire.

- Modèle linéaire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (\varepsilon_i)_{i=1,\dots,n} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2)$$

$$\text{et} \quad \left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Résultat

$$\hat{\beta}_1 = r \sqrt{\frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}}$$

- Le test de student implémenté dans `lm` est équivalent au test de Pearson `cor.test`.

- On constate bien que le test de corrélation de Pearson est paramétrique. Pour obtenir une statistique libre en loi sous H_0, il faut se "débarrasser" des valeurs prises par les variables. Une possibilité est de passer par les rangs des variables.

Lien entre corrélation de Pearson et régression linéaire.

- Modèle linéaire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (\varepsilon_i)_{i=1,\dots,n} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2)$$

$$\text{et} \quad \left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Résultat

$$\hat{\beta}_1 = r \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

- Le test de student implémenté dans `lm` est équivalent au test de Pearson `cor.test`.
- On constate bien que le test de corrélation de Pearson est paramétrique. Pour obtenir une statistique libre en loi sous H_0, il faut se "débarrasser" des valeurs prises par les variables. Une possibilité est de passer par les rangs des variables.

Lien entre corrélation de Pearson et régression linéaire.

- Modèle linéaire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (\varepsilon_i)_{i=1,\dots,n} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2)$$

$$\text{et} \quad \left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Résultat

$$\hat{\beta}_1 = r \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

- Le test de student implémenté dans `lm` est équivalent au test de Pearson `cor.test`.

- On constate bien que le test de corrélation de Pearson est paramétrique. Pour obtenir une statistique libre en loi sous H_0, il faut se "débarrasser" des valeurs prises par les variables. Une possibilité est de passer par les rangs des variables.

Lien entre corrélation de Pearson et régression linéaire.

- Modèle linéaire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (\varepsilon_i)_{i=1,\dots,n} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2)$$

$$\text{et} \quad \left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Résultat

$$\hat{\beta}_1 = r \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

- Le test de student implémenté dans `lm` est équivalent au test de Pearson `cor.test`.

- On constate bien que le test de corrélation de Pearson est paramétrique. Pour obtenir une statistique libre en loi sous H_0, il faut se "débarrasser" des valeurs prises par les variables. Une possibilité est de passer par les rangs des variables.

Test de corrélation des rangs de Spearman

- On teste donc H_0 "X_i, Y_i non corrélées" contre H_1 : "X_i, Y_i sont en relation monotone".
- Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux échantillons appariés et $(R_1, \ldots, R_n)$, $(S_1, \ldots, S_n)$ leurs statistiques de rang. Le coefficient de corrélation de Spearman est égal au coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (S_i - \bar{S})^2}}.$$

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi, et on rejette H_0 pour des grandes valeurs de $|\rho|$.
- Dans R : fonction `cor.test` avec `'spearman'`

Test de corrélation des rangs de Spearman

- On teste donc H_0 "X_i, Y_i non corrélées" contre H_1 : "X_i, Y_i sont en relation monotone".
- Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux échantillons appariés et $(R_1, \ldots, R_n)$, $(S_1, \ldots, S_n)$ leurs statistiques de rang. Le coefficient de corrélation de Spearman est égal au coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (S_i - \bar{S})^2}}.$$

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi, et on rejette H_0 pour des grandes valeurs de $|\rho|$.
- Dans R : fonction `cor.test` avec `'spearman'`

Test de corrélation des rangs de Spearman

- On teste donc H_0 "X_i, Y_i non corrélées" contre H_1 : "X_i, Y_i sont en relation monotone".
- Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux échantillons appariés et $(R_1, \ldots, R_n)$, $(S_1, \ldots, S_n)$ leurs statistiques de rang. Le coefficient de corrélation de Spearman est égal au coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (S_i - \bar{S})^2}}.$$

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi, et on rejette H_0 pour des grandes valeurs de $|\rho|$.
- Dans R : fonction `cor.test` avec `'spearman'`

Test de corrélation des rangs de Spearman

- On teste donc H_0 "X_i, Y_i non corrélées" contre H_1 : "X_i, Y_i sont en relation monotone".
- Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux échantillons appariés et $(R_1, \ldots, R_n)$, $(S_1, \ldots, S_n)$ leurs statistiques de rang. Le coefficient de corrélation de Spearman est égal au coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (S_i - \bar{S})^2}}.$$

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi, et on rejette H_0 pour des grandes valeurs de $|\rho|$.
- Dans R : fonction `cor.test` avec 'spearman'

Test de corrélation des rangs de Spearman

- On teste donc H_0 "X_i, Y_i non corrélées" contre H_1 : "X_i, Y_i sont en relation monotone".
- Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux échantillons appariés et $(R_1, \ldots, R_n)$, $(S_1, \ldots, S_n)$ leurs statistiques de rang. Le coefficient de corrélation de Spearman est égal au coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (S_i - \bar{S})^2}}.$$

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi, et on rejette H_0 pour des grandes valeurs de $|\rho|$.
- Dans R : fonction `cor.test` avec 'spearman'

Test de corrélation des rangs de Spearman

- On teste donc H_0 "X_i, Y_i non corrélées" contre H_1 : "X_i, Y_i sont en relation monotone".
- Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux échantillons appariés et $(R_1, \ldots, R_n)$, $(S_1, \ldots, S_n)$ leurs statistiques de rang. Le coefficient de corrélation de Spearman est égal au coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (S_i - \bar{S})^2}}.$$

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi, et on rejette H_0 pour des grandes valeurs de $|\rho|$.
- Dans R : fonction `cor.test` avec 'spearman'

Test de corrélation des rangs de Spearman

- On teste donc H_0 "X_i, Y_i non corrélées" contre H_1 : "X_i, Y_i sont en relation monotone".
- Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux échantillons appariés et $(R_1, \ldots, R_n)$, $(S_1, \ldots, S_n)$ leurs statistiques de rang. Le coefficient de corrélation de Spearman est égal au coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (S_i - \bar{S})^2}}.$$

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi, et on rejette H_0 pour des grandes valeurs de $|\rho|$.
- Dans R : fonction `cor.test` avec 'spearman'

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Conclusion sur les tests non paramétriques

- Tests paramétriques : on suppose que les variables suivent une loi paramétrique (souvent gaussienne ou associée)
  - Approximation gaussienne raisonnable pour de grands échantillons
  - Il est indispensable de vérifier les hypothèses paramétriques pour de petits échantillons !
- Tests non paramétriques : la distribution de la statistique de test sous H_0 est libre en loi i.e. elle ne dépend pas de la loi des variables considérées.
  - Préférable pour de petits échantillons ou quand les approximations paramétriques sont mauvaises.
  - La distribution sous H_0 pour les grandes valeurs de n peut-etre calculée par une approximation gaussienne basée sur le TCL. Mais ceci est différent d'un test paramétrique car on ne suppose pas que les variables sont gaussiennes.
- Contrairement à l'estimation de fonction, les tests NP sont à privilégier pour de petites tailles d'échantillon !

Estimation de densité

- Contexte et exemple introductif
- Décomposition biais-variance : calculs
- Estimation par projection
- Estimation par noyaux
- Conclusion
- Vitesse minimax

Contexte de l'estimation de densité (univariée)

- Densité d'une variable X :
  $f(x) = \lim_{\delta x \to 0} \mathbb{P}\left[X \in [x - \delta x, x + \delta x]\right] / 2\delta x$
- Observations : $X_1, \ldots, X_n$ v.a. i.i.d. réelles de fdr F et admettant une densité $f = F'$.
- But : estimer (à partir des observations) f en faisant le moins d'hypothèses possibles sur cette densité.
- Typiquement, on supposera que $f \in \mathcal{F}$ espace fonctionnel et on notera $\hat{f}_n$ un estimateur de f.

Objectifs

Obtenir des informations de nature géométrique sur la distribution des variables. Ex :

- Combien de modes ?
- Zones peu denses ? très denses ?

Introduire les notions (calculs plus simples qu'en régression)

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Mesure de la qualité d'un estimateur : risque

- Fonction de perte sur $\mathcal{F}$: mesure l'écart entre $\hat{f}_n$ et f. Ex :
  - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$.
  - $d(f,g) = \|f-g\|_\infty$.
  - $d(f,g) = (f(x_0)-g(x_0))^2$ où x_0 fixé.
- L'erreur $d(\hat{f}_n, f)$ dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}[d(\hat{f}_n, f)].$$

  C'est l'erreur moyenne commise en estimant f par $\hat{f}_n$.
- On va considérer deux fonctions de risque :
  - Risque quadratique intégré : MISE = mean integrated squared error.

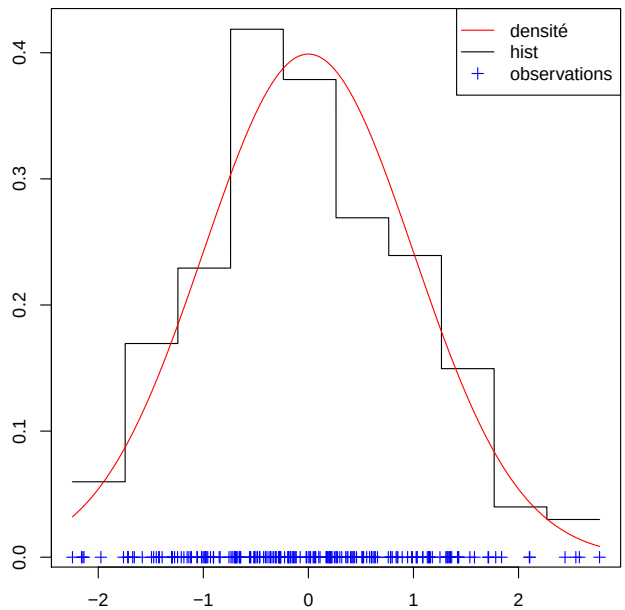
$$R(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|_2^2 = \mathbb{E}\int_x (\hat{f}_n(x) - f(x))^2 dx.$$

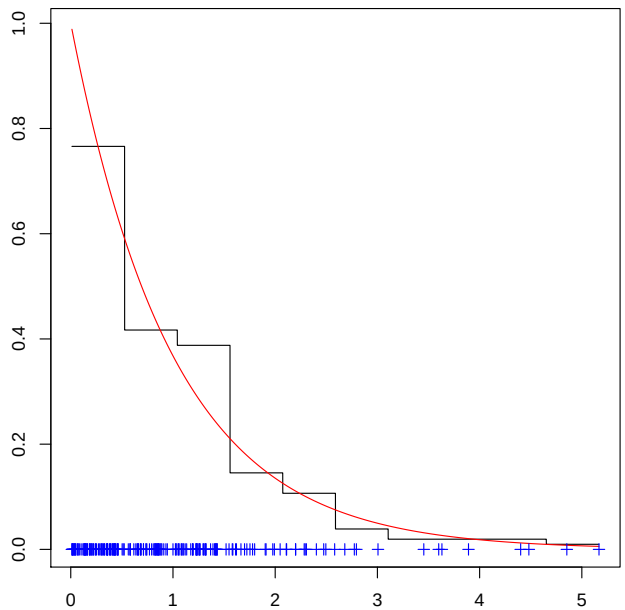
  - Risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right]$$

Exemple introductif : histogrammes réguliers à D intervalles

- Soit $(X_i)_{i=1,\dots,n}$ i.i.d. de densité f. On considére l'estimateur $\hat{f}^D$ par histogramme régulier à D morceaux
- L'intervalle d'estimation est découpé en D sous-intervalles de même longueur
- Chaque "morceau" de l'histogramme est égal à "proportion d'observations dans l'intervalle $\times$ cte"
- Exemples :





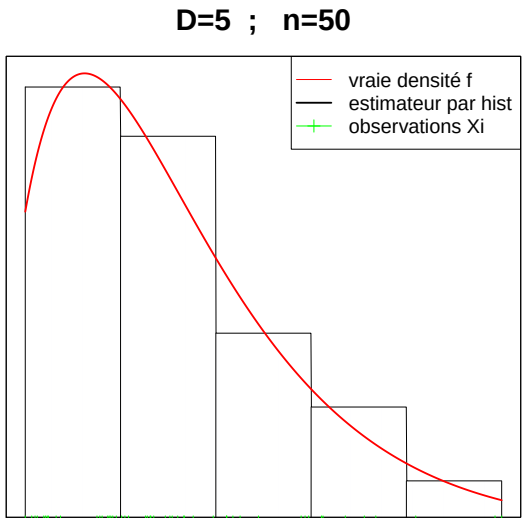
Erreur d'estimation de deux natures : biais et variance

- Soit f^D l'histogramme à D morceaux qui approche le mieux f

  $\hookrightarrow$ f^D est inconnu

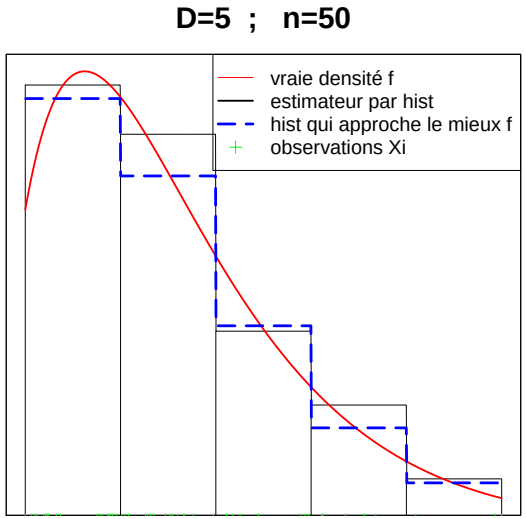
  $\hookrightarrow$ On montrera que $f_D = \mathbb{E}[\hat{f}_D]$.
- $d(f, f^D)$ = biais : mesure comment l'espace des histogrammes à D morceaux approche f.
- $\mathbb{E}\left[d(\hat{f}^D, f^D)\right]$ = variance : mesure la somme de l'erreur d'estimation de chaque morceau de l'histogramme.

Erreur d'estimation de deux natures : biais et variance



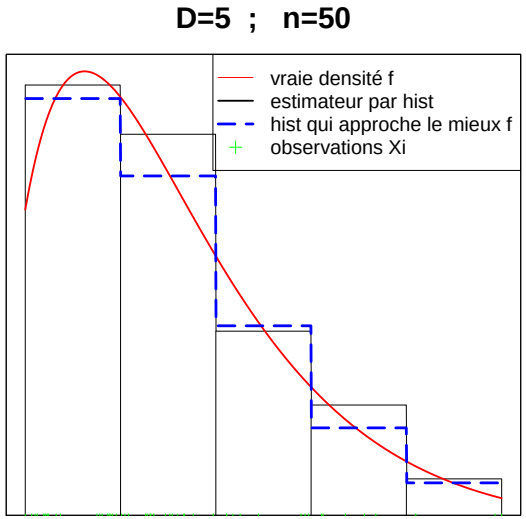
- Soit f^D l'histogramme à D morceaux qui approche le mieux f
  $\hookrightarrow$ f^D est inconnu
  $\hookrightarrow$ On montrera que $f_D = \mathbb{E}[\hat{f}_D]$.
- $d(f, f^D)$ = biais : mesure comment l'espace des histogrammes à D morceaux approche f.
- $\mathbb{E}\left[d(\hat{f}^D, f^D)\right]$ = variance : mesure la somme de l'erreur d'estimation de chaque morceau de l'histogramme.

Erreur d'estimation de deux natures : biais et variance



- Soit f^D l'histogramme à D morceaux qui approche le mieux f
  $\hookrightarrow$ f^D est inconnu
  $\hookrightarrow$ On montrera que $f_D = \mathbb{E}[\hat{f}_D]$.
- $d(f, f^D)$ = biais : mesure comment l'espace des histogrammes à D morceaux approche f.
- $\mathbb{E}\left[d(\hat{f}^D, f^D)\right]$ = variance : mesure la somme de l'erreur d'estimation de chaque morceau de l'histogramme.

Erreur d'estimation de deux natures : biais et variance



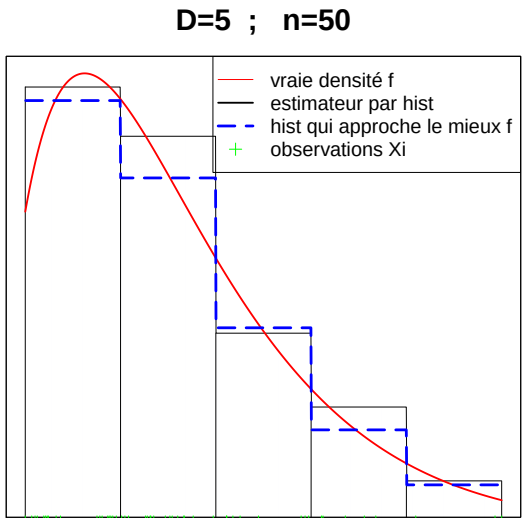
- Soit f^D l'histogramme à D morceaux qui approche le mieux f

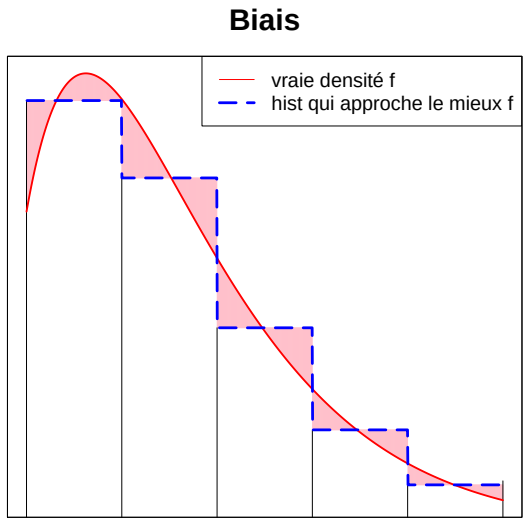
  $\hookrightarrow$ f^D est inconnu

  $\hookrightarrow$ On montrera que $f_D = \mathbb{E}[\hat{f}_D]$.

- $d(f, f^D)$ = biais : mesure comment l'espace des histogrammes à D morceaux approche f.
- $\mathbb{E}\left[d(\hat{f}^D, f^D)\right]$ = variance : mesure la somme de l'erreur d'estimation de chaque morceau de l'histogramme.

Erreur d'estimation de deux natures : biais et variance



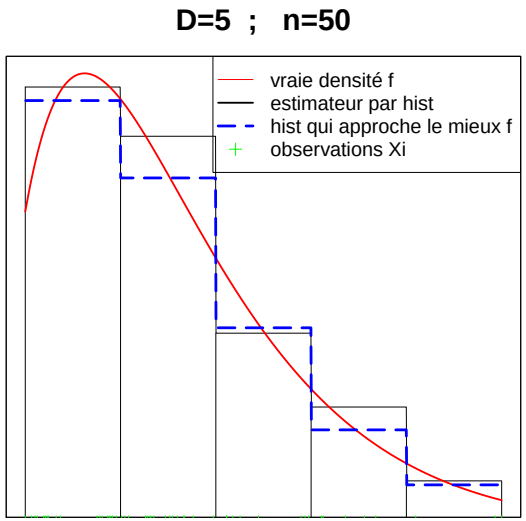


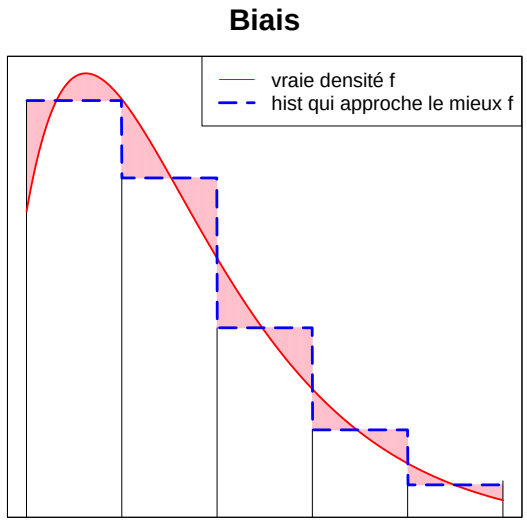
- Soit f^D l'histogramme à D morceaux qui approche le mieux f

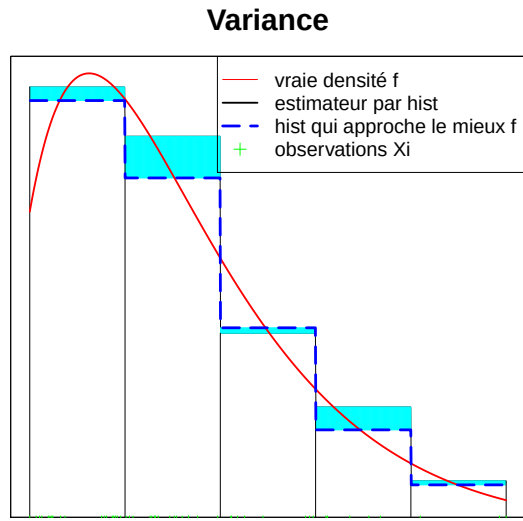
  $\hookrightarrow$ f^D est inconnu

  $\hookrightarrow$ On montrera que $f_D = \mathbb{E}[\hat{f}_D]$.
- $d(f, f^D)$ = biais : mesure comment l'espace des histogrammes à D morceaux approche f.
- $\mathbb{E}\left[d(\hat{f}^D, f^D)\right]$ = variance : mesure la somme de l'erreur d'estimation de chaque morceau de l'histogramme.

Erreur d'estimation de deux natures : biais et variance





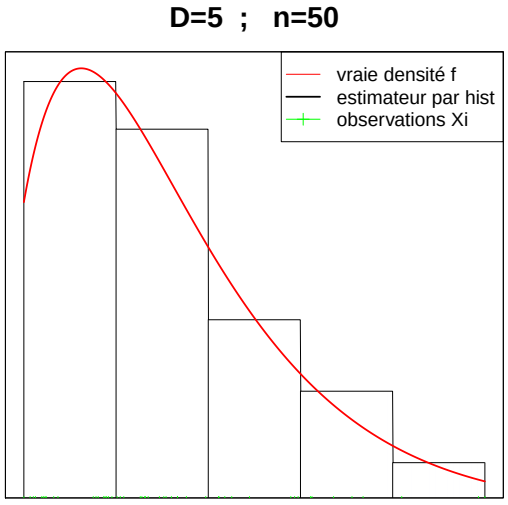


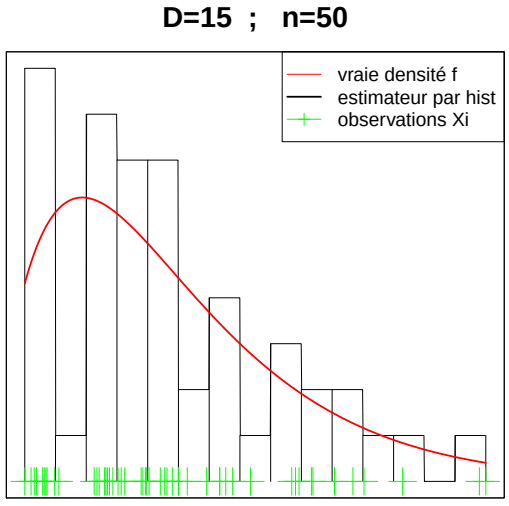
- Soit f^D l'histogramme à D morceaux qui approche le mieux f

  $\hookrightarrow$ f^D est inconnu

  $\hookrightarrow$ On montrera que $f_D = \mathbb{E}[\hat{f}_D]$.
- $d(f, f^D)$ = biais : mesure comment l'espace des histogrammes à D morceaux approche f.
- $\mathbb{E}\left[d(\hat{f}^D, f^D)\right]$ = variance : mesure la somme de l'erreur d'estimation de chaque morceau de l'histogramme.

Décompo biais-variance : dimension et taille d'échantillon

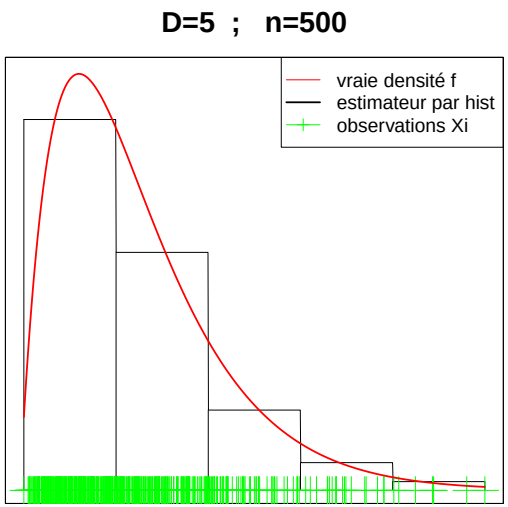


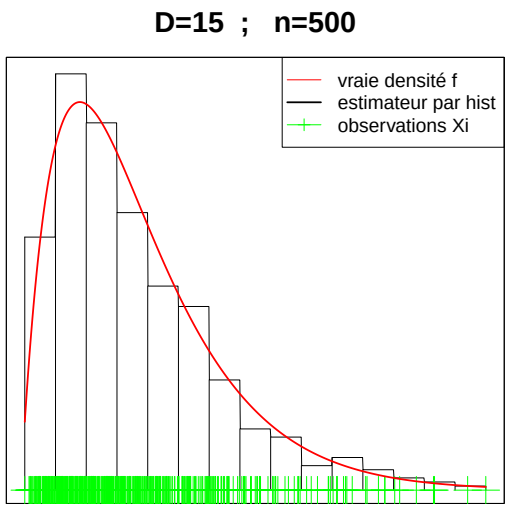


- Biais $d(f, f^D)$
  - $\searrow$ quand D $\nearrow$
  - ne dépend pas de n
- Variance $d(\hat{f}^D, f^D)$
  - $\nearrow$ quand D $\nearrow$
  - $\searrow$ quand n $\nearrow$
- En connaissant la vraie fonction f (ce qui n'est pas le cas en estimation !) on voit que :
  - Pour $n = 50$: meilleur estimateur pour $D = 5$
  - Pour $n = 500$: meilleur estimateur pour $D = 15$

Décompo biais-variance : dimension et taille d'échantillon

- Biais $d(f, f^D)$
  - $\searrow$ quand D $\nearrow$
  - ne dépend pas de n
- Variance $d(\hat{f}^D, f^D)$
  - $\nearrow$ quand D $\nearrow$
  - $\searrow$ quand n $\nearrow$

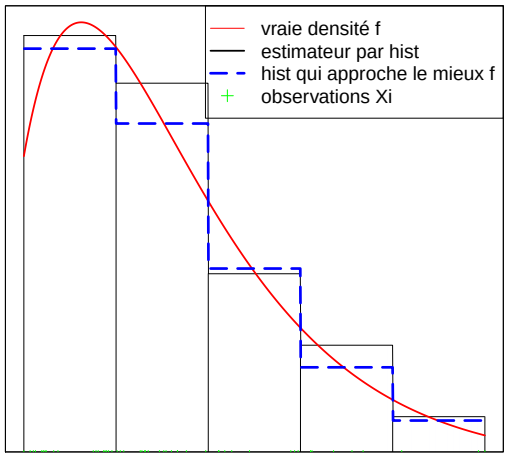




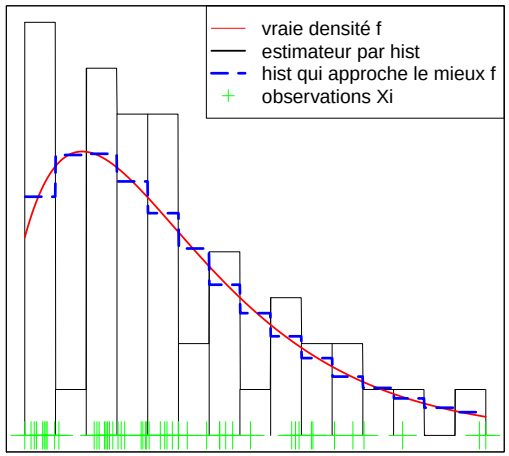
- En connaissant la vraie fonction f (ce qui n'est pas le cas en estimation !) on voit que :
  - Pour $n = 50$: meilleur estimateur pour $D = 5$
  - Pour $n = 500$: meilleur estimateur pour $D = 15$

Décompo biais-variance : dimension et taille d'échantillon

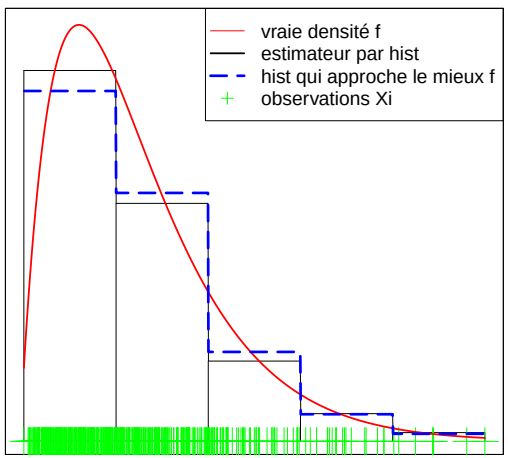
D=5 ; n=50



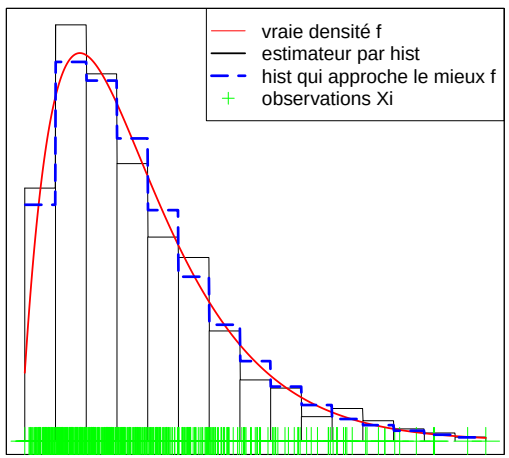
D=15 ; n=50



D=5 ; n=500



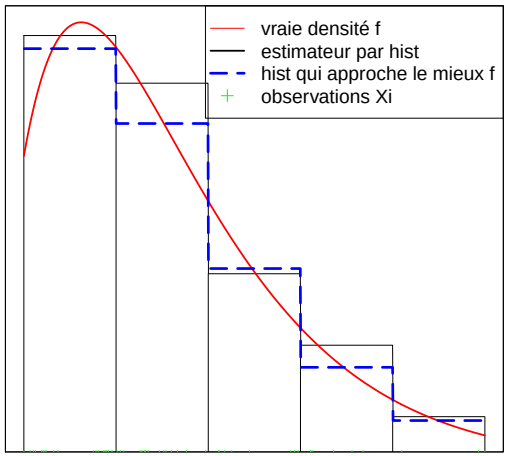
D=15 ; n=500



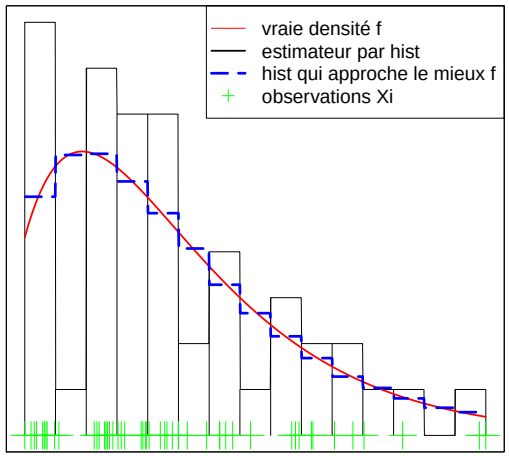
- Biais $d(f, f^D)$
  - $\searrow$ quand D $\nearrow$
  - ne dépend pas de n
- Variance $d(\hat{f}^D, f^D)$
  - $\nearrow$ quand D $\nearrow$
  - $\searrow$ quand n $\nearrow$
- En connaissant la vraie fonction f (ce qui n'est pas le cas en estimation !) on voit que :
  - Pour $n = 50$: meilleur estimateur pour $D = 5$
  - Pour $n = 500$: meilleur estimateur pour $D = 15$

Décompo biais-variance : dimension et taille d'échantillon

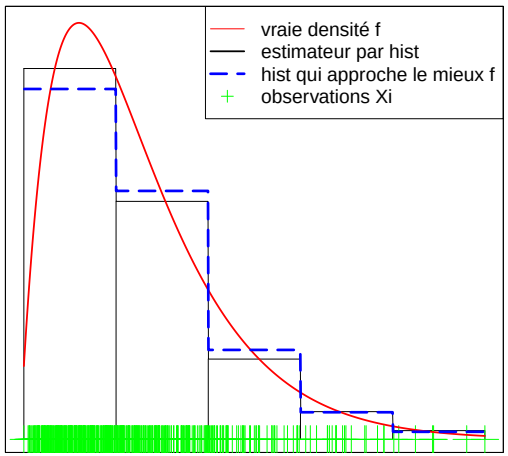
D=5 ; n=50



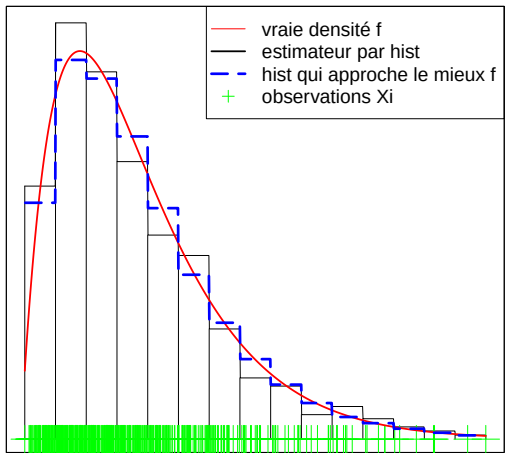
D=15 ; n=50



D=5 ; n=500



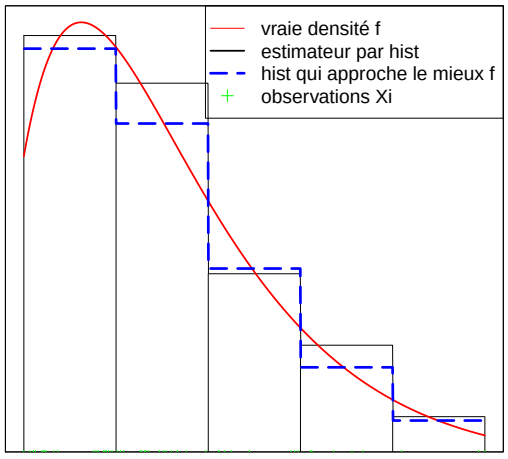
D=15 ; n=500



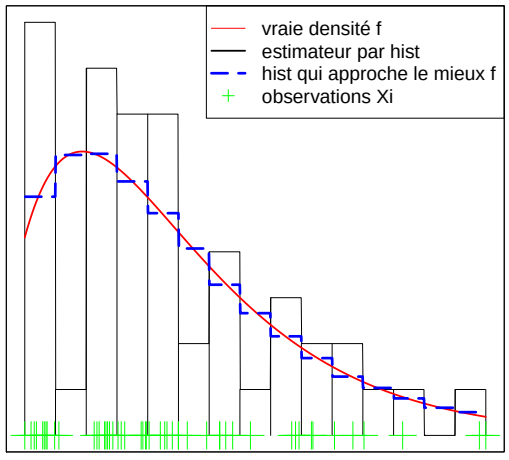
- Biais $d(f, f^D)$
  - $\searrow$ quand D $\nearrow$
  - ne dépend pas de n
- Variance $d(\hat{f}^D, f^D)$
  - $\nearrow$ quand D $\nearrow$
  - $\searrow$ quand n $\nearrow$
- En connaissant la vraie fonction f (ce qui n'est pas le cas en estimation !) on voit que :
  - Pour $n = 50$: meilleur estimateur pour $D = 5$
  - Pour $n = 500$: meilleur estimateur pour $D = 15$

Décompo biais-variance : dimension et taille d'échantillon

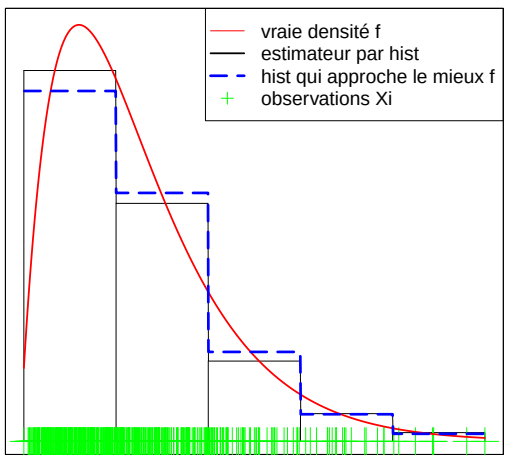
D=5 ; n=50



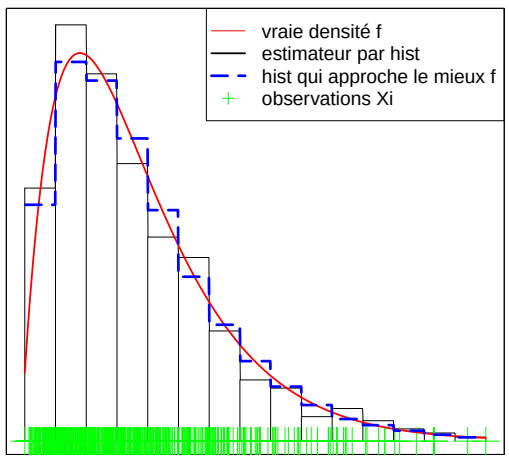
D=15 ; n=50



D=5 ; n=500



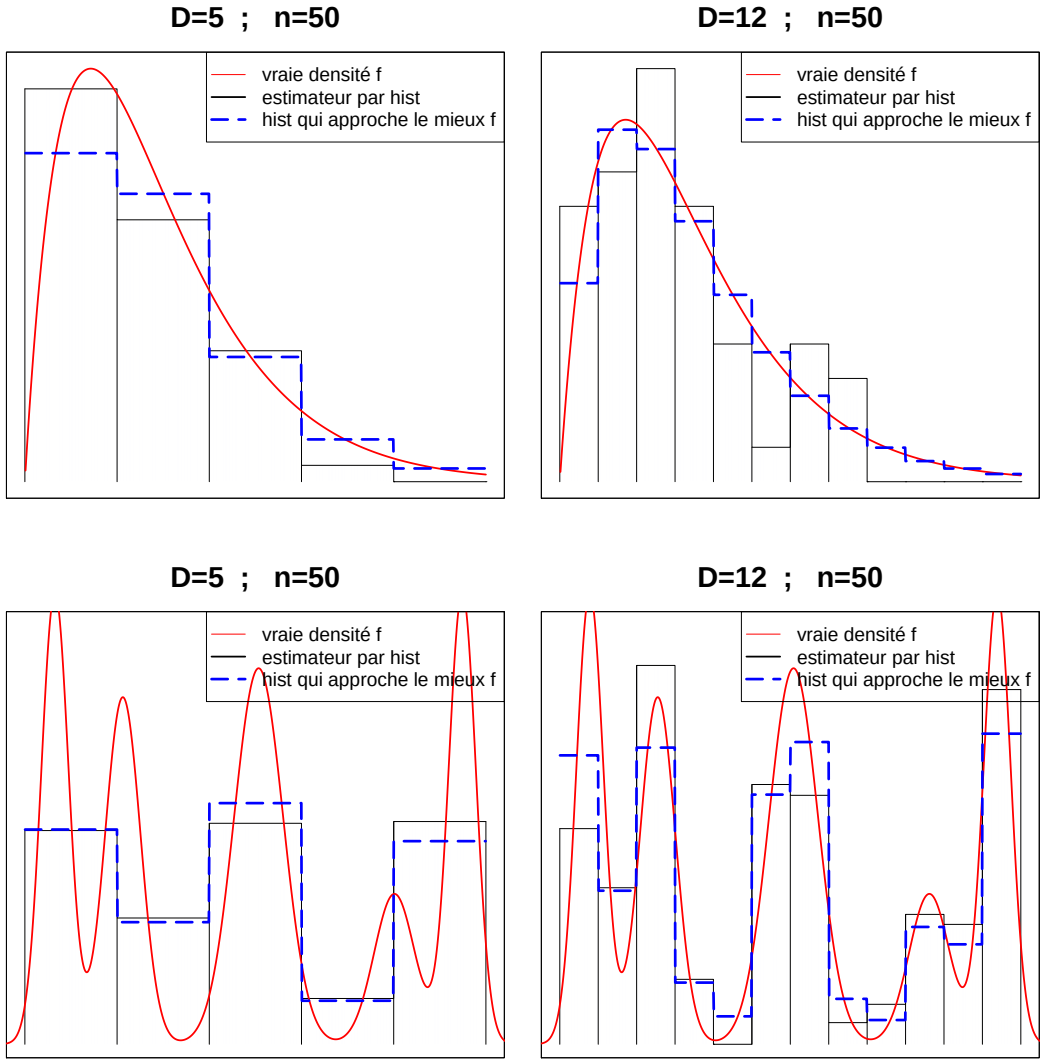
D=15 ; n=500



- Biais $d(f, f^D)$
  - $\searrow$ quand D $\nearrow$
  - ne dépend pas de n
- Variance $d(\hat{f}^D, f^D)$
  - $\nearrow$ quand D $\nearrow$
  - $\searrow$ quand n $\nearrow$
- En connaissant la vraie fonction f (ce qui n'est pas le cas en estimation !) on voit que :
  - Pour $n = 50$: meilleur estimateur pour $D = 5$
  - Pour $n = 500$: meilleur estimateur pour $D = 15$

Décomposition biais-variance et régularité

- Le D optimal dépend également de la régularité de la fonction

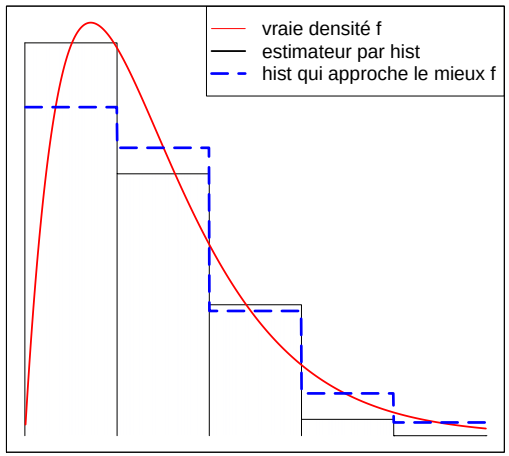


- Plus la fonction est irrégulière, plus le D optimal est grand.

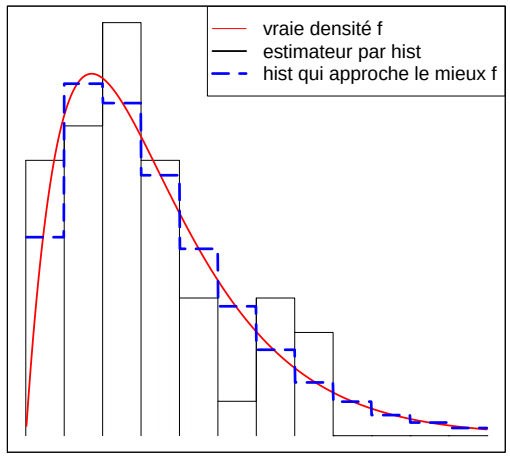
Décomposition biais-variance et régularité

- Le D optimal dépend également de la régularité de la fonction

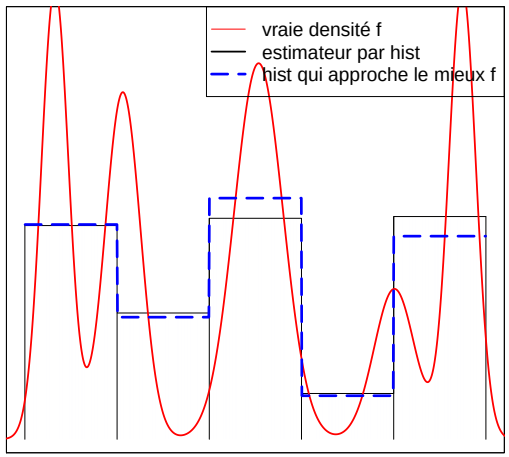
D=5 ; n=50



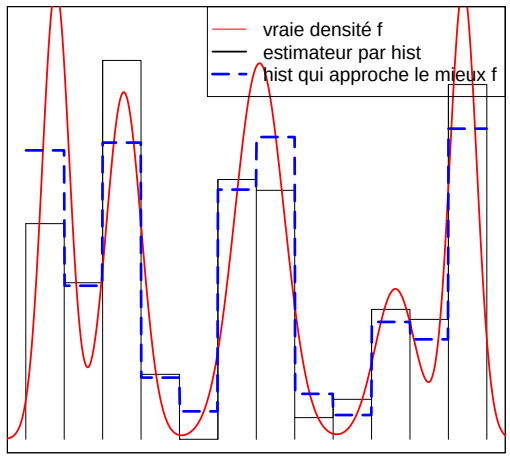
D=12 ; n=50



D=5 ; n=50



D=12 ; n=50



- Plus la fonction est irrégulière, plus le D optimal est grand.

Décomposition biais-variance du MSE (risque quadratique en x_0)

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[\left(\hat{f}_n(x_0) - f(x_0)\right)^2\right]$$

$$= \mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2\right] + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right] + \text{dp}$$

Or $\left(\hat{f}_n(x_0) - \mathbb{E}\left[\hat{f}_n(x_0)\right]\right)^2$ est déterministe donc $\mathbb{E}$ disparaît, et le double produit vérifie

$$\begin{aligned} \text{dp} &= 2\mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right] \\ &= 2\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right] = 0 \end{aligned}$$

Donc $$R_{x_0}(\hat{f}_n, f) = \left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2 + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right]$$

$$= \text{Biais} + \text{Var}(\hat{f}_n(x_0)).$$

Décomposition biais-variance du MSE (risque quadratique en x_0)

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[\left(\hat{f}_n(x_0) - f(x_0)\right)^2\right]$$

$$= \mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2\right] + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right] + \text{dp}$$

Or $\left(\hat{f}_n(x_0) - \mathbb{E}\left[\hat{f}_n(x_0)\right]\right)^2$ est déterministe donc $\mathbb{E}$ disparaît, et le double produit vérifie

$$\text{dp} = 2\mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right]$$

$$= 2\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right] = 0$$

$$\text{Donc} \quad R_{x_0}(\hat{f}_n, f) = \left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2 + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right]$$

$$= \text{Biais} + \text{Var}(\hat{f}_n(x_0)).$$

Décomposition biais-variance du MSE (risque quadratique en x_0)

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[\left(\hat{f}_n(x_0) - f(x_0)\right)^2\right]$$

$$= \mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2\right] + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right] + \text{dp}$$

Or $\left(\hat{f}_n(x_0) - \mathbb{E}\left[\hat{f}_n(x_0)\right]\right)^2$ est déterministe donc $\mathbb{E}$ disparaît, et le double produit vérifie

$$\begin{aligned} \text{dp} &= 2\mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right] \\ &= 2\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right] = 0 \end{aligned}$$

Donc $R_{x_0}(\hat{f}_n, f) = \left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2 + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right]$

$= \text{Biais} + \text{Var}(\hat{f}_n(x_0))$.

Décomposition biais-variance du MSE (risque quadratique en x_0)

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}\left[\left(\hat{f}_n(x_0) - f(x_0)\right)^2\right]$$

$$= \mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2\right] + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right] + \text{dp}$$

Or $\left(\hat{f}_n(x_0) - \mathbb{E}\left[\hat{f}_n(x_0)\right]\right)^2$ est déterministe donc $\mathbb{E}$ disparaît, et le double produit vérifie

$$\begin{aligned} \text{dp} &= 2\mathbb{E}\left[\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right] \\ &= 2\left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)\mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)\right] = 0 \end{aligned}$$

Donc $$R_{x_0}(\hat{f}_n, f) = \left(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right)^2 + \mathbb{E}\left[\left(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right)^2\right]$$

$$= \text{Biais} + \text{Var}(\hat{f}_n(x_0)).$$

Décomposition "biais-variance" du MISE (risque intégré)

- Par un calcul similaire au MSE, on obtient pour le MISE

$$\begin{aligned} R(\hat{f}_n, f) &= \mathbb{E}\left[\|\hat{f}_n - f\|_2^2\right] = \mathbb{E}\left[\int_x \left(\hat{f}_n(x) - f(x)\right)^2 dx\right] \\ &= \|\mathbb{E}(\hat{f}_n) - f\|_2^2 + \mathbb{E}\left[\|\hat{f}_n - \mathbb{E}(\hat{f}_n)\|_2^2\right] \\ &= \text{Biais} + \text{variance}. \end{aligned}$$

Compromis biais/variance

- L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.
- On pourra accepter des estimateurs biaisés mais de variance petite, tels que le risque quadratique soit contrôlé.

Oracle

Idéalement, le meilleur estimateur pour le risque R, est

$$f_n^\star = \operatorname*{Argmin}_{\hat{f}_n} R(\hat{f}_n, f).$$

- Pbm : $R(\hat{f}_n, f)$ dépend de la densité f inconnue et n'est donc pas calculable ; $f_n^\star$ n'est pas un estimateur, c'est un oracle.
- Souvent, on dispose d'une famille d'estimateurs $\hat{f}_{\lambda,n}$ dépendants d'un paramètre λ (ex :partition). Alors

$$f_n^\star = \hat{f}_{\lambda^\star,n} \quad \text{avec} \quad \lambda^\star = \operatorname*{Argmin}_{\lambda} R(\hat{f}_{\lambda,n}, f).$$

mais $\hat{f}_{\lambda^\star,n}$ n'est pas un estimateur. On veut sélectionner le meilleur λ à partir de l'étude du risque de $\hat{f}_{\lambda,n}$.

Compromis biais/variance

- L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.
- On pourra accepter des estimateurs biaisés mais de variance petite, tels que le risque quadratique soit contrôlé.

Oracle

Idéalement, le meilleur estimateur pour le risque R, est

$$f_n^\star = \operatorname*{Argmin}_{\hat{f}_n} R(\hat{f}_n, f).$$

- Pbm : $R(\hat{f}_n, f)$ dépend de la densité f inconnue et n'est donc pas calculable ; $f_n^\star$ n'est pas un estimateur, c'est un oracle.
- Souvent, on dispose d'une famille d'estimateurs $\hat{f}_{\lambda,n}$ dépendants d'un paramètre λ (ex :partition). Alors

$$f_n^\star = \hat{f}_{\lambda^\star,n} \quad \text{avec} \quad \lambda^\star = \operatorname*{Argmin}_{\lambda} R(\hat{f}_{\lambda,n}, f).$$

mais $\hat{f}_{\lambda^\star,n}$ n'est pas un estimateur. On veut sélectionner le meilleur λ à partir de l'étude du risque de $\hat{f}_{\lambda,n}$.

Compromis biais/variance

- L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.
- On pourra accepter des estimateurs biaisés mais de variance petite, tels que le risque quadratique soit contrôlé.

Oracle

Idéalement, le meilleur estimateur pour le risque R, est

$$f_n^\star = \operatorname*{Argmin}_{\hat{f}_n} R(\hat{f}_n, f).$$

- Pbm : $R(\hat{f}_n, f)$ dépend de la densité f inconnue et n'est donc pas calculable ; $f_n^\star$ n'est pas un estimateur, c'est un oracle.
- Souvent, on dispose d'une famille d'estimateurs $\hat{f}_{\lambda,n}$ dépendants d'un paramètre λ (ex :partition). Alors

$$f_n^\star = \hat{f}_{\lambda^\star,n} \quad \text{avec} \quad \lambda^\star = \operatorname*{Argmin}_{\lambda} R(\hat{f}_{\lambda,n}, f).$$

mais $\hat{f}_{\lambda^\star,n}$ n'est pas un estimateur. On veut sélectionner le meilleur λ à partir de l'étude du risque de $\hat{f}_{\lambda,n}$.

Compromis biais/variance

- L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.
- On pourra accepter des estimateurs biaisés mais de variance petite, tels que le risque quadratique soit contrôlé.

Oracle

Idéalement, le meilleur estimateur pour le risque R, est

$$f_n^\star = \operatorname*{Argmin}_{\hat{f}_n} R(\hat{f}_n, f).$$

- Pbm : $R(\hat{f}_n, f)$ dépend de la densité f inconnue et n'est donc pas calculable ; $f_n^\star$ n'est pas un estimateur, c'est un oracle.
- Souvent, on dispose d'une famille d'estimateurs $\hat{f}_{\lambda,n}$ dépendants d'un paramètre λ (ex :partition). Alors

$$f_n^\star = \hat{f}_{\lambda^\star,n} \quad \text{avec} \quad \lambda^\star = \operatorname*{Argmin}_{\lambda} R(\hat{f}_{\lambda,n}, f).$$

mais $\hat{f}_{\lambda^\star,n}$ n'est pas un estimateur. On veut sélectionner le meilleur λ à partir de l'étude du risque de $\hat{f}_{\lambda,n}$.

Compromis biais/variance

- L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.
- On pourra accepter des estimateurs biaisés mais de variance petite, tels que le risque quadratique soit contrôlé.

Oracle

Idéalement, le meilleur estimateur pour le risque R, est

$$f_n^\star = \operatorname*{Argmin}_{\hat{f}_n} R(\hat{f}_n, f).$$

- Pbm : $R(\hat{f}_n, f)$ dépend de la densité f inconnue et n'est donc pas calculable ; $f_n^\star$ n'est pas un estimateur, c'est un oracle.
- Souvent, on dispose d'une famille d'estimateurs $\hat{f}_{\lambda,n}$ dépendants d'un paramètre λ (ex :partition). Alors

$$f_n^\star = \hat{f}_{\lambda^\star,n} \quad \text{avec} \quad \lambda^\star = \operatorname*{Argmin}_{\lambda} R(\hat{f}_{\lambda,n}, f).$$

mais $\hat{f}_{\lambda^\star,n}$ n'est pas un estimateur. On veut sélectionner le meilleur λ à partir de l'étude du risque de $\hat{f}_{\lambda,n}$

Compromis biais/variance

- L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.
- On pourra accepter des estimateurs biaisés mais de variance petite, tels que le risque quadratique soit contrôlé.

Oracle

Idéalement, le meilleur estimateur pour le risque R, est

$$f_n^\star = \operatorname*{Argmin}_{\hat{f}_n} R(\hat{f}_n, f).$$

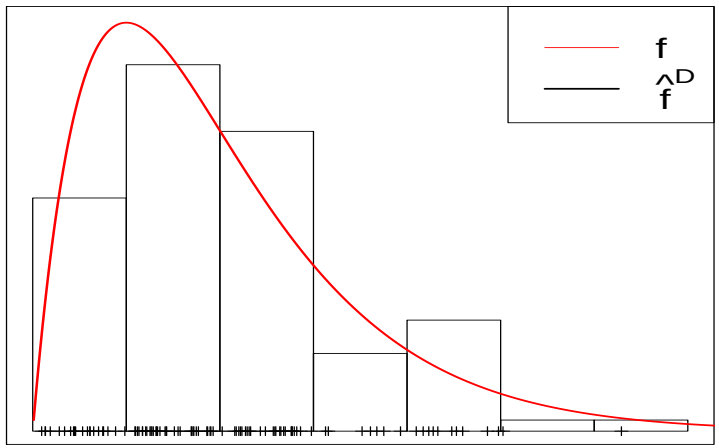
- Pbm : $R(\hat{f}_n, f)$ dépend de la densité f inconnue et n'est donc pas calculable ; $f_n^\star$ n'est pas un estimateur, c'est un oracle.
- Souvent, on dispose d'une famille d'estimateurs $\hat{f}_{\lambda,n}$ dépendants d'un paramètre λ (ex :partition). Alors

$$f_n^\star = \hat{f}_{\lambda^\star,n} \quad \text{avec} \quad \lambda^\star = \operatorname*{Argmin}_{\lambda} R(\hat{f}_{\lambda,n}, f).$$

mais $\hat{f}_{\lambda^\star,n}$ n'est pas un estimateur. On veut sélectionner le meilleur λ à partir de l'étude du risque de $\hat{f}_{\lambda,n}$

Estimateurs par histogramme

- $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$
- Estimateur par histogramme de pas $1/D$:



Pour tout $j = 1, \ldots, D$, soit

$$\hat{d}_j = \frac{1}{n * \text{longueur}(I_j)} \#\{i, X_i \in I_j\}$$

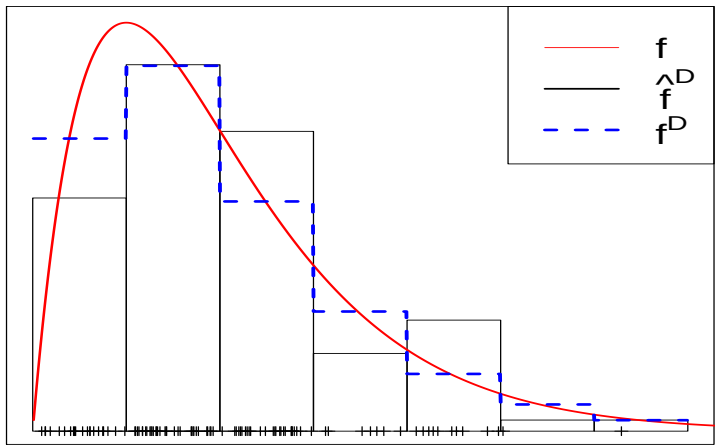
où $I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$, alors $\hat{f}^D = \sum_{j=1}^{D} \hat{d}_j 1_{I_j}$.

- $\hat{d}_j$ estimateur sans biais de $d_j = \mathbb{P}[X \in I_j]/\text{longueur}(I_j)$.
- $f^D = \sum_{j=1}^{D} d_j 1_{I_j}$ est l'histogramme régulier de pas $1/D$ le plus proche de f pour la distance L^2 :

$\Rightarrow$ f^D = projecteur de f sur l'espace des histogrammes de pas $1/D$: vect $\left(1_{I_j}, j = 1, \ldots, D\right)$.

Estimateurs par histogramme

- $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$
- Estimateur par histogramme de pas $1/D$:



Pour tout $j = 1, \ldots, D$, soit

$$\hat{d}_j = \frac{1}{n * \text{longueur}(I_j)} \#\{i, X_i \in I_j\}$$

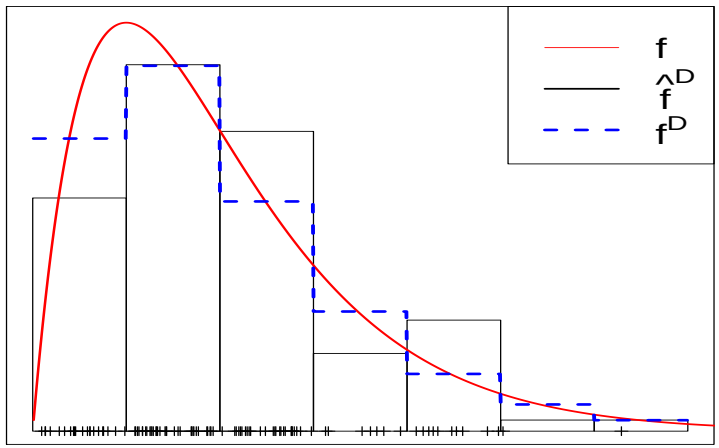
où $I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$, alors $\hat{f}^D = \sum_{j=1}^{D} \hat{d}_j 1_{I_j}$.

- $\hat{d}_j$ estimateur sans biais de $d_j = \mathbb{P}[X \in I_j]/\text{longueur}(I_j)$.
- $f^D = \sum_{j=1}^{D} d_j 1_{I_j}$ est l'histogramme régulier de pas $1/D$ le plus proche de f pour la distance L^2 :

$\Rightarrow$ f^D = projecteur de f sur l'espace des histogrammes de pas $1/D$: vect $\left(1_{I_j}, j = 1, \ldots, D\right)$.

Estimateurs par histogramme

- $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$
- Estimateur par histogramme de pas $1/D$:



Pour tout $j = 1, \ldots, D$, soit

$$\hat{d}_j = \frac{1}{n * \text{longueur}(I_j)} \#\{i, X_i \in I_j\}$$

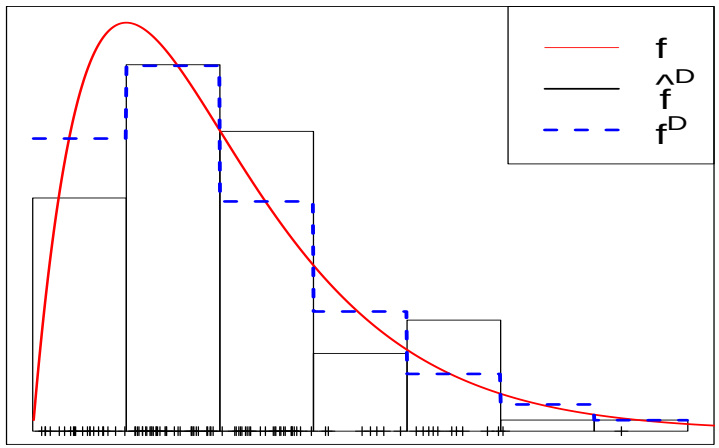
où $I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$, alors $\hat{f}^D = \sum_{j=1}^{D} \hat{d}_j 1_{I_j}$.

- $\hat{d}_j$ estimateur sans biais de $d_j = \mathbb{P}[X \in I_j]/\text{longueur}(I_j)$.
- $f^D = \sum_{j=1}^{D} d_j 1_{I_j}$ est l'histogramme régulier de pas $1/D$ le plus proche de f pour la distance L^2 :

$\Rightarrow$ f^D = projecteur de f sur l'espace des histogrammes de pas $1/D$: vect $\left(1_{I_j}, j = 1, \ldots, D\right)$.

Estimateurs par histogramme

- $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$
- Estimateur par histogramme de pas $1/D$:



Pour tout $j = 1, \ldots, D$, soit

$$\hat{d}_j = \frac{1}{n * \text{longueur}(I_j)} \#\{i, X_i \in I_j\}$$

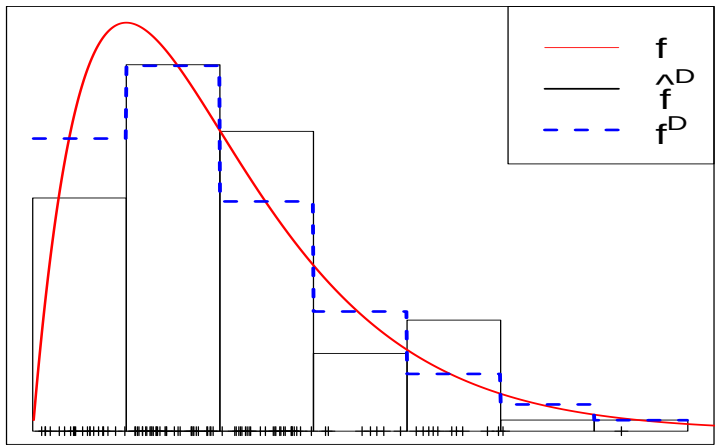
où $I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$, alors $\hat{f}^D = \sum_{j=1}^{D} \hat{d}_j 1_{I_j}$.

- $\hat{d}_j$ estimateur sans biais de $d_j = \mathbb{P}[X \in I_j]/\text{longueur}(I_j)$.
- $f^D = \sum_{j=1}^{D} d_j 1_{I_j}$ est l'histogramme régulier de pas $1/D$ le plus proche de f pour la distance L^2 :

$\Rightarrow$ f^D = projecteur de f sur l'espace des histogrammes de pas $1/D$: vect $\left(1_{I_j}, j = 1, \ldots, D\right)$.

Estimateurs par histogramme

- $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$
- Estimateur par histogramme de pas $1/D$:



Pour tout $j = 1, \ldots, D$, soit

$$\hat{d}_j = \frac{1}{n * \text{longueur}(I_j)} \#\{i, X_i \in I_j\}$$

où $I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$, alors $\hat{f}^D = \sum_{j=1}^{D} \hat{d}_j 1_{I_j}$.

- $\hat{d}_j$ estimateur sans biais de $d_j = \mathbb{P}[X \in I_j]/\text{longueur}(I_j)$.
- $f^D = \sum_{j=1}^{D} d_j 1_{I_j}$ est l'histogramme régulier de pas $1/D$ le plus proche de f pour la distance L^2 :

$\Rightarrow$ f^D = projecteur de f sur l'espace des histogrammes de pas $1/D$: vect $\left(1_{I_j}, j = 1, \ldots, D\right)$.

Estimateurs par projection : principe

- Soit $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$.
- Soit $\{\phi_j\}_{j=1,\ldots,D}$ une famille de fonctions de $L^2([0,1])$ orthonormée pour le p.s. L^2 :

$$\langle g, h \rangle = \int_0^1 g(x)h(x)dx \quad \text{et} \quad \|g\|^2 = \langle g, g \rangle$$

- Soit f_D le projecteur de f sur le modèle $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$.

$$f_D(x) = \sum_{j=1}^{D} \langle f, \phi_j \rangle \phi_j(x)$$

- On calcule un estimateur sans biais de f_D : $\hat{f}_D$.
- MISE de $\hat{f}_D$:

$$\begin{aligned} \mathbb{E}[\|\hat{f}_D - f\|^2] &= \|\mathbb{E}(\hat{f}_D) - f\|^2 + \mathbb{E}[\|\hat{f}_D - \mathbb{E}(\hat{f}_D)\|^2] \\ &= \underbrace{\|f_D - f\|^2}_{\text{biais}} + \underbrace{\mathbb{E}[\|\hat{f}_D - f_D\|^2]}_{\text{variance}} \end{aligned}$$

Estimateurs par projection : principe

- Soit $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$.
- Soit $\{\phi_j\}_{j=1,\ldots,D}$ une famille de fonctions de $L^2([0,1])$ orthonormée pour le p.s. L^2 :

$$\langle g, h \rangle = \int_0^1 g(x)h(x)dx \quad \text{et} \quad \|g\|^2 = \langle g, g \rangle$$

- Soit f_D le projecteur de f sur le modèle $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$.

$$f_D(x) = \sum_{j=1}^{D} \langle f, \phi_j \rangle \phi_j(x)$$

- On calcule un estimateur sans biais de f_D : $\hat{f}_D$.
- MISE de $\hat{f}_D$:

$$\begin{aligned} \mathbb{E}[\|\hat{f}_D - f\|^2] &= \|\mathbb{E}(\hat{f}_D) - f\|^2 + \mathbb{E}[\|\hat{f}_D - \mathbb{E}(\hat{f}_D)\|^2] \\ &= \underbrace{\|f_D - f\|^2}_{\text{biais}} + \underbrace{\mathbb{E}[\|\hat{f}_D - f_D\|^2]}_{\text{variance}} \end{aligned}$$

Estimateurs par projection : principe

- Soit $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$.
- Soit $\{\phi_j\}_{j=1,\ldots,D}$ une famille de fonctions de $L^2([0,1])$ orthonormée pour le p.s. L^2 :

$$\langle g, h \rangle = \int_0^1 g(x)h(x)dx \quad \text{et} \quad \|g\|^2 = \langle g, g \rangle$$

- Soit f_D le projecteur de f sur le modèle $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$.

$$f_D(x) = \sum_{j=1}^{D} \langle f, \phi_j \rangle \phi_j(x)$$

- On calcule un estimateur sans biais de f_D : $\hat{f}_D$.
- MISE de $\hat{f}_D$:

$$\begin{aligned} \mathbb{E}[\|\hat{f}_D - f\|^2] &= \|\mathbb{E}(\hat{f}_D) - f\|^2 + \mathbb{E}[\|\hat{f}_D - \mathbb{E}(\hat{f}_D)\|^2] \\ &= \underbrace{\|f_D - f\|^2}_{\text{biais}} + \underbrace{\mathbb{E}[\|\hat{f}_D - f_D\|^2]}_{\text{variance}} \end{aligned}$$

Estimateurs par projection : principe

- Soit $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$.
- Soit $\{\phi_j\}_{j=1,\ldots,D}$ une famille de fonctions de $L^2([0,1])$ orthonormée pour le p.s. L^2 :

$$\langle g, h \rangle = \int_0^1 g(x)h(x)dx \quad \text{et} \quad \|g\|^2 = \langle g, g \rangle$$

- Soit f_D le projecteur de f sur le modèle $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$.

$$f_D(x) = \sum_{j=1}^{D} \langle f, \phi_j \rangle \phi_j(x)$$

- On calcule un estimateur sans biais de f_D : $\hat{f}_D$.
- MISE de $\hat{f}_D$:

$$\begin{aligned} \mathbb{E}[\|\hat{f}_D - f\|^2] &= \|\mathbb{E}(\hat{f}_D) - f\|^2 + \mathbb{E}[\|\hat{f}_D - \mathbb{E}(\hat{f}_D)\|^2] \\ &= \underbrace{\|f_D - f\|^2}_{\text{biais}} + \underbrace{\mathbb{E}[\|\hat{f}_D - f_D\|^2]}_{\text{variance}} \end{aligned}$$

Estimateurs par projection : principe

- Soit $X_1, \ldots, X_n$ i.i.d de densité $f \in L^2([0,1])$.
- Soit $\{\phi_j\}_{j=1,\ldots,D}$ une famille de fonctions de $L^2([0,1])$ orthonormée pour le p.s. L^2 :

$$\langle g, h \rangle = \int_0^1 g(x)h(x)dx \quad \text{et} \quad \|g\|^2 = \langle g, g \rangle$$

- Soit f_D le projecteur de f sur le modèle $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$.

$$f_D(x) = \sum_{j=1}^{D} \langle f, \phi_j \rangle \phi_j(x)$$

- On calcule un estimateur sans biais de f_D : $\hat{f}_D$.
- MISE de $\hat{f}_D$:

$$\begin{aligned} \mathbb{E}[\|\hat{f}_D - f\|^2] &= \|\mathbb{E}(\hat{f}_D) - f\|^2 + \mathbb{E}[\|\hat{f}_D - \mathbb{E}(\hat{f}_D)\|^2] \\ &= \underbrace{\|f_D - f\|^2}_{\text{biais}} + \underbrace{\mathbb{E}[\|\hat{f}_D - f_D\|^2]}_{\text{variance}} \end{aligned}$$

Exemples de bases d'approximation

- Histogrammes réguliers de pas $1/D$. Pour tout $j = 1, \ldots, D$,

$$\phi_j(x) = \sqrt{D} 1_{[j-1/D, j/D[}(x)$$

alors $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$ est l'ensemble des histogrammes réguliers de pas $1/D$.

- Base de fonctions linéaires par morceaux de pas $1/d$: $S_D = \text{vect}\{\phi_{j,0}, \phi_{j,1}, j = 1, \ldots, d\}$ avec

$$\begin{cases} \phi_{j,0}(x) = 1_{[j-1/d, j/d[}(x) \\ \phi_{j,1}(x) = (x - a_j) 1_{[j-1/d, j/d[(}(x) \quad \text{avec} \quad a_j = (2j-1)/d \end{cases}$$

- Splines de pas $1/d$ et de degré 1 : base linéaire par morceaux modifiée pour contraindre l'estimateur à être continu.

- Base trigonométrique : $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$ avec

$$\begin{cases} \phi_1(x) = 1 \\ \phi_{2j+1}(x) = \sqrt{2} \sin(2\pi j x) \\ \phi_{2j}(x) = \sqrt{2} \cos(2\pi j x) \end{cases}$$

Exemples de bases d'approximation

- Histogrammes réguliers de pas $1/D$. Pour tout $j = 1, \dots, D$,

$$\phi_j(x) = \sqrt{D} 1_{[j-1/D, j/D[}(x)$$

alors $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ est l'ensemble des histogrammes réguliers de pas $1/D$.

- Base de fonctions linéaires par morceaux de pas $1/d$: $S_D = \text{vect}\{\phi_{j,0}, \phi_{j,1}, j = 1, \dots, d\}$ avec

$$\begin{cases} \phi_{j,0}(x) = 1_{[j-1/d, j/d[}(x) \\ \phi_{j,1}(x) = (x - a_j) 1_{[j-1/d, j/d[(}(x) \quad \text{avec} \quad a_j = (2j-1)/d \end{cases}$$

- Splines de pas $1/d$ et de degré 1 : base linéaire par morceaux modifiée pour contraindre l'estimateur à être continu.

- Base trigonométrique : $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ avec

$$\begin{cases} \phi_1(x) = 1 \\ \phi_{2j+1}(x) = \sqrt{2} \sin(2\pi j x) \\ \phi_{2j}(x) = \sqrt{2} \cos(2\pi j x) \end{cases}$$

Exemples de bases d'approximation

- Histogrammes réguliers de pas $1/D$. Pour tout $j = 1, \dots, D$,

$$\phi_j(x) = \sqrt{D} 1_{[j-1/D, j/D[}(x)$$

alors $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ est l'ensemble des histogrammes réguliers de pas $1/D$.

- Base de fonctions linéaires par morceaux de pas $1/d$: $S_D = \text{vect}\{\phi_{j,0}, \phi_{j,1}, j = 1, \dots, d\}$ avec

$$\begin{cases} \phi_{j,0}(x) = 1_{[j-1/d, j/d[}(x) \\ \phi_{j,1}(x) = (x - a_j) 1_{[j-1/d, j/d[(}(x) \quad \text{avec} \quad a_j = (2j-1)/d \end{cases}$$

- Splines de pas $1/d$ et de degré 1 : base linéaire par morceaux modifiée pour contraindre l'estimateur à être continu.

- Base trigonométrique : $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ avec

$$\begin{cases} \phi_1(x) = 1 \\ \phi_{2j+1}(x) = \sqrt{2} \sin(2\pi j x) \\ \phi_{2j}(x) = \sqrt{2} \cos(2\pi j x) \end{cases}$$

Exemples de bases d'approximation

- Histogrammes réguliers de pas $1/D$. Pour tout $j = 1, \dots, D$,

$$\phi_j(x) = \sqrt{D} 1_{[j-1/D, j/D[}(x)$$

alors $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ est l'ensemble des histogrammes réguliers de pas $1/D$.

- Base de fonctions linéaires par morceaux de pas $1/d$: $S_D = \text{vect}\{\phi_{j,0}, \phi_{j,1}, j = 1, \dots, d\}$ avec

$$\begin{cases} \phi_{j,0}(x) = 1_{[j-1/d, j/d[}(x) \\ \phi_{j,1}(x) = (x - a_j) 1_{[j-1/d, j/d[(}(x) \quad \text{avec} \quad a_j = (2j-1)/d \end{cases}$$

- Splines de pas $1/d$ et de degré 1 : base linéaire par morceaux modifiée pour contraindre l'estimateur à être continu.

- Base trigonométrique : $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ avec

$$\begin{cases} \phi_1(x) = 1 \\ \phi_{2j+1}(x) = \sqrt{2} \sin(2\pi j x) \\ \phi_{2j}(x) = \sqrt{2} \cos(2\pi j x) \end{cases}$$

Exemples de bases d'approximation

- Histogrammes réguliers de pas $1/D$. Pour tout $j = 1, \dots, D$,

$$\phi_j(x) = \sqrt{D} 1_{[j-1/D, j/D[}(x)$$

alors $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ est l'ensemble des histogrammes réguliers de pas $1/D$.

- Base de fonctions linéaires par morceaux de pas $1/d$: $S_D = \text{vect}\{\phi_{j,0}, \phi_{j,1}, j = 1, \dots, d\}$ avec

$$\begin{cases} \phi_{j,0}(x) = 1_{[j-1/d, j/d[}(x) \\ \phi_{j,1}(x) = (x - a_j) 1_{[j-1/d, j/d[(}(x) \quad \text{avec} \quad a_j = (2j-1)/d \end{cases}$$

- Splines de pas $1/d$ et de degré 1 : base linéaire par morceaux modifiée pour contraindre l'estimateur à être continu.

- Base trigonométrique : $S_D = \text{vect}\{\phi_j, j = 1, \dots, D\}$ avec

$$\begin{cases} \phi_1(x) = 1 \\ \phi_{2j+1}(x) = \sqrt{2} \sin(2\pi j x) \\ \phi_{2j}(x) = \sqrt{2} \cos(2\pi j x) \end{cases}$$

- Propriété : Pour les bases précédentes, on peut montrer par des calculs simples que :

$$\left\| \sum_{j=1}^{D} \phi_j^2 \right\|_\infty \leq KD$$

  oú K dépend de la nature de la base (histogramme, trigonométrique, etc) mais pas de D.

- Exples Pour les histogrammes et la base trigonométrique, $K = 1$.

- Remarque On peut également considérer des bases irrégulières.

Ex : option "`regular`" et "`irregular`" du package `histogram`

Histogramme régulier $D = 7$ Histogramme irrégulier $D = 7$

- Propriété : Pour les bases précédentes, on peut montrer par des calculs simples que :

$$\left\| \sum_{j=1}^{D} \phi_j^2 \right\|_\infty \leq KD$$

  oú K dépend de la nature de la base (histogramme, trigonométrique, etc) mais pas de D.

- Exples Pour les histogrammes et la base trigonométrique, $K = 1$.

- Remarque On peut également considérer des bases irrégulières.

Ex : option "`regular`" et "`irregular`" du package `histogram`

Histogramme régulier $D = 7$ Histogramme irrégulier $D = 7$

- Propriété : Pour les bases précédentes, on peut montrer par des calculs simples que :

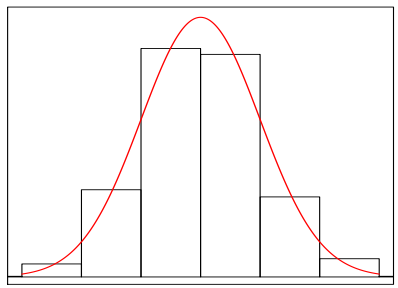
$$\left\| \sum_{j=1}^{D} \phi_j^2 \right\|_\infty \leq KD$$

  où K dépend de la nature de la base (histogramme, trigonométrique, etc) mais pas de D.

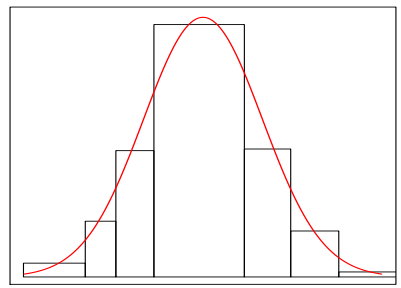
- Exples Pour les histogrammes et la base trigonométrique, $K = 1$.

- Remarque On peut également considérer des bases irrégulières.

Ex : option "`regular`" et "`irregular`" du package `histogram`



Histogramme régulier $D = 7$



Histogramme irrégulier $D = 7$

Construction des estimateurs par projection

- Soit $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$, et soit $f_D = \sum_{j=1}^{D} \theta_j \phi_j$ le projecteur de f sur S_D
- Soit $f = f_D + f_{S_D^\perp}$, alors comme la famille $\{\phi_j\}$ est orthonormée

$$\langle f, \phi_j \rangle = \langle f_D, \phi_j \rangle + \langle f_{S_D^\perp}, \phi_j \rangle = \sum_{k=1}^{D} \theta_k \langle \phi_k, \phi_j \rangle + 0 = \theta_j$$

- D'où :

$$\theta_j = \int \phi_j(x) f(x)\, dx = \mathbb{E}[\phi_j(X_1)] \quad \text{estimé par} \quad \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i)$$

et $\hat{f}_D(x) = \sum_{j=1}^{D} \hat{\theta}_j \phi_j(x)$ est l'estimateur par projection de f sur l'espace S_D.

Construction des estimateurs par projection

- Soit $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$, et soit $f_D = \sum_{j=1}^{D} \theta_j \phi_j$ le projecteur de f sur S_D
- Soit $f = f_D + f_{S_D^\perp}$, alors comme la famille $\{\phi_j\}$ est orthonormée

$$\langle f, \phi_j \rangle = \langle f_D, \phi_j \rangle + \langle f_{S_D^\perp}, \phi_j \rangle = \sum_{k=1}^{D} \theta_k \langle \phi_k, \phi_j \rangle + 0 = \theta_j$$

- D'où :

$$\theta_j = \int \phi_j(x) f(x)\, dx = \mathbb{E}[\phi_j(X_1)] \quad \text{estimé par} \quad \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i)$$

et $\hat{f}_D(x) = \sum_{j=1}^{D} \hat{\theta}_j \phi_j(x)$ est l'estimateur par projection de f sur l'espace S_D.

Construction des estimateurs par projection

- Soit $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$, et soit $f_D = \sum_{j=1}^{D} \theta_j \phi_j$ le projecteur de f sur S_D
- Soit $f = f_D + f_{S_D^\perp}$, alors comme la famille $\{\phi_j\}$ est orthonormée

$$\langle f, \phi_j \rangle = \langle f_D, \phi_j \rangle + \langle f_{S_D^\perp}, \phi_j \rangle = \sum_{k=1}^{D} \theta_k \langle \phi_k, \phi_j \rangle + 0 = \theta_j$$

- D'où :

$$\theta_j = \int \phi_j(x) f(x)\, dx = \mathbb{E}[\phi_j(X_1)] \quad \text{estimé par} \quad \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i)$$

et $\hat{f}_D(x) = \sum_{j=1}^{D} \hat{\theta}_j \phi_j(x)$ est l'estimateur par projection de f sur l'espace S_D.

Construction des estimateurs par projection

- Soit $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$, et soit $f_D = \sum_{j=1}^{D} \theta_j \phi_j$ le projecteur de f sur S_D
- Soit $f = f_D + f_{S_D^\perp}$, alors comme la famille $\{\phi_j\}$ est orthonormée

$$\langle f, \phi_j \rangle = \langle f_D, \phi_j \rangle + \langle f_{S_D^\perp}, \phi_j \rangle = \sum_{k=1}^{D} \theta_k \langle \phi_k, \phi_j \rangle + 0 = \theta_j$$

- D'où :

$$\theta_j = \int \phi_j(x) f(x) dx = \mathbb{E}[\phi_j(X_1)] \quad \text{estimé par} \quad \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i)$$

et $\hat{f}_D(x) = \sum_{j=1}^{D} \hat{\theta}_j \phi_j(x)$ est l'estimateur par projection de f sur l'espace S_D.

Construction des estimateurs par projection

- Soit $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$, et soit $f_D = \sum_{j=1}^{D} \theta_j \phi_j$ le projecteur de f sur S_D
- Soit $f = f_D + f_{S_D^\perp}$, alors comme la famille $\{\phi_j\}$ est orthonormée

$$\langle f, \phi_j \rangle = \langle f_D, \phi_j \rangle + \langle f_{S_D^\perp}, \phi_j \rangle = \sum_{k=1}^{D} \theta_k \langle \phi_k, \phi_j \rangle + 0 = \theta_j$$

- D'où :

$$\theta_j = \int \phi_j(x) f(x) dx = \mathbb{E}[\phi_j(X_1)] \quad \text{estimé par} \quad \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i)$$

et $\hat{f}_D(x) = \sum_{j=1}^{D} \hat{\theta}_j \phi_j(x)$ est l'estimateur par projection de f sur l'espace S_D.

Variance du MISE

La base (ϕ_j) est orthonormée pour le p.s. L^2, donc :

$$\|\hat{f}_D - f_D\|^2 = \Big\| \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)\phi_j \Big\|^2 = \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)^2$$

$$\begin{aligned}
\text{Var(MISE)} = \mathbb{E}[\|\hat{f}_D - f_D\|^2] &= \sum_{j=1}^{D} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] = \sum_{j=1}^{D} \text{Var}(\hat{\theta}_j) \\
&= \sum_{j=1}^{D} \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\right] = \frac{1}{n}\sum_{j=1}^{D} \text{Var}[\phi_j(X_1)] \\
&\leq \frac{1}{n}\sum_{j=1}^{D} \mathbb{E}\left[\phi_j^2(X_1)\right]
\end{aligned}$$

Or $\| \sum_{j=1}^{D} \phi_j(.)^2 \|_\infty \leq K$ avec K indépendant de D, donc,

$$\mathbb{E}[\|\hat{f}_D - f_D\|^2] \leq K\frac{D}{n}$$

Variance du MISE

La base (ϕ_j) est orthonormée pour le p.s. L^2, donc :

$$\|\hat{f}_D - f_D\|^2 = \left\| \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)\phi_j \right\|^2 = \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)^2$$

$$\begin{aligned}
\text{Var(MISE)} = \mathbb{E}[\|\hat{f}_D - f_D\|^2] &= \sum_{j=1}^{D} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] = \sum_{j=1}^{D} \text{Var}(\hat{\theta}_j) \\
&= \sum_{j=1}^{D} \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\right] = \frac{1}{n}\sum_{j=1}^{D} \text{Var}[\phi_j(X_1)] \\
&\leq \frac{1}{n}\sum_{j=1}^{D} \mathbb{E}\left[\phi_j^2(X_1)\right]
\end{aligned}$$

Or $\| \sum_{j=1}^{D} \phi_j(.)^2 \|_\infty \leq K$ avec K indépendant de D, donc,

$$\mathbb{E}[\|\hat{f}_D - f_D\|^2] \leq K\frac{D}{n}$$

Variance du MISE

La base (ϕ_j) est orthonormée pour le p.s. L^2, donc :

$$\|\hat{f}_D - f_D\|^2 = \left\| \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)\phi_j \right\|^2 = \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)^2$$

$$\begin{aligned}
\text{Var(MISE)} = \mathbb{E}[\|\hat{f}_D - f_D\|^2] &= \sum_{j=1}^{D} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] = \sum_{j=1}^{D} \text{Var}(\hat{\theta}_j) \\
&= \sum_{j=1}^{D} \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\right] = \frac{1}{n}\sum_{j=1}^{D} \text{Var}[\phi_j(X_1)] \\
&\leq \frac{1}{n}\sum_{j=1}^{D} \mathbb{E}\left[\phi_j^2(X_1)\right]
\end{aligned}$$

Or $\| \sum_{j=1}^{D} \phi_j(.)^2 \|_\infty \leq K$ avec K indépendant de D, donc,

$$\mathbb{E}[\|\hat{f}_D - f_D\|^2] \leq K \frac{D}{n}$$

Variance du MISE

La base (ϕ_j) est orthonormée pour le p.s. L^2, donc :

$$\|\hat{f}_D - f_D\|^2 = \Big\| \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)\phi_j \Big\|^2 = \sum_{j=1}^{D} (\hat{\theta}_j - \theta_j)^2$$

$$\begin{aligned}
\text{Var(MISE)} = \mathbb{E}[\|\hat{f}_D - f_D\|^2] &= \sum_{j=1}^{D} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] = \sum_{j=1}^{D} \text{Var}(\hat{\theta}_j) \\
&= \sum_{j=1}^{D} \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\right] = \frac{1}{n}\sum_{j=1}^{D} \text{Var}[\phi_j(X_1)] \\
&\leq \frac{1}{n}\sum_{j=1}^{D} \mathbb{E}\left[\phi_j^2(X_1)\right]
\end{aligned}$$

Or $\| \sum_{j=1}^{D} \phi_j(.)^2 \|_\infty \leq K$ avec K indépendant de D, donc,

$$\mathbb{E}[\|\hat{f}_D - f_D\|^2] \leq K\frac{D}{n} \quad \Rightarrow \quad \text{la variance augmente avec } D$$

Biais du MISE : heuristique

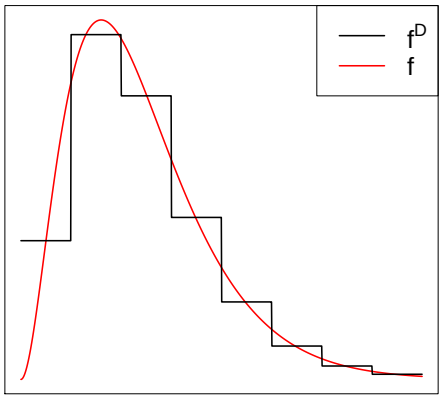
- Le biais diminue quand la régularité de f augmente
- Exples : histogramme régulier et base trigonométrique ($D = 8$).

f plus régulière f moins régulière

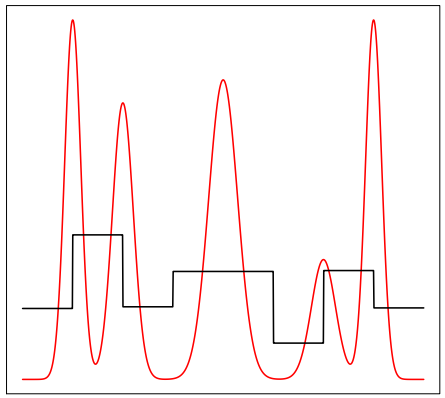
Biais du MISE : heuristique

- Le biais diminue quand la régularité de f augmente
- Exples : histogramme régulier et base trigonométrique ($D = 8$).

f plus régulière



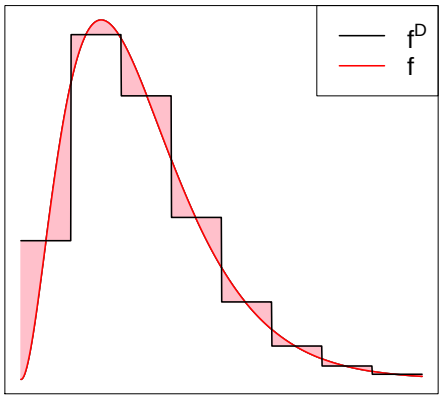
f moins régulière



Biais du MISE : heuristique

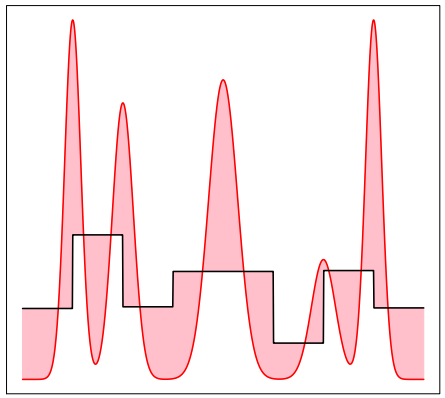
- Le biais diminue quand la régularité de f augmente
- Exples : histogramme régulier et base trigonométrique ($D = 8$).

f plus régulière



biais=0.09

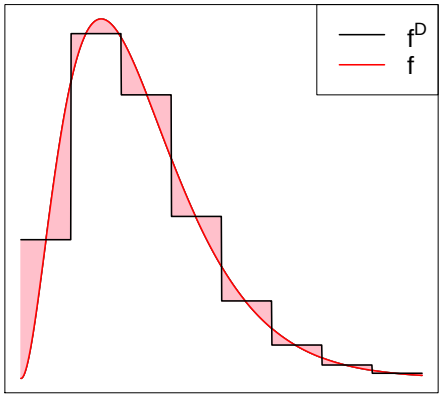
f moins régulière


biais=1.24

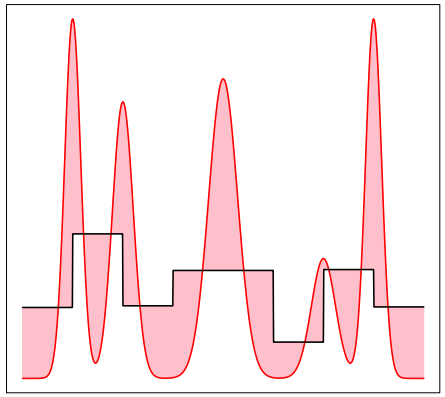
Biais du MISE : heuristique

- Le biais diminue quand la régularité de f augmente
- Exples : histogramme régulier et base trigonométrique ($D = 8$).

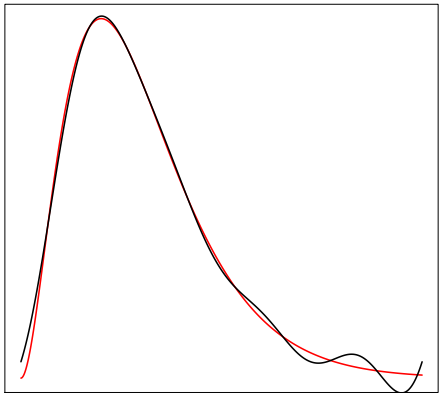
f plus régulière f moins régulière

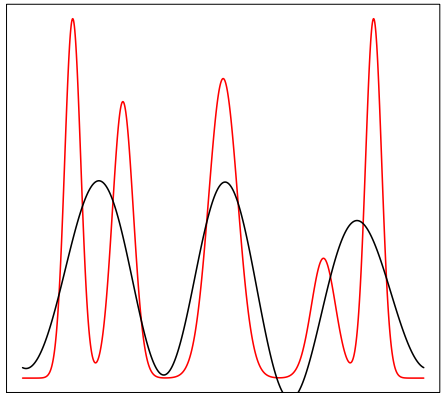


biais=0.09



biais=1.24

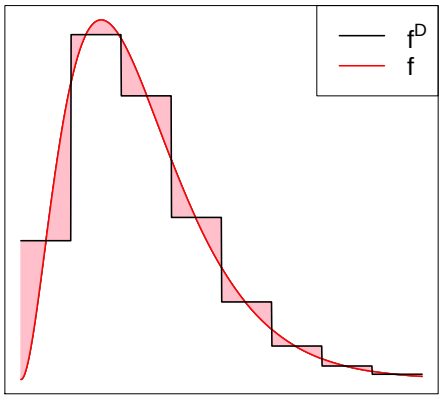




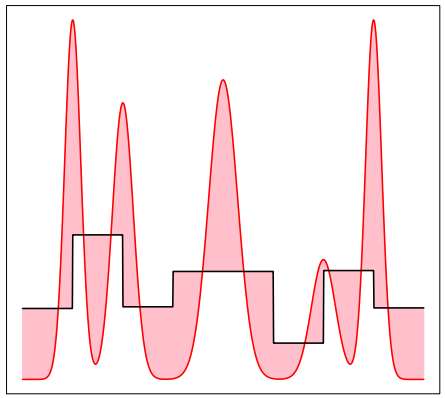
Biais du MISE : heuristique

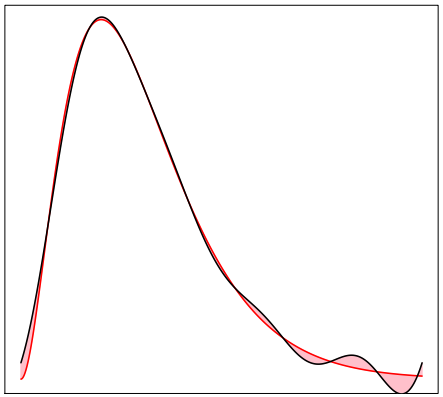
- Le biais diminue quand la régularité de f augmente
- Exples : histogramme régulier et base trigonométrique ($D = 8$).

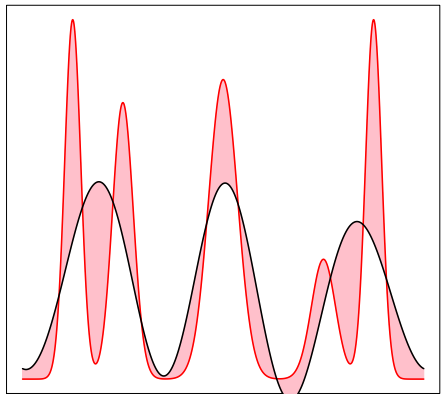
f plus régulière f moins régulière



biais=0.09


biais=1.24


biais=0.003


biais=0.8

Espaces de régularités

Espaces de fonctions k-dérivables

(i) **Régularité globale** : $\mathcal{A}_1(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_L^2 \leq L\}$

D'après l'égalité de Parseval, soit f^* la transformée de Fourier de f,

$$f \in \mathcal{A}_1(k, L) \quad \Leftrightarrow \quad \|f^{(k)}\| = \int (f^*(\lambda))^2 \lambda^{2k} d\lambda \leq L^2$$

(ii) **Régularité locale** : $\mathcal{A}_2(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_\infty \leq L\}$

Généralisation à une régularité $\alpha > 0$

(i) Espaces de Sobolev : soit $\alpha > 0$,

$$\mathcal{S}(\alpha, L) = \{f, f \in \mathcal{C}^r, \int (f^*(\lambda))^2 \lambda^{2\alpha} d\lambda \leq L^2\}$$

(ii) Espaces de Holder : soit $\alpha > 0$ et r le plus petit entier strictement inférieur à α

$$\mathcal{H}(\alpha, L) = \{f, f \in \mathcal{C}^r, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\alpha - r}, \forall x, y\}$$

Espaces de régularités

Espaces de fonctions k-dérivables

(i) **Régularité globale** : $\mathcal{A}_1(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_L^2 \leq L\}$

D'après l'égalité de Parseval, soit f^* la transformée de Fourier de f,

$$f \in \mathcal{A}_1(k, L) \quad \Leftrightarrow \quad \|f^{(k)}\| = \int (f^*(\lambda))^2 \lambda^{2k} d\lambda \leq L^2$$

(ii) **Régularité locale** : $\mathcal{A}_2(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_\infty \leq L\}$

Généralisation à une régularité $\alpha > 0$

(i) Espaces de Sobolev : soit $\alpha > 0$,

$$\mathcal{S}(\alpha, L) = \{f, f \in \mathcal{C}^r, \int (f^*(\lambda))^2 \lambda^{2\alpha} d\lambda \leq L^2\}$$

(ii) Espaces de Holder : soit $\alpha > 0$ et r le plus petit entier strictement inférieur à α

$$\mathcal{H}(\alpha, L) = \{f, f \in \mathcal{C}^r, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\alpha - r}, \forall x, y\}$$

Espaces de régularités

Espaces de fonctions k-dérivables

(i) **Régularité globale** : $\mathcal{A}_1(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_L^2 \leq L\}$

D'après l'égalité de Parseval, soit f^* la transformée de Fourier de f,

$$f \in \mathcal{A}_1(k, L) \quad \Leftrightarrow \quad \|f^{(k)}\| = \int (f^*(\lambda))^2 \lambda^{2k} d\lambda \leq L^2$$

(ii) **Régularité locale** : $\mathcal{A}_2(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_\infty \leq L\}$

Généralisation à une régularité $\alpha > 0$

(i) Espaces de Sobolev : soit $\alpha > 0$,

$$\mathcal{S}(\alpha, L) = \{f, f \in \mathcal{C}^r, \int (f^*(\lambda))^2 \lambda^{2\alpha} d\lambda \leq L^2\}$$

(ii) Espaces de Holder : soit $\alpha > 0$ et r le plus petit entier strictement inférieur à α

$$\mathcal{H}(\alpha, L) = \{f, f \in \mathcal{C}^r, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\alpha - r}, \forall x, y\}$$

Espaces de régularités

Espaces de fonctions k-dérivables

(i) **Régularité globale** : $\mathcal{A}_1(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_L^2 \leq L\}$

D'après l'égalité de Parseval, soit f^* la transformée de Fourier de f,

$$f \in \mathcal{A}_1(k, L) \quad \Leftrightarrow \quad \|f^{(k)}\| = \int (f^*(\lambda))^2 \lambda^{2k} d\lambda \leq L^2$$

(ii) **Régularité locale** : $\mathcal{A}_2(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_\infty \leq L\}$

Généralisation à une régularité $\alpha > 0$

(i) Espaces de Sobolev : soit $\alpha > 0$,

$$\mathcal{S}(\alpha, L) = \{f, f \in \mathcal{C}^r, \int (f^*(\lambda))^2 \lambda^{2\alpha} d\lambda \leq L^2\}$$

(ii) Espaces de Holder : soit $\alpha > 0$ et r le plus petit entier strictement inférieur à α

$$\mathcal{H}(\alpha, L) = \{f, f \in \mathcal{C}^r, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\alpha - r}, \forall x, y\}$$

Espaces de régularités

Espaces de fonctions k-dérivables

(i) **Régularité globale** : $\mathcal{A}_1(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_L^2 \leq L\}$

D'après l'égalité de Parseval, soit f^* la transformée de Fourier de f,

$$f \in \mathcal{A}_1(k, L) \quad \Leftrightarrow \quad \|f^{(k)}\| = \int (f^*(\lambda))^2 \lambda^{2k} d\lambda \leq L^2$$

(ii) **Régularité locale** : $\mathcal{A}_2(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_\infty \leq L\}$

Généralisation à une régularité $\alpha > 0$

(i) Espaces de Sobolev : soit $\alpha > 0$,

$$\mathcal{S}(\alpha, L) = \{f, f \in \mathcal{C}^r, \int (f^*(\lambda))^2 \lambda^{2\alpha} d\lambda \leq L^2\}$$

(ii) Espaces de Holder : soit $\alpha > 0$ et r le plus petit entier strictement inférieur à α

$$\mathcal{H}(\alpha, L) = \{f, f \in \mathcal{C}^r, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\alpha - r}, \forall x, y\}$$

Espaces de régularités

Espaces de fonctions k-dérivables

(i) **Régularité globale** : $\mathcal{A}_1(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_L^2 \leq L\}$

D'après l'égalité de Parseval, soit f^* la transformée de Fourier de f,

$$f \in \mathcal{A}_1(k, L) \quad \Leftrightarrow \quad \|f^{(k)}\| = \int (f^*(\lambda))^2 \lambda^{2k} d\lambda \leq L^2$$

(ii) **Régularité locale** : $\mathcal{A}_2(k, L) = \{f \in \mathcal{C}^k, \|f^{(k)}\|_\infty \leq L\}$

Généralisation à une régularité $\alpha > 0$

(i) Espaces de Sobolev : soit $\alpha > 0$,

$$\mathcal{S}(\alpha, L) = \{f, f \in \mathcal{C}^r, \int (f^*(\lambda))^2 \lambda^{2\alpha} d\lambda \leq L^2\}$$

(ii) Espaces de Holder : soit $\alpha > 0$ et r le plus petit entier strictement inférieur à α

$$\mathcal{H}(\alpha, L) = \{f, f \in \mathcal{C}^r, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\alpha - r}, \forall x, y\}$$

Résultats

- Par des calculs simples, on montre les résultats suivants :
  - Soit $f \in \mathcal{H}(\alpha, L)$ avec $\alpha \in]0, 1]$, et f_D sa projection sur la base d'histogramme régulier de pas $1/D$, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

  - Soit $f \in \mathcal{S}(\alpha, L)$, et f_D sa projection sur la base trigonométrique de dimension D, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

- Plus généralement, pour les bases d'approximation "classiques" et les espaces de régularité correspondants, on a :

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

où f_D est le projecteur de f sur l'espace d'approximation de dimension D et α la régularité de f.

$\hookrightarrow$ Le biais diminue quand D augmente

Résultats

- Par des calculs simples, on montre les résultats suivants :
  - Soit $f \in \mathcal{H}(\alpha, L)$ avec $\alpha \in]0, 1]$, et f_D sa projection sur la base d'histogramme régulier de pas $1/D$, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

  - Soit $f \in \mathcal{S}(\alpha, L)$, et f_D sa projection sur la base trigonométrique de dimension D, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

- Plus généralement, pour les bases d'approximation "classiques" et les espaces de régularité correspondants, on a :

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

où f_D est le projecteur de f sur l'espace d'approximation de dimension D et α la régularité de f.

$\hookrightarrow$ Le biais diminue quand D augmente

Résultats

- Par des calculs simples, on montre les résultats suivants :
  - Soit $f \in \mathcal{H}(\alpha, L)$ avec $\alpha \in]0, 1]$, et f_D sa projection sur la base d'histogramme régulier de pas $1/D$, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

  - Soit $f \in \mathcal{S}(\alpha, L)$, et f_D sa projection sur la base trigonométrique de dimension D, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

- Plus généralement, pour les bases d'approximation "classiques" et les espaces de régularité correspondants, on a :

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

où f_D est le projecteur de f sur l'espace d'approximation de dimension D et α la régularité de f.

$\hookrightarrow$ Le biais diminue quand D augmente

Résultats

- Par des calculs simples, on montre les résultats suivants :
  - Soit $f \in \mathcal{H}(\alpha, L)$ avec $\alpha \in]0, 1]$, et f_D sa projection sur la base d'histogramme régulier de pas $1/D$, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

  - Soit $f \in \mathcal{S}(\alpha, L)$, et f_D sa projection sur la base trigonométrique de dimension D, alors

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

- Plus généralement, pour les bases d'approximation "classiques" et les espaces de régularité correspondants, on a :

$$\|f - f_D\|^2 \leq L^2 D^{-2\alpha}$$

où f_D est le projecteur de f sur l'espace d'approximation de dimension D et α la régularité de f.

$\hookrightarrow$ Le biais diminue quand D augmente

Décomposition bias-variance : conclusion

- $\mathbb{E}[\|\hat{f}_D - f\|^2] =$ biais + variance avec,

$$\begin{cases} \text{Variance} \leq \frac{KD}{n} & \nearrow \text{ quand } D \nearrow \\ \text{Biais} \leq L^2 D^{-2\alpha} & \searrow \text{ quand } D \nearrow \end{cases}$$

- Dimension de l'espace d'approximation optimal ou oracle :

$$D^{**} = \arg\min \left\{ \frac{KD}{n} + D^{-2\alpha} \right\} \quad \Rightarrow \quad D^{**} = Cn^{1/(2\alpha+1)}$$

$\hookrightarrow D^{**}$ dépend de la régularité inconnue de la fonction

- Vitesse de convergence pour l'oracle :

$$\mathbb{E}[\|\hat{f}_{D^{**}} - f\|^2] \propto n^{-2\alpha/(2\alpha+1)}$$

- **Remarque :** $\frac{KD}{n} + D^{-2\alpha}$ est un majorant du MISE, mais on peut montrer que cette majoration atteint la vitesse optimale.
- Sélection de modèle : critère de sélection automatique de D en fonction des données.

Décomposition bias-variance : conclusion

- $\mathbb{E}[\|\hat{f}_D - f\|^2] =$ biais + variance avec,

$$\begin{cases} \text{Variance} \leq \frac{KD}{n} & \nearrow \text{ quand } D \nearrow \\ \text{Biais} \leq L^2 D^{-2\alpha} & \searrow \text{ quand } D \nearrow \end{cases}$$

- Dimension de l'espace d'approximation optimal ou oracle :

$$D^{**} = \arg\min \left\{ \frac{KD}{n} + D^{-2\alpha} \right\} \quad \Rightarrow \quad D^{**} = Cn^{1/(2\alpha+1)}$$

$\hookrightarrow D^{**}$ dépend de la régularité inconnue de la fonction

- Vitesse de convergence pour l'oracle :

$$\mathbb{E}[\|\hat{f}_{D^{**}} - f\|^2] \propto n^{-2\alpha/(2\alpha+1)}$$

- **Remarque :** $\frac{KD}{n} + D^{-2\alpha}$ est un majorant du MISE, mais on peut montrer que cette majoration atteint la vitesse optimale.
- Sélection de modèle : critère de sélection automatique de D en fonction des données.

Décomposition bias-variance : conclusion

- $\mathbb{E}[\|\hat{f}_D - f\|^2] =$ biais + variance avec,

$$\begin{cases} \text{Variance} \leq \frac{KD}{n} & \nearrow \text{ quand } D \nearrow \\ \text{Biais} \leq L^2 D^{-2\alpha} & \searrow \text{ quand } D \nearrow \end{cases}$$

- Dimension de l'espace d'approximation optimal ou oracle :

$$D^{**} = \arg\min \left\{ \frac{KD}{n} + D^{-2\alpha} \right\} \quad \Rightarrow \quad D^{**} = Cn^{1/(2\alpha+1)}$$

$\hookrightarrow D^{**}$ dépend de la régularité inconnue de la fonction

- Vitesse de convergence pour l'oracle :

$$\mathbb{E}[\|\hat{f}_{D^{**}} - f\|^2] \propto n^{-2\alpha/(2\alpha+1)}$$

- **Remarque :** $\frac{KD}{n} + D^{-2\alpha}$ est un majorant du MISE, mais on peut montrer que cette majoration atteint la vitesse optimale.
- Sélection de modèle : critère de sélection automatique de D en fonction des données.

Décomposition bias-variance : conclusion

- $\mathbb{E}[\|\hat{f}_D - f\|^2] =$ biais + variance avec,

$$\begin{cases} \text{Variance} \leq \frac{KD}{n} & \nearrow \text{ quand } D \nearrow \\ \text{Biais} \leq L^2 D^{-2\alpha} & \searrow \text{ quand } D \nearrow \end{cases}$$

- Dimension de l'espace d'approximation optimal ou oracle :

$$D^{**} = \arg\min \left\{ \frac{KD}{n} + D^{-2\alpha} \right\} \quad \Rightarrow \quad D^{**} = Cn^{1/(2\alpha+1)}$$

$\hookrightarrow D^{**}$ dépend de la régularité inconnue de la fonction

- Vitesse de convergence pour l'oracle :

$$\mathbb{E}[\|\hat{f}_{D^{**}} - f\|^2] \propto n^{-2\alpha/(2\alpha+1)}$$

- **Remarque :** $\frac{KD}{n} + D^{-2\alpha}$ est un majorant du MISE, mais on peut montrer que cette majoration atteint la vitesse optimale.
- Sélection de modèle : critère de sélection automatique de D en fonction des données.

Décomposition bias-variance : conclusion

- $\mathbb{E}[\|\hat{f}_D - f\|^2] =$ biais + variance avec,

$$\begin{cases} \text{Variance} \leq \frac{KD}{n} & \nearrow \text{ quand } D \nearrow \\ \text{Biais} \leq L^2 D^{-2\alpha} & \searrow \text{ quand } D \nearrow \end{cases}$$

- Dimension de l'espace d'approximation optimal ou oracle :

$$D^{**} = \arg\min \left\{ \frac{KD}{n} + D^{-2\alpha} \right\} \quad \Rightarrow \quad D^{**} = Cn^{1/(2\alpha+1)}$$

$\hookrightarrow D^{**}$ dépend de la régularité inconnue de la fonction

- Vitesse de convergence pour l'oracle :

$$\mathbb{E}[\|\hat{f}_{D^{**}} - f\|^2] \propto n^{-2\alpha/(2\alpha+1)}$$

- **Remarque :** $\frac{KD}{n} + D^{-2\alpha}$ est un majorant du MISE, mais on peut montrer que cette majoration atteint la vitesse optimale.
- Sélection de modèle : critère de sélection automatique de D en fonction des données.

Sélection de modèle : heuristique

- Oracle :

$$D^* = \arg \min_{D \in \mathbb{N}^*} \left(\text{biais} + \frac{KD}{n} \right) \quad \text{avec}$$

$$\begin{cases} \text{biais} = \|f - f^D\|^2 = \|f\|^2 - \|f^D\|^2 = -\|f^D\|^2 + cte & \text{(Pythagore)} \\ \text{var} \leq \frac{KD}{n} \end{cases}$$

- On montre que

$$-\|\hat{f}_D\|^2 + \frac{KD}{n}$$

est un estimateur sans biais du biais (à cte additive près).

- On sélectionne alors le modèle :

$$\widehat{D} = \arg \min_D \left\{ -\|\hat{f}_D\|^2 + pen(D) \right\}$$

avec $pen(D) = 2KD/n$.

Théorème

Il existe une constante $C > 1$, et une constante $C' > 0$ telles que

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}} - f\right\|^2\right] \leq C \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{KD}{n}\right\} + \frac{C'}{n} \qquad (2)$$

- Soit D^* l'oracle alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} = \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{2KD}{n}\right\}$$

- De plus, soit la α régularité (inconnue) de f, alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} \propto n^{-2\alpha/(2\alpha+1)}$$

et $C'/n = \mathrm{o}(n^{-2\alpha/(2\alpha+1)})$.

- (2) est appelée Inégalité oracle. Elle stipule qu'à constante multiplicative près, l'estimateur de sélection de modèle est aussi performant que le meilleur modèle dans la collection.
- L'estimateur $\hat{f}_{\hat{D}}$ est dit adaptatif car il s'adapte à la régularité inconnue de f.

Théorème

Il existe une constante $C > 1$, et une constante $C' > 0$ telles que

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}} - f\right\|^2\right] \leq C \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{KD}{n}\right\} + \frac{C'}{n} \qquad (2)$$

- Soit D^* l'oracle alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} = \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{2KD}{n}\right\}$$

- De plus, soit la α régularité (inconnue) de f, alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} \propto n^{-2\alpha/(2\alpha+1)}$$

et $C'/n = \mathrm{o}(n^{-2\alpha/(2\alpha+1)})$.

- (2) est appelée Inégalité oracle. Elle stipule qu'à constante multiplicative près, l'estimateur de sélection de modèle est aussi performant que le meilleur modèle dans la collection.
- L'estimateur $\hat{f}_{\hat{D}}$ est dit adaptatif car il s'adapte à la régularité inconnue de f.

Théorème

Il existe une constante $C > 1$, *et une constante* $C' > 0$ *telles que*

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}} - f\right\|^2\right] \leq C \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{KD}{n}\right\} + \frac{C'}{n} \qquad (2)$$

- Soit D^* l'oracle alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} = \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{2KD}{n}\right\}$$

- De plus, soit la α régularité (inconnue) de f, alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} \propto n^{-2\alpha/(2\alpha+1)}$$

et $C'/n = \mathrm{o}(n^{-2\alpha/(2\alpha+1)})$.

- (2) est appelée Inégalité oracle. Elle stipule qu'à constante multiplicative près, l'estimateur de sélection de modèle est aussi performant que le meilleur modèle dans la collection.
- L'estimateur $\hat{f}_{\hat{D}}$ est dit adaptatif car il s'adapte à la régularité inconnue de f.

Théorème

Il existe une constante $C > 1$, *et une constante* $C' > 0$ *telles que*

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}} - f\right\|^2\right] \leq C \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{KD}{n}\right\} + \frac{C'}{n} \qquad (2)$$

- Soit D^* l'oracle alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} = \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{2KD}{n}\right\}$$

- De plus, soit la α régularité (inconnue) de f, alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} \propto n^{-2\alpha/(2\alpha+1)}$$

  et $C'/n = \mathrm{o}(n^{-2\alpha/(2\alpha+1)})$.
- (2) est appelée Inégalité oracle. Elle stipule qu'à constante multiplicative près, l'estimateur de sélection de modèle est aussi performant que le meilleur modèle dans la collection.
- L'estimateur $\hat{f}_{\hat{D}}$ est dit adaptatif car il s'adapte à la régularité inconnue de f.

Théorème

Il existe une constante $C > 1$, et une constante $C' > 0$ telles que

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}} - f\right\|^2\right] \leq C \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{KD}{n}\right\} + \frac{C'}{n} \qquad (2)$$

- Soit D^* l'oracle alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} = \inf_{D=1,\dots,n}\left\{\|f_D - f\|^2 + \frac{2KD}{n}\right\}$$

- De plus, soit la α régularité (inconnue) de f, alors

$$\|f_{D^*} - f\|^2 + \frac{2KD^*}{n} \propto n^{-2\alpha/(2\alpha+1)}$$

  et $C'/n = \mathrm{o}(n^{-2\alpha/(2\alpha+1)})$.
- (2) est appelée Inégalité oracle. Elle stipule qu'à constante multiplicative près, l'estimateur de sélection de modèle est aussi performant que le meilleur modèle dans la collection.
- L'estimateur $\hat{f}_{\hat{D}}$ est dit adaptatif car il s'adapte à la régularité inconnue de f.

Sélection de modèle : commentaires

- Dans les exemples précédents, le choix d'un modèle est équivalent au choix d'une dimension D, mais la sélection de modèle se généralise au cas de plusieurs modèles de même dimension.
- D'autres procédures de choix de D existent, basées sur des heuristiques différentes. Exples
  - Sélection de modèle pour les histogrammes irréguliers : $pen(m) = \log(D/n)KD/n$
  - Règle empirique de Sturges pour les histogrammes réguliers (fondé sur la loi normale) : $D = 1 + \log_2 n$
  - ...

Sélection de modèle : commentaires

- Dans les exemples précédents, le choix d'un modèle est équivalent au choix d'une dimension D, mais la sélection de modèle se généralise au cas de plusieurs modèles de même dimension.
- D'autres procédures de choix de D existent, basées sur des heuristiques différentes. Exples
  - Sélection de modèle pour les histogrammes irréguliers :
    $$pen(m) = \log(D/n)KD/n$$
  - Règle empirique de Sturges pour les histogrammes réguliers (fondé sur la loi normale) : $D = 1 + \log_2 n$
  - ...

Sélection de modèle : commentaires

- Dans les exemples précédents, le choix d'un modèle est équivalent au choix d'une dimension D, mais la sélection de modèle se généralise au cas de plusieurs modèles de même dimension.
- D'autres procédures de choix de D existent, basées sur des heuristiques différentes. Exples
  - Sélection de modèle pour les histogrammes irréguliers :
    $$pen(m) = \log(D/n)KD/n$$
  - Règle empirique de Sturges pour les histogrammes réguliers (fondé sur la loi normale) : $D = 1 + \log_2 n$
  - ...

Sélection de modèle : commentaires

- Dans les exemples précédents, le choix d'un modèle est équivalent au choix d'une dimension D, mais la sélection de modèle se généralise au cas de plusieurs modèles de même dimension.
- D'autres procédures de choix de D existent, basées sur des heuristiques différentes. Exples
  - Sélection de modèle pour les histogrammes irréguliers :
    $$pen(m) = \log(D/n)KD/n$$
  - Règle empirique de Sturges pour les histogrammes réguliers (fondé sur la loi normale) : $D = 1 + \log_2 n$
  - ...

Sélection de modèle : commentaires

- Dans les exemples précédents, le choix d'un modèle est équivalent au choix d'une dimension D, mais la sélection de modèle se généralise au cas de plusieurs modèles de même dimension.
- D'autres procédures de choix de D existent, basées sur des heuristiques différentes. Exples
  - Sélection de modèle pour les histogrammes irréguliers :
    $$pen(m) = \log(D/n)KD/n$$
  - Règle empirique de Sturges pour les histogrammes réguliers (fondé sur la loi normale) : $D = 1 + \log_2 n$
  - ...

Estimation par noyau

- Soit $X_1, \ldots, X_n$ i.i.d. de densité f et de fdr F.
- Pour h assez petit, on a

$$f(x) = F'(x) \simeq \frac{F(x+h) - F(x-h)}{2h}.$$

- D'où l'estimateur

$$\begin{aligned} \hat{f}_h(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} 1\{X_i \in]x-h; x+h]\} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right) \end{aligned}$$

où $K_0(u) = 1_{]-1;1]}(u)/2$ est le noyau rectangulaire.

- $\hat{f}_h(x)$ est la fréquence des observations dans l'intervalle $]x-h, x+h]$
- Le paramètre $h > 0$ est appelé fenêtre.

Estimation par noyau

- Soit $X_1, \ldots, X_n$ i.i.d. de densité f et de fdr F.
- Pour h assez petit, on a

$$f(x) = F'(x) \simeq \frac{F(x+h) - F(x-h)}{2h}.$$

- D'où l'estimateur

$$\begin{aligned} \hat{f}_h(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} 1\{X_i \in]x-h; x+h]\} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right) \end{aligned}$$

où $K_0(u) = 1_{]-1;1]}(u)/2$ est le noyau rectangulaire.

- $\hat{f}_h(x)$ est la fréquence des observations dans l'intervalle $]x-h, x+h]$
- Le paramètre $h > 0$ est appelé fenêtre.

Estimation par noyau

- Soit $X_1, \ldots, X_n$ i.i.d. de densité f et de fdr F.
- Pour h assez petit, on a

$$f(x) = F'(x) \simeq \frac{F(x+h) - F(x-h)}{2h}.$$

- D'où l'estimateur

$$\begin{aligned} \hat{f}_h(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} 1\{X_i \in]x-h; x+h]\} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right) \end{aligned}$$

où $K_0(u) = 1_{]-1;1]}(u)/2$ est le noyau rectangulaire.

- $\hat{f}_h(x)$ est la fréquence des observations dans l'intervalle $]x-h, x+h]$
- Le paramètre $h > 0$ est appelé fenêtre.

Estimation par noyau

- Soit $X_1, \ldots, X_n$ i.i.d. de densité f et de fdr F.
- Pour h assez petit, on a

$$f(x) = F'(x) \simeq \frac{F(x+h) - F(x-h)}{2h}.$$

- D'où l'estimateur

$$\begin{aligned} \hat{f}_h(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} 1\{X_i \in]x-h; x+h]\} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right) \end{aligned}$$

où $K_0(u) = 1_{]-1;1]}(u)/2$ est le noyau rectangulaire.

- $\hat{f}_h(x)$ est la fréquence des observations dans l'intervalle $]x-h, x+h]$
- Le paramètre $h > 0$ est appelé fenêtre.

Estimation par noyau

- Soit $X_1, \ldots, X_n$ i.i.d. de densité f et de fdr F.
- Pour h assez petit, on a

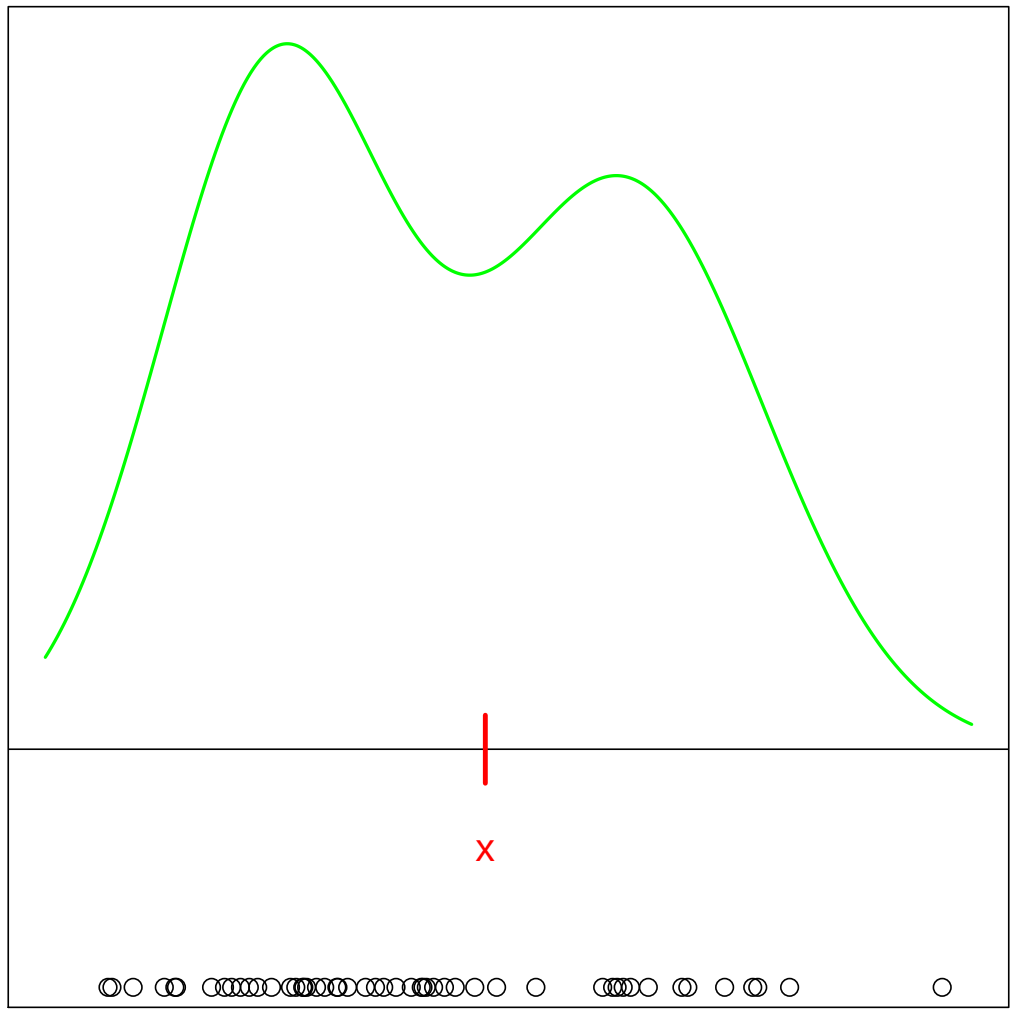
$$f(x) = F'(x) \simeq \frac{F(x+h) - F(x-h)}{2h}.$$

- D'où l'estimateur

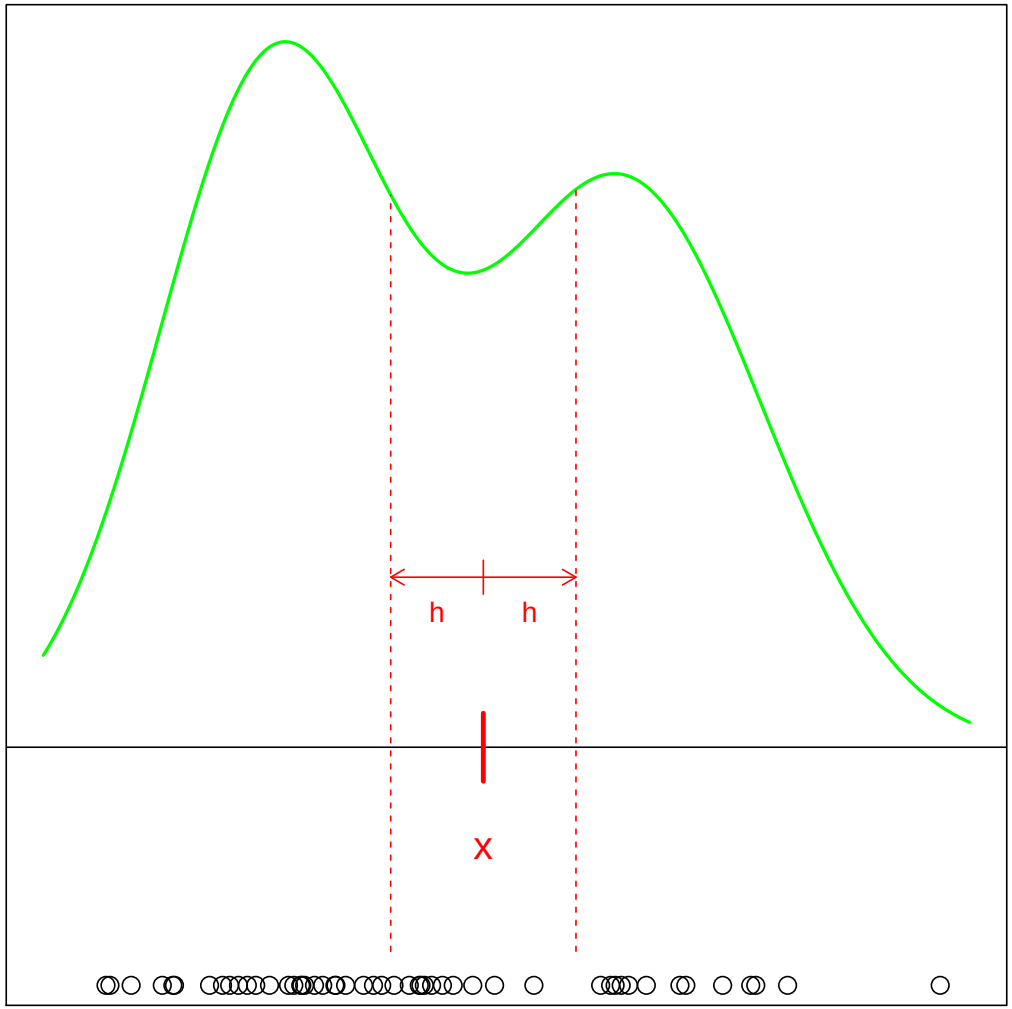
$$\begin{aligned} \hat{f}_h(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} 1\{X_i \in]x-h; x+h]\} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right) \end{aligned}$$

  où $K_0(u) = 1_{]-1;1]}(u)/2$ est le noyau rectangulaire.
- $\hat{f}_h(x)$ est la fréquence des observations dans l'intervalle $]x-h, x+h]$
- Le paramètre $h > 0$ est appelé fenêtre.

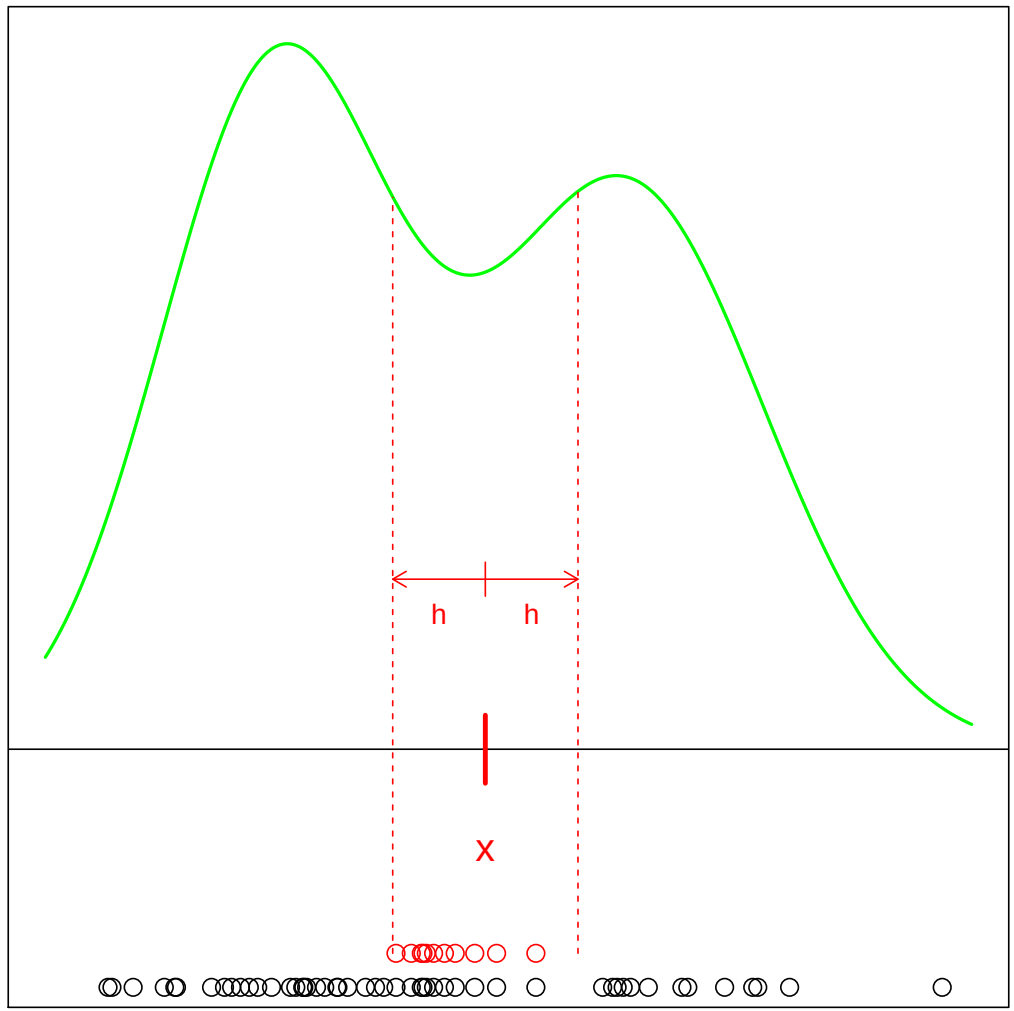
Estimation par noyau



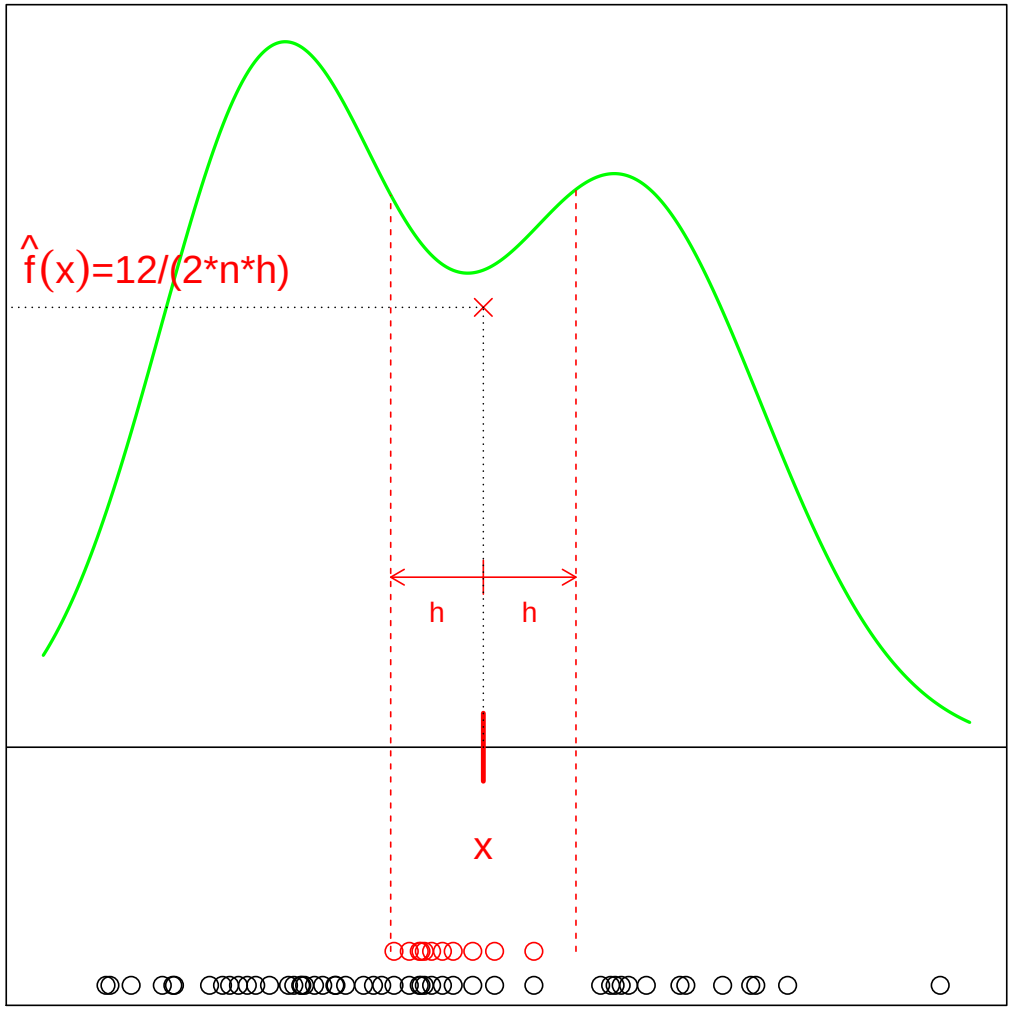
Estimation par noyau



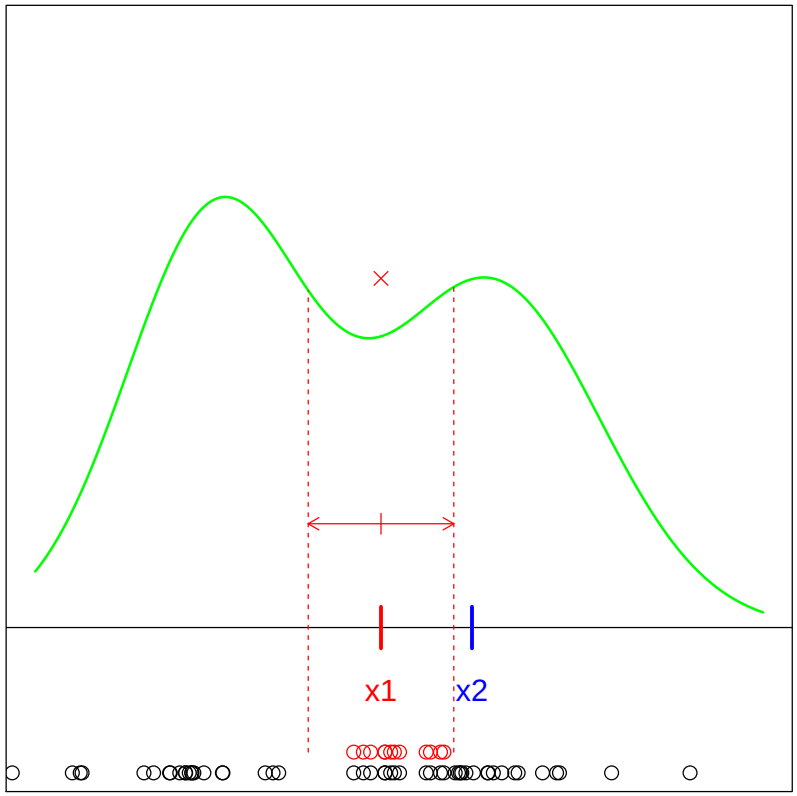
Estimation par noyau



Estimation par noyau



Mise en perspective : histogrammes et estimateurs à noyau

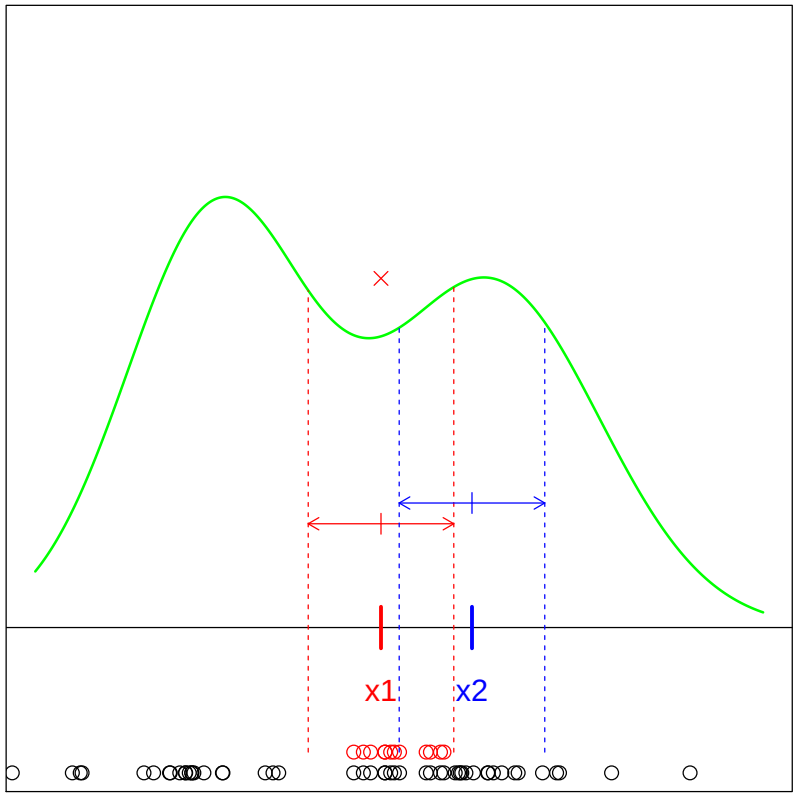


Estimateur à noyau

Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau

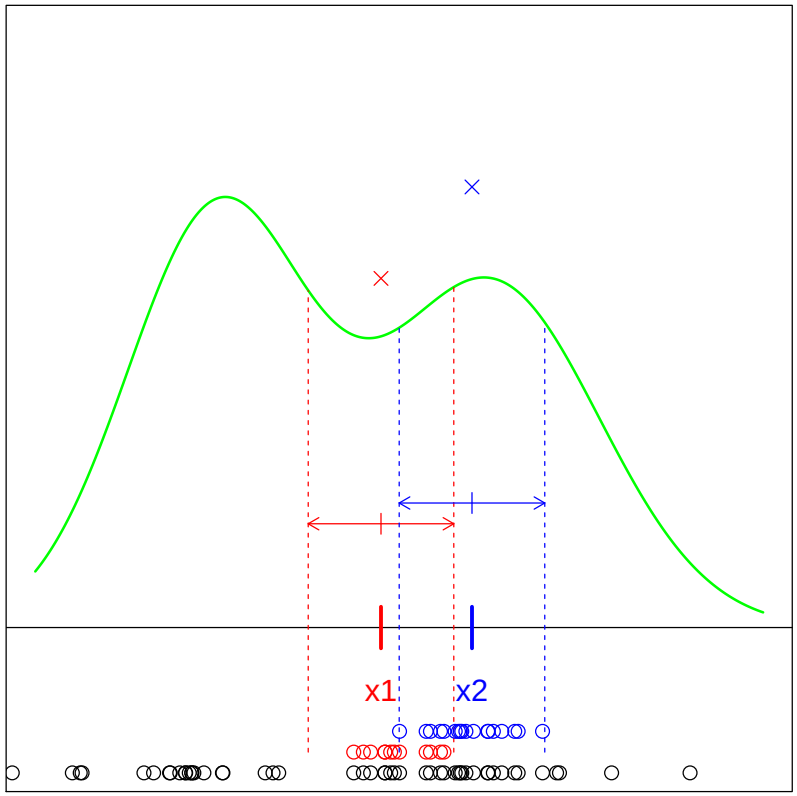


Estimateur à noyau

Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau

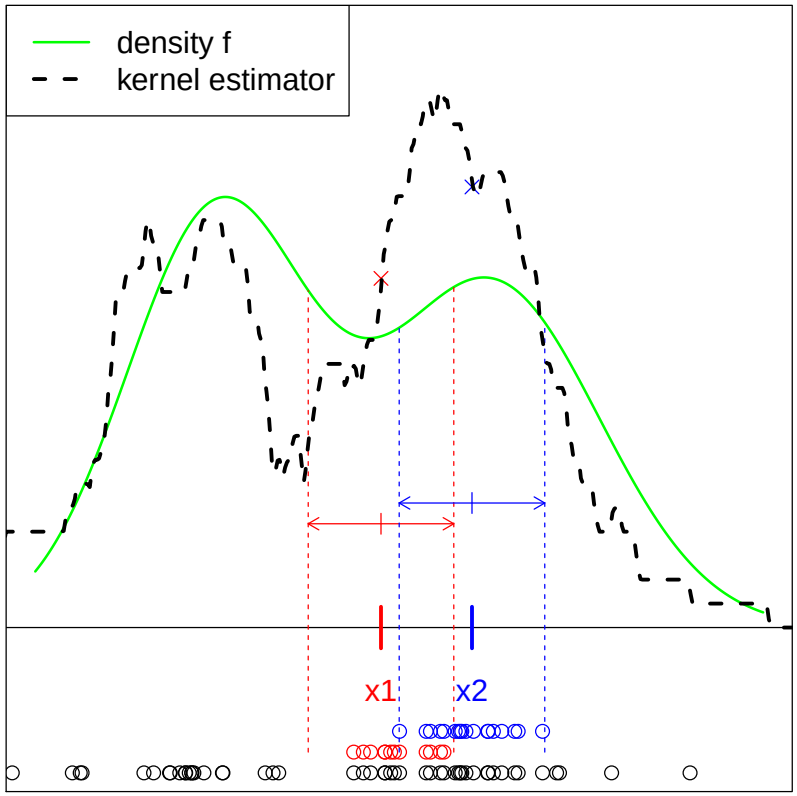


Estimateur à noyau

Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau

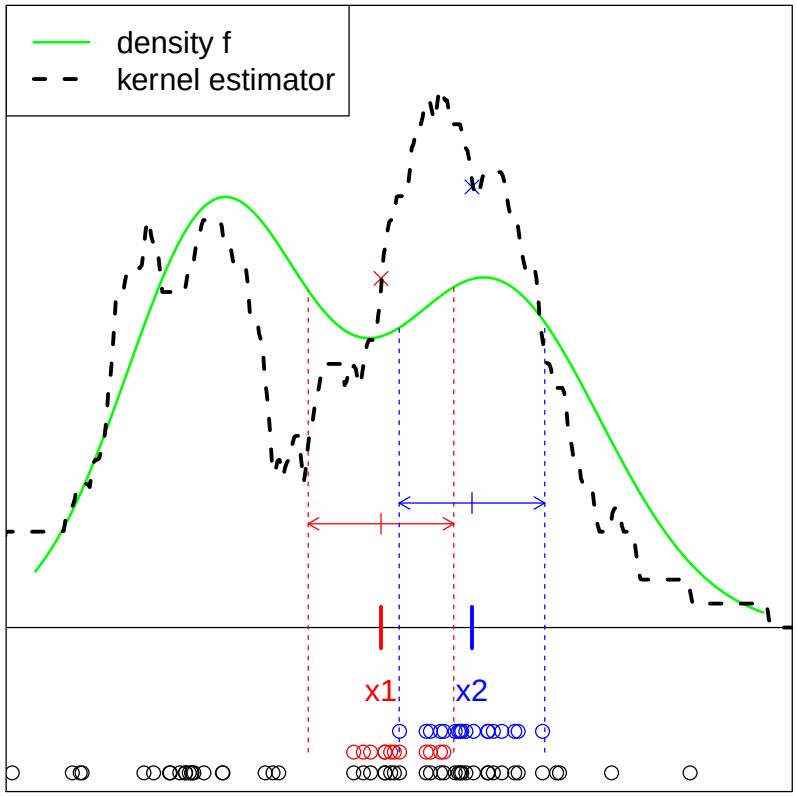


Estimateur à noyau

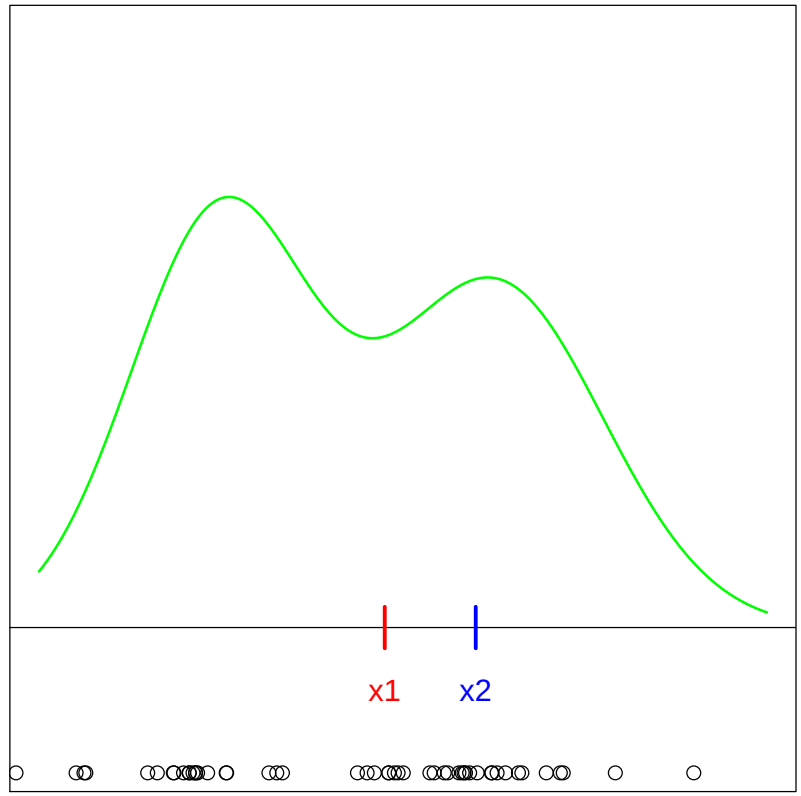
Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau



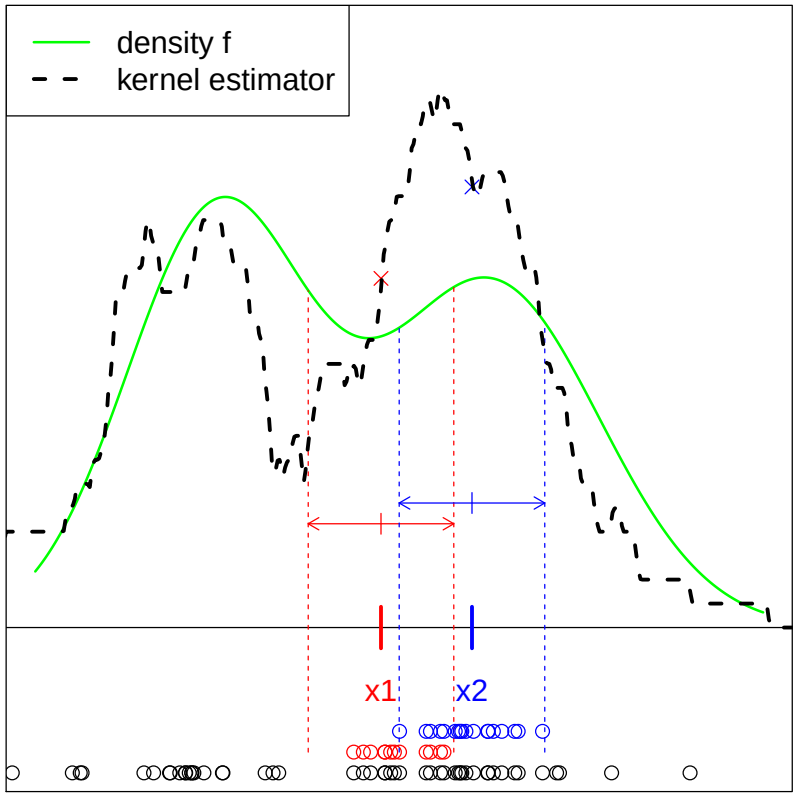
Estimateur à noyau



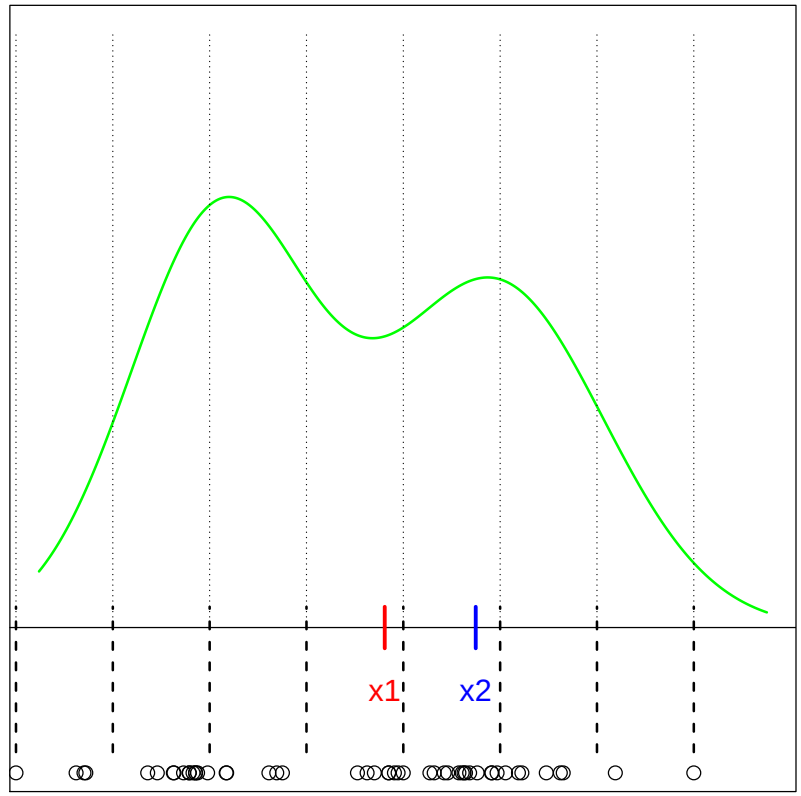
Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau



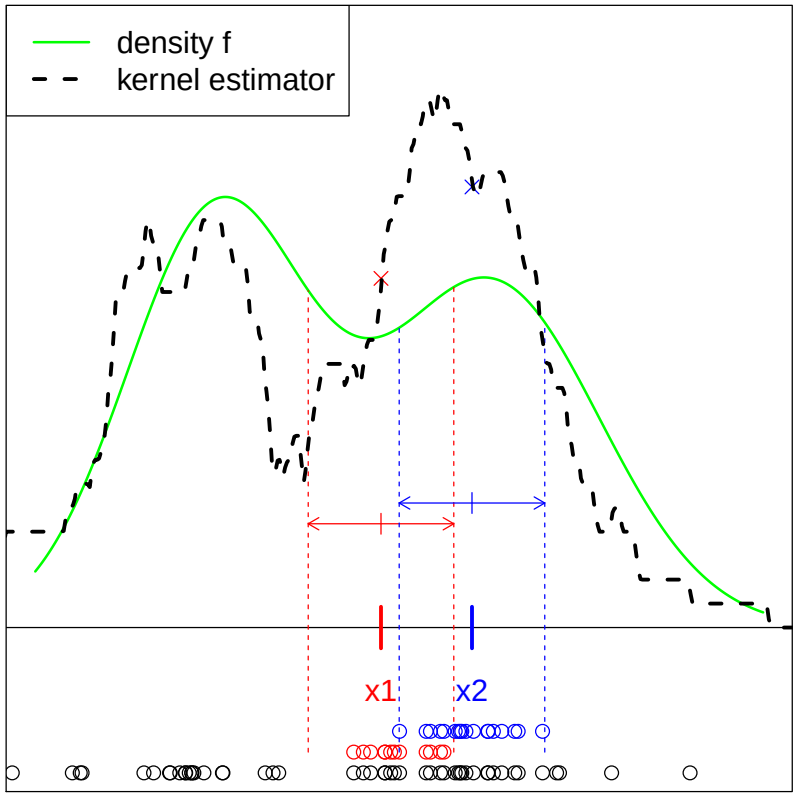
Estimateur à noyau



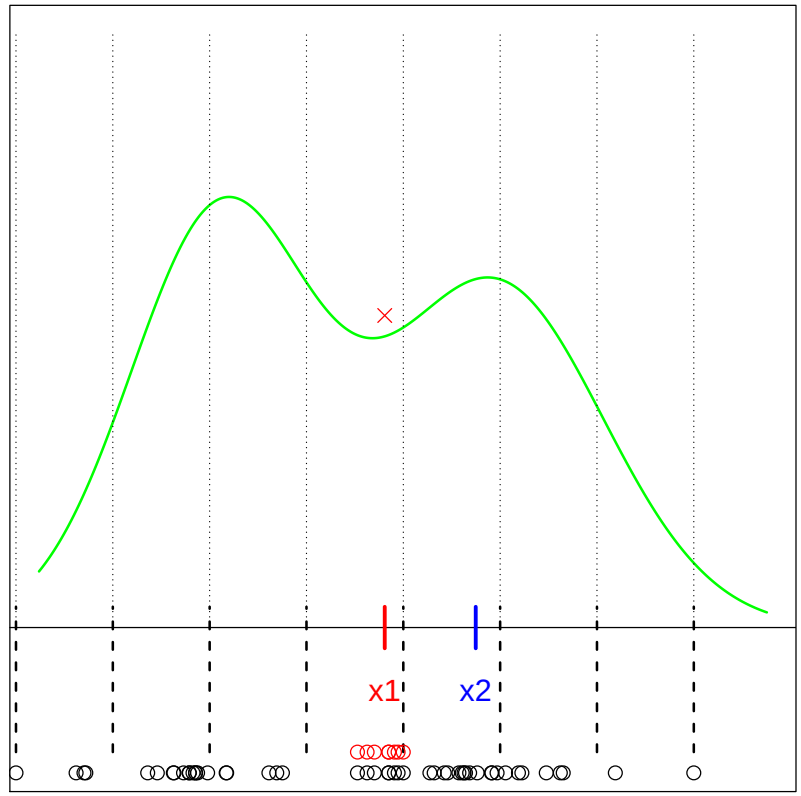
Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau



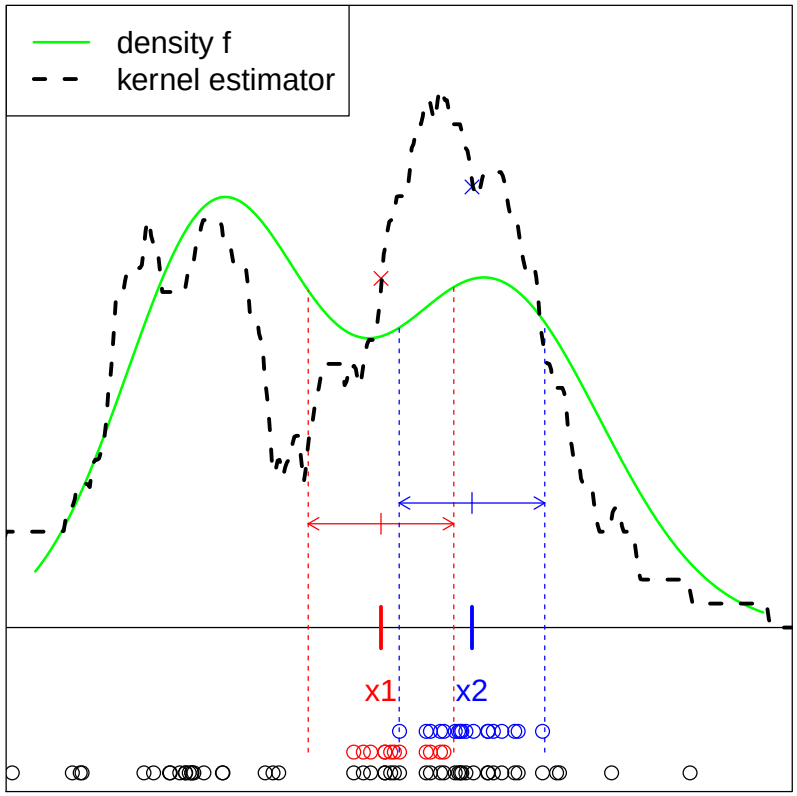
Estimateur à noyau



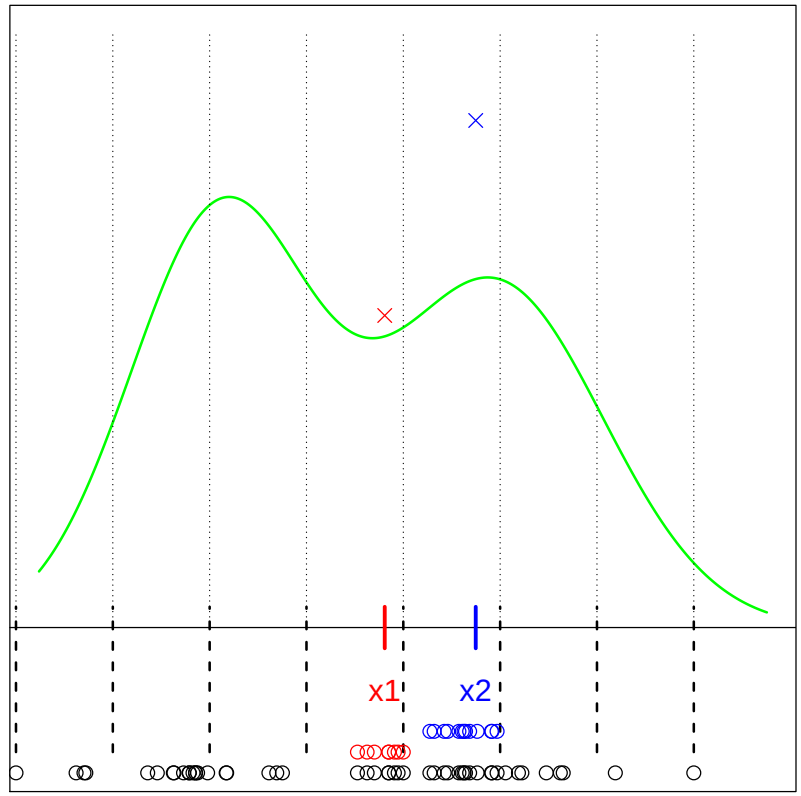
Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau



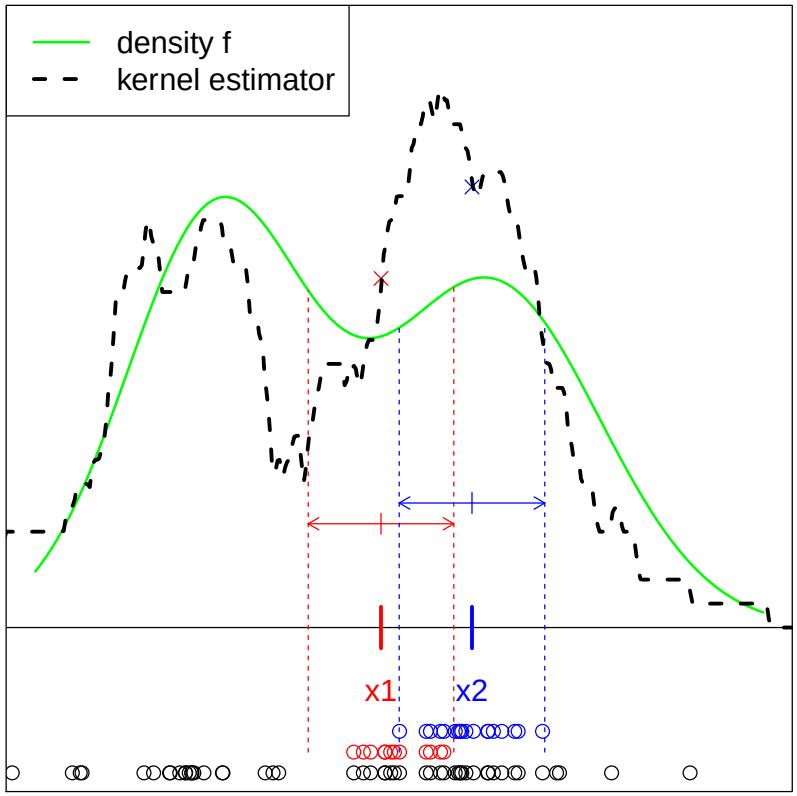
Estimateur à noyau



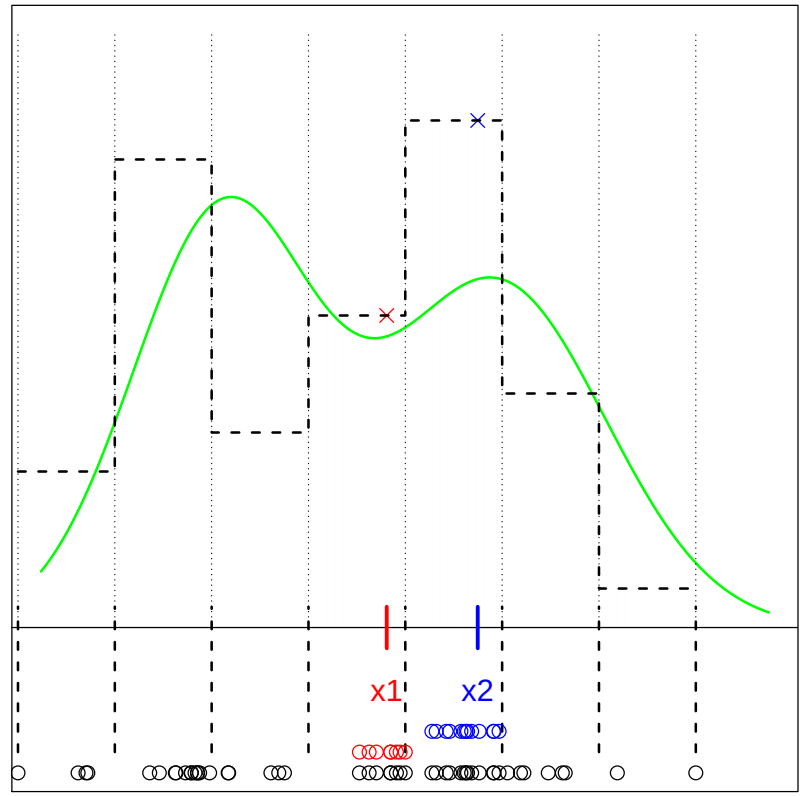
Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau



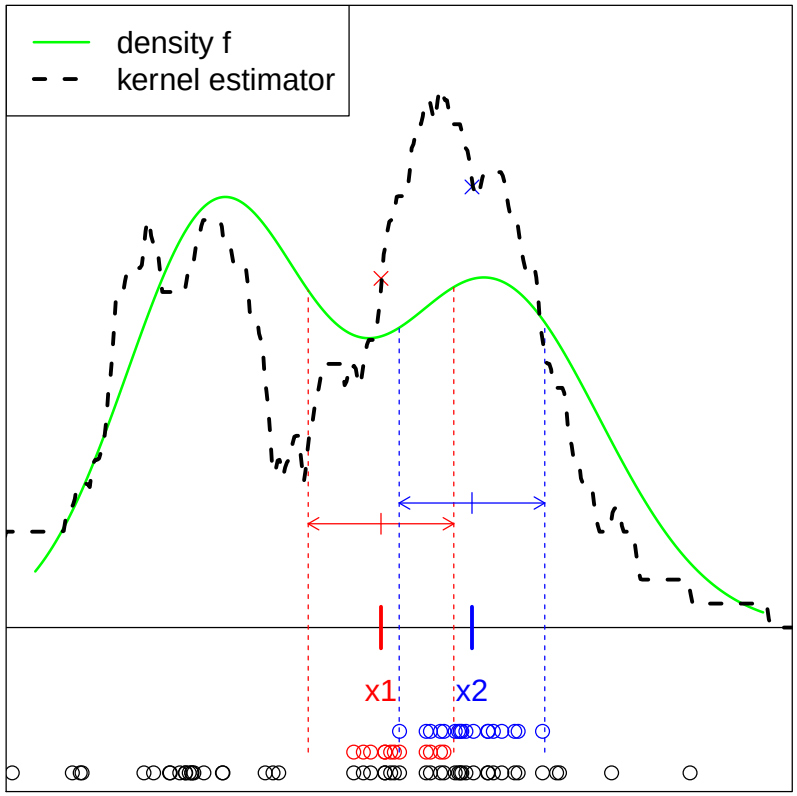
Estimateur à noyau



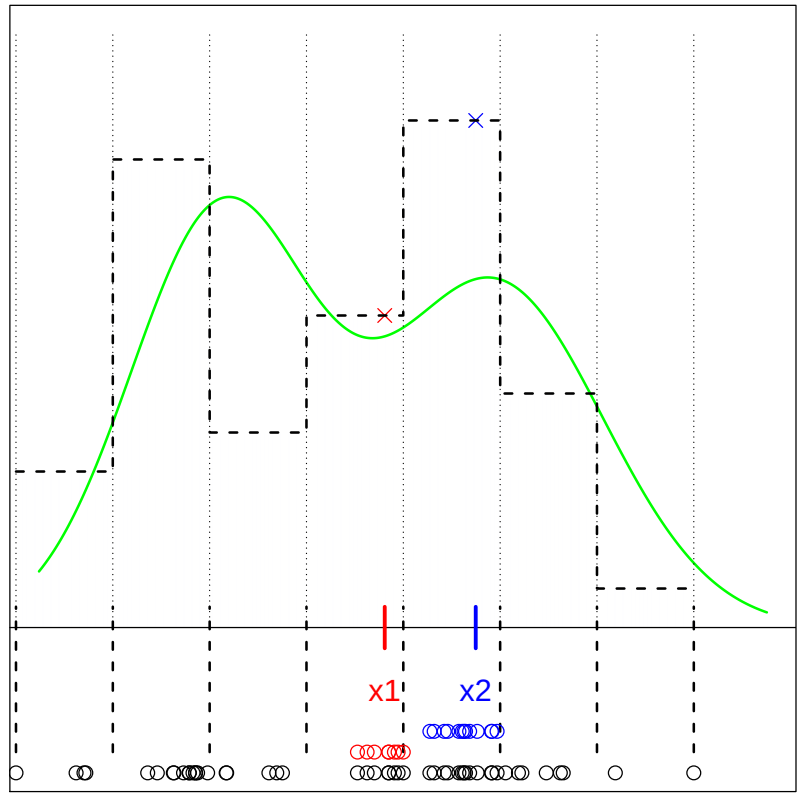
Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau



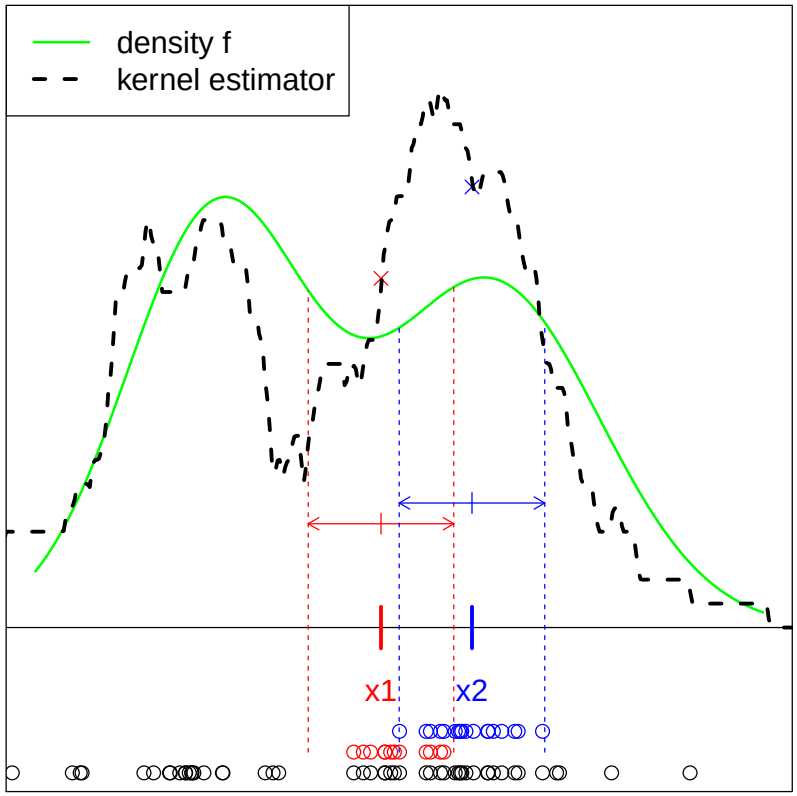
Estimateur à noyau



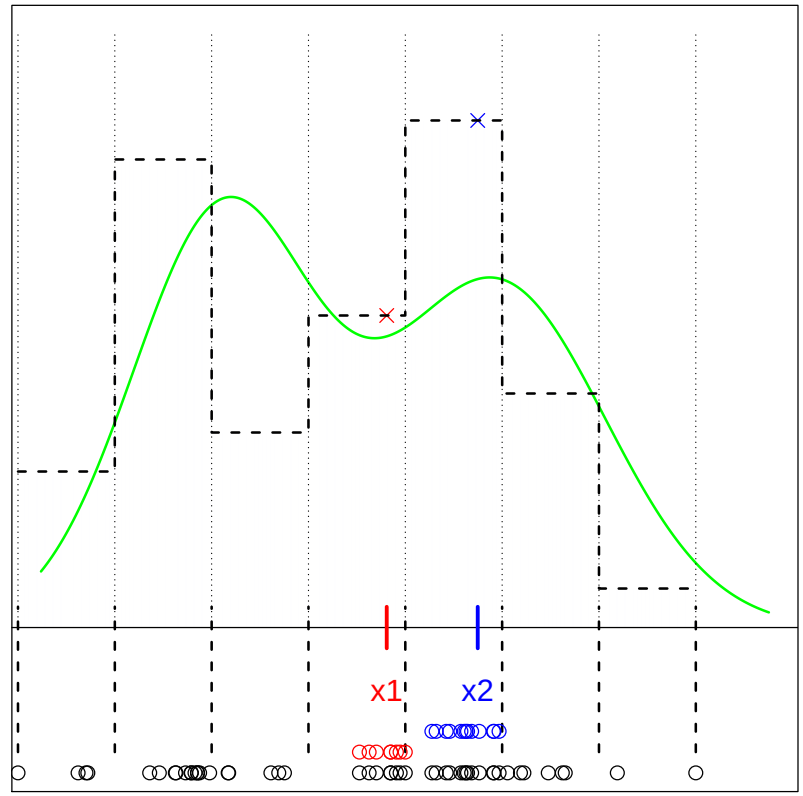
Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Mise en perspective : histogrammes et estimateurs à noyau



Estimateur à noyau



Histogramme

- Histogramme : moyenne sur un intervalle fixé
- Noyau : moyenne sur un intervalle glissant
- Remarque : l'estimateur par noyau fait des "sauts".

Estimateur par noyaux : cas général

- Estimateur par noyau rectangulaire :

$$\hat{f}_h(x) = \frac{1}{h} \sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right) \quad \text{avec} \quad K_0(u) = 1_{]-1;1]}(u)/2$$

- Généralisation : $\hat{f}_h(x) = \frac{1}{h} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$

  avec $K : \mathbb{R} \to \mathbb{R}$ tel que $\int K = 1$.
- Exemples :
- Le noyau donne un poids plus fort aux observations X_i qui sont proches de x

Estimateur par noyaux : cas général

- Estimateur par noyau rectangulaire :

$$\hat{f}_h(x) = \frac{1}{h} \sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right) \quad \text{avec} \quad K_0(u) = 1_{]-1;1]}(u)/2$$

- Généralisation : $\hat{f}_h(x) = \frac{1}{h} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$

  avec $K : \mathbb{R} \to \mathbb{R}$ tel que $\int K = 1$.
- Exemples :
- Le noyau donne un poids plus fort aux observations X_i qui sont proches de x

Estimateur par noyaux : cas général

- Estimateur par noyau rectangulaire :

$$\hat{f}_h(x) = \frac{1}{h} \sum_{i=1}^{n} K_0 \left(\frac{X_i - x}{h} \right) \quad \text{avec} \quad K_0(u) = 1_{]-1;1]}(u)/2$$

- Généralisation : $\hat{f}_h(x) = \frac{1}{h} \sum_{i=1}^{n} K \left(\frac{X_i - x}{h} \right)$

  avec $K : \mathbb{R} \to \mathbb{R}$ tel que $\int K = 1$.

- Exemples :
- Le noyau donne un poids plus fort aux observations X_i qui sont proches de x

Estimateur par noyaux : cas général

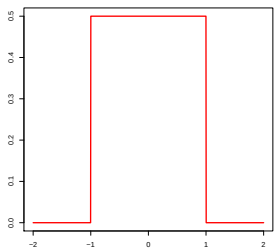
- Estimateur par noyau rectangulaire :

$$\hat{f}_h(x) = \frac{1}{h}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right) \quad \text{avec} \quad K_0(u) = 1_{]-1;1]}(u)/2$$

- Généralisation : $\hat{f}_h(x) = \frac{1}{h}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$

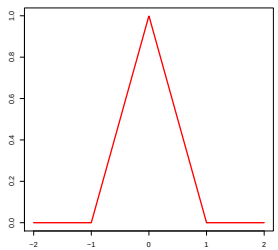
  avec $K : \mathbb{R} \to \mathbb{R}$ tel que $\int K = 1$.

- Exemples :



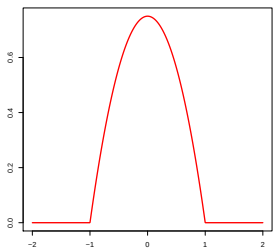
Noyau rectangulaire

$1_{]-1;1]}(u)/2$



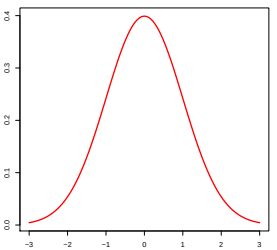
Noyau triangulaire

$(1 - |u|)1_{|u|\leq 1}$



Noyau d'Epanechnikov

$0.75 * (1 - u^2)1_{|u|\leq 1}$



Noyau gaussien

$e^{(-u^2/2)}/\sqrt{2\pi}$

- Le noyau donne un poids plus fort aux observations X_i qui sont proches de x

Estimateur par noyaux : cas général

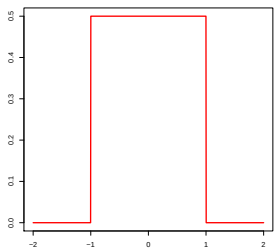
- Estimateur par noyau rectangulaire :

$$\hat{f}_h(x) = \frac{1}{h}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right) \quad \text{avec} \quad K_0(u) = 1_{]-1;1]}(u)/2$$

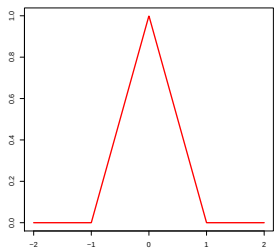
- Généralisation : $\hat{f}_h(x) = \frac{1}{h}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$

  avec $K : \mathbb{R} \to \mathbb{R}$ tel que $\int K = 1$.

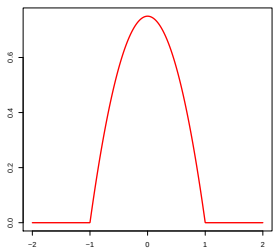
- Exemples :



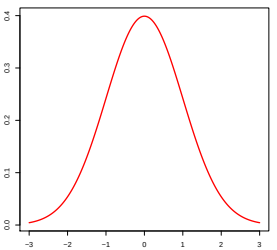
Noyau rectangulaire
$1_{]-1;1]}(u)/2$



Noyau triangulaire
$(1 - |u|)1_{|u|\leq 1}$



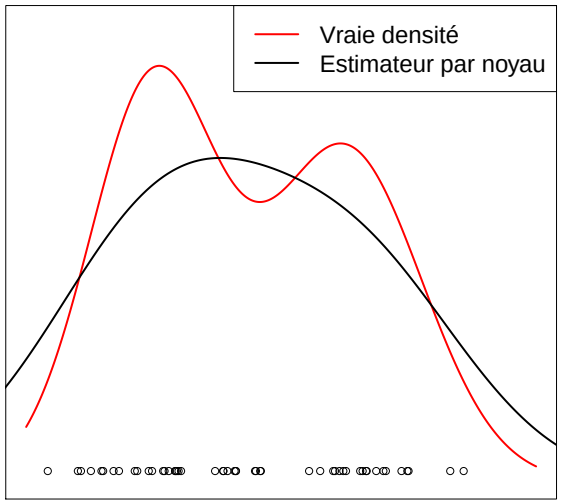
Noyau d'Epanechnikov
$0.75 * (1 - u^2)1_{|u|\leq 1}$



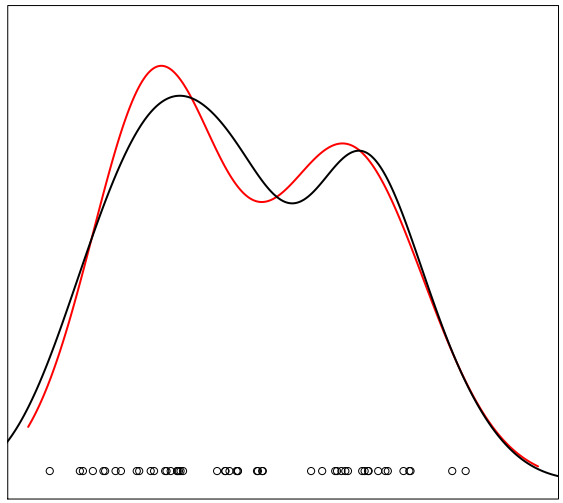
Noyau gaussien
$e^{(-u^2/2)}/\sqrt{2\pi}$

- Le noyau donne un poids plus fort aux observations X_i qui sont proches de x

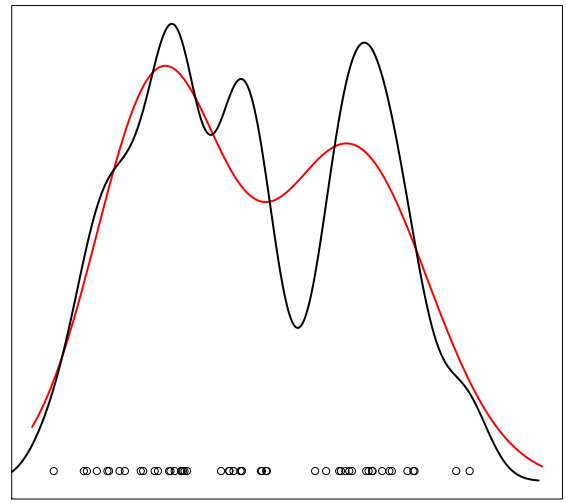
Effet de la variation de h sur l'estimateur à noyau



h grand



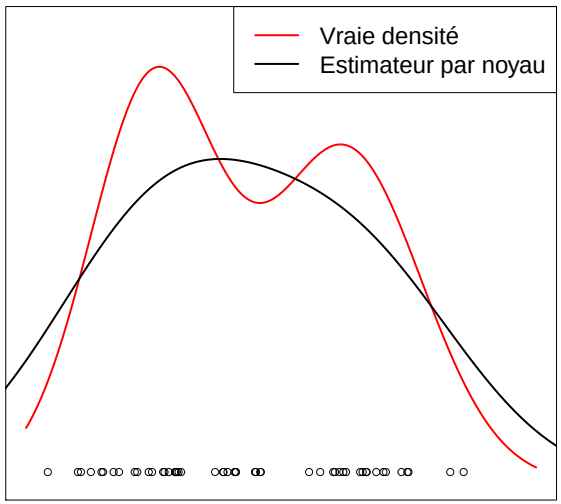
h optimal

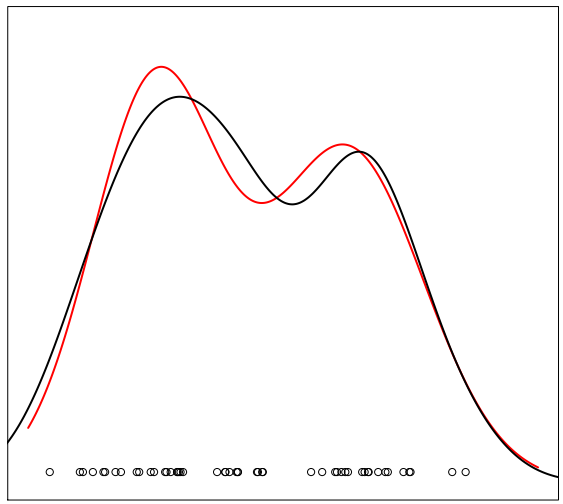


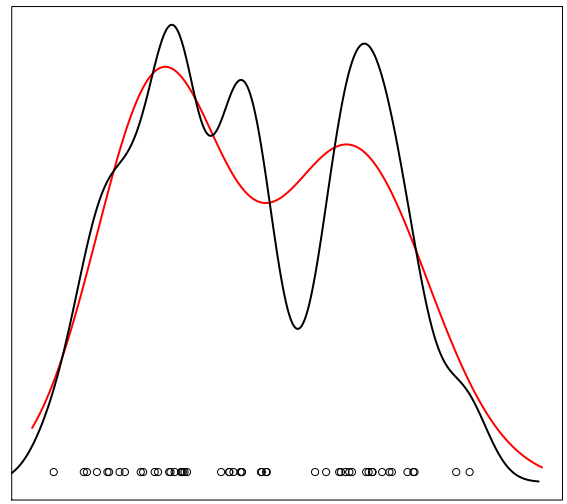
h petit

- Plus h est grand, plus l'estimateur est lissé.
- La fenêtre h est l'équivalent de $1/D$ en projection, où D est la dimension du modèle.

Effet de la variation de h sur l'estimateur à noyau



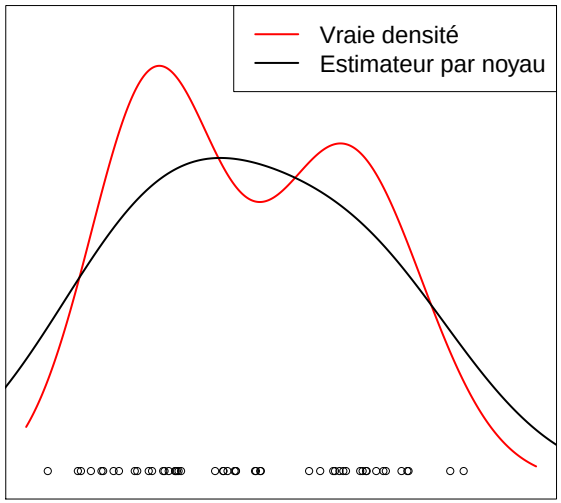


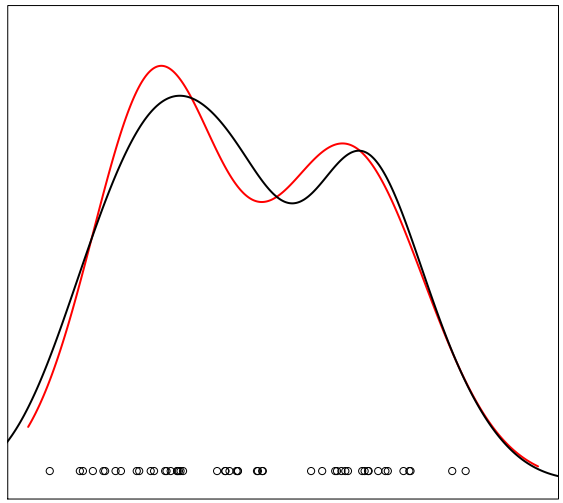


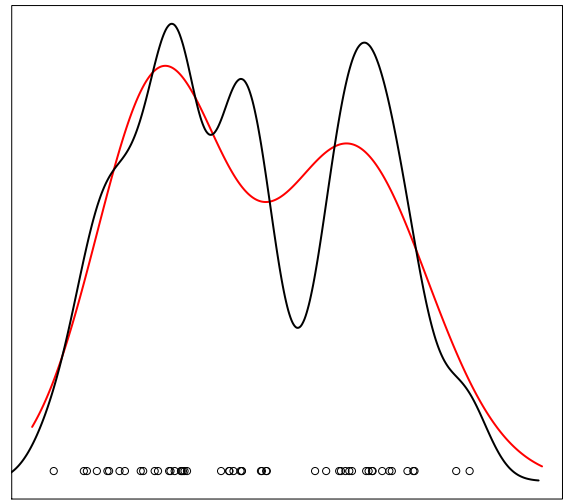
h grand h optimal h petit

- Plus h est grand, plus l'estimateur est lissé.
- La fenêtre h est l'équivalent de $1/D$ en projection, où D est la dimension du modèle.

Effet de la variation de h sur l'estimateur à noyau







h grand h optimal h petit

- Plus h est grand, plus l'estimateur est lissé.
- La fenêtre h est l'équivalent de $1/D$ en projection, où D est la dimension du modèle.

MSE : heuristique avec le noyau rectangulaire

Soit x_0 fixé,

$$MSE = \mathbb{E}[(\hat{f}_h - f)^2(x_0)] = (f_h - f(x_0))^2 + \mathsf{var}(\hat{f}_h(x_0))$$

avec

$$\hat{f}_h(x_0) = \frac{1}{2h} \sum_{i=1}^{n} 1_{X_i \in [x_0-h, x_0+h]}$$

et $f_h(x_0) = \mathbb{E}[\hat{f}_h(x_0)]$.

Variance

$$\mathsf{var}(\hat{f}_h(x_0)) = \frac{1}{4nh^2} \mathsf{var}\left(1_{X_1 \in [x_0-h, x_0+h]}\right)$$

$1_{X_1 \in [x_0-h, x_0+h]}$ est une variable de Bernouilli de paramètre :

$$p = \mathbb{P}\left[X_1 \in [x_0 - h, x_0 + h]\right] = \int_{x_0-h}^{x_0+h} f(x)\,dx.$$

En supposant h petit, $p \simeq 2hf(x_0)$, et $1 - p \simeq 1 - 2hf(x_0) \simeq 1$.
De plus, la variance d'une Bernouilli vaut $p(1-p) < 1/4$, d'où :

$$\mathsf{var}(\hat{f}_h(x_0)) \simeq \frac{\mathsf{cte}}{nh}$$

MSE : heuristique avec le noyau rectangulaire

Soit x_0 fixé,

$$MSE = \mathbb{E}[(\hat{f}_h - f)^2(x_0)] = (f_h - f(x_0))^2 + \text{var}(\hat{f}_h(x_0))$$

avec

$$\hat{f}_h(x_0) = \frac{1}{2h} \sum_{i=1}^{n} 1_{X_i \in [x_0-h, x_0+h]}$$

et $f_h(x_0) = \mathbb{E}[\hat{f}_h(x_0)]$.

Variance

$$\text{var}(\hat{f}_h(x_0)) = \frac{1}{4nh^2} \text{var}\left(1_{X_1 \in [x_0-h, x_0+h]}\right)$$

$1_{X_1 \in [x_0-h, x_0+h]}$ est une variable de Bernouilli de paramètre :

$$p = \mathbb{P}\left[X_1 \in [x_0 - h, x_0 + h]\right] = \int_{x_0-h}^{x_0+h} f(x)\,dx.$$

En supposant h petit, $p \simeq 2hf(x_0)$, et $1 - p \simeq 1 - 2hf(x_0) \simeq 1$.
De plus, la variance d'une Bernouilli vaut $p(1-p) < 1/4$, d'où :

$$\text{var}(\hat{f}_h(x_0)) \simeq \frac{\text{cte}}{nh}$$

MSE : heuristique avec le noyau rectangulaire (2)

- $\hat{f}_h(x_0) = (1/2nh) \sum_{i=1}^n 1_{X_i \in [x_0-h, x_0+h]}$ donc

$$f_h(x_0) = \mathbb{E}[\hat{f}_h(x_0)] = \frac{1}{2h}\mathbb{P}[X_1 \in [x_0-h, x_0+h]] = \frac{1}{2h}\int_{x_0-h}^{x_0+h} f(x)\,dx$$

$\hookrightarrow$ $f_h(x_0)$ est la moyenne de f sur l'intervalle $[x_0 - h, x_0 + h]$.

- Si $f \in \mathcal{C}^1$ et $\|f'\|_\infty < \infty$, d'après la formule de Taylor,pour tout $x \in [x_0 - h, x_0 + h]$ il existe x_1 tq

$$f(x) = f(x_0) + (x - x_0)f'(x_1)$$

- Biais : si f est bornée presque surement,

$$\begin{aligned} |f_h(x_0) - f(x_0)| &= \left| \frac{1}{2h}\int_{x_0-h}^{x_0+h} \left(f(x_0) + (x - x_0)f'(x_1)\right) dx - f(x_0) \right| \\ &\leq \frac{1}{2h}\|f'\|_\infty \left| \int_{x_0-h}^{x_0+h} |x - x_0|\,dx \right| = \|f'\|_\infty \frac{2h^2}{2h} = \text{cte} \times h \end{aligned}$$

MSE : heuristique avec le noyau rectangulaire (2)

- $\hat{f}_h(x_0) = (1/2nh) \sum_{i=1}^n 1_{X_i \in [x_0-h, x_0+h]}$ donc

$$f_h(x_0) = \mathbb{E}[\hat{f}_h(x_0)] = \frac{1}{2h}\mathbb{P}[X_1 \in [x_0-h, x_0+h]] = \frac{1}{2h}\int_{x_0-h}^{x_0+h} f(x)\,dx$$

$\hookrightarrow$ $f_h(x_0)$ est la moyenne de f sur l'intervalle $[x_0 - h, x_0 + h]$.

- Si $f \in \mathcal{C}^1$ et $\|f'\|_\infty < \infty$, d'après la formule de Taylor,pour tout $x \in [x_0 - h, x_0 + h]$ il existe x_1 tq

$$f(x) = f(x_0) + (x - x_0)f'(x_1)$$

- Biais : si f est bornée presque surement,

$$\begin{aligned} |f_h(x_0) - f(x_0)| &= \left| \frac{1}{2h}\int_{x_0-h}^{x_0+h} \left(f(x_0) + (x - x_0)f'(x_1)\right) dx - f(x_0) \right| \\ &\leq \frac{1}{2h}\|f'\|_\infty \left| \int_{x_0-h}^{x_0+h} |x - x_0|\,dx \right| = \|f'\|_\infty \frac{2h^2}{2h} = \text{cte} \times h \end{aligned}$$

MSE : heuristique avec le noyau rectangulaire (2)

- $\hat{f}_h(x_0) = (1/2nh) \sum_{i=1}^n 1_{X_i \in [x_0-h, x_0+h]}$ donc

$$f_h(x_0) = \mathbb{E}[\hat{f}_h(x_0)] = \frac{1}{2h}\mathbb{P}[X_1 \in [x_0-h, x_0+h]] = \frac{1}{2h}\int_{x_0-h}^{x_0+h} f(x)\,dx$$

$\hookrightarrow$ $f_h(x_0)$ est la moyenne de f sur l'intervalle $[x_0 - h, x_0 + h]$.

- Si $f \in \mathcal{C}^1$ et $\|f'\|_\infty < \infty$, d'après la formule de Taylor,pour tout $x \in [x_0 - h, x_0 + h]$ il existe x_1 tq

$$f(x) = f(x_0) + (x - x_0)f'(x_1)$$

- Biais : si f est bornée presque surement,

$$\begin{aligned} |f_h(x_0) - f(x_0)| &= \left| \frac{1}{2h}\int_{x_0-h}^{x_0+h} \left(f(x_0) + (x - x_0)f'(x_1)\right) dx - f(x_0) \right| \\ &\leq \frac{1}{2h}\|f'\|_\infty \left| \int_{x_0-h}^{x_0+h} |x - x_0|\,dx \right| = \|f'\|_\infty \frac{2h^2}{2h} = \text{cte} \times h \end{aligned}$$

MSE : heuristique avec le noyau rectangulaire (2)

- $\hat{f}_h(x_0) = (1/2nh) \sum_{i=1}^n 1_{X_i \in [x_0-h, x_0+h]}$ donc

$$f_h(x_0) = \mathbb{E}[\hat{f}_h(x_0)] = \frac{1}{2h}\mathbb{P}[X_1 \in [x_0-h, x_0+h]] = \frac{1}{2h}\int_{x_0-h}^{x_0+h} f(x)\,dx$$

$\hookrightarrow$ $f_h(x_0)$ est la moyenne de f sur l'intervalle $[x_0 - h, x_0 + h]$.

- Si $f \in \mathcal{C}^1$ et $\|f'\|_\infty < \infty$, d'après la formule de Taylor,pour tout $x \in [x_0 - h, x_0 + h]$ il existe x_1 tq

$$f(x) = f(x_0) + (x - x_0)f'(x_1)$$

- Biais : si f est bornée presque surement,

$$\begin{aligned} |f_h(x_0) - f(x_0)| &= \left| \frac{1}{2h}\int_{x_0-h}^{x_0+h} \left(f(x_0) + (x - x_0)f'(x_1)\right) dx - f(x_0)\right| \\ &\leq \frac{1}{2h}\|f'\|_\infty \left|\int_{x_0-h}^{x_0+h} |x - x_0|\,dx\right| = \|f'\|_\infty \frac{2h^2}{2h} = \text{cte} \times h \end{aligned}$$

MSE des estimateurs par noyaux : cas général

Notons

$$\mathcal{C}^{\ell *} = \{f \in \mathcal{C}^{\ell}, \|f^{(\ell)}\|_{\infty} < \infty.$$

- Biais Si $f \in \mathcal{C}^{\ell *}$ et K est suffisamment régulier (K "noyau d'ordre ℓ") alors

$$(f_h(x_0) - f(x_0))^2 \leq \kappa_0 h^{2\ell}$$

↪ Remarque : Analogie avec le biais des estimateurs par projection :

$$D^{-2\alpha} \leftrightarrow h^{2\ell}$$

- Variance Si f est bornée est $\int K^2(u)du < \infty$,

$$\text{Var}(\hat{f}_h(x)) \leq \frac{\kappa_1}{nh}$$

avec $\kappa_1 = \|f\|_{\infty} \int K^2(u)du$.

MSE des estimateurs par noyaux : cas général

Notons

$$\mathcal{C}^{\ell *} = \{f \in \mathcal{C}^{\ell}, \|f^{(\ell)}\|_{\infty} < \infty.$$

- Biais Si $f \in \mathcal{C}^{\ell *}$ et K est suffisamment régulier (K "noyau d'ordre ℓ") alors

$$(f_h(x_0) - f(x_0))^2 \leq \kappa_0 h^{2\ell}$$

$\hookrightarrow$ Remarque : Analogie avec le biais des estimateurs par projection :

$$D^{-2\alpha} \leftrightarrow h^{2\ell}$$

- Variance Si f est bornée est $\int K^2(u)du < \infty$,

$$\text{Var}(\hat{f}_h(x)) \leq \frac{\kappa_1}{nh}$$

avec $\kappa_1 = \|f\|_{\infty} \int K^2(u)du$.

MSE des estimateurs par noyaux : cas général

Notons

$$\mathcal{C}^{\ell *} = \{f \in \mathcal{C}^{\ell}, \|f^{(\ell)}\|_{\infty} < \infty.$$

- Biais Si $f \in \mathcal{C}^{\ell *}$ et K est suffisamment régulier (K "noyau d'ordre ℓ") alors

$$(f_h(x_0) - f(x_0))^2 \leq \kappa_0 h^{2\ell}$$

$\hookrightarrow$ Remarque : Analogie avec le biais des estimateurs par projection :

$$D^{-2\alpha} \leftrightarrow h^{2\ell}$$

- Variance Si f est bornée est $\int K^2(u)du < \infty$,

$$\text{Var}(\hat{f}_h(x)) \leq \frac{\kappa_1}{nh}$$

avec $\kappa_1 = \|f\|_{\infty} \int K^2(u)du$.

Majoration de la variance : calculs

Si f est bornée est $\int K^2(u)du < \infty$,

$$\begin{aligned}
\mathsf{Var}(\hat{f}_h(x)) &= \frac{1}{h^2}\mathsf{Var}\left[\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)\right] \\
&= \frac{1}{nh^2}\mathsf{Var}\left[K\left(\frac{X_i - x_0}{h}\right)\right] \\
&\leq \frac{1}{nh^2}\mathbb{E}\left[K^2\left(\frac{X_i - x_0}{h}\right)\right] \\
&= \frac{1}{nh^2}\int K^2\left(\frac{x - x_0}{h}\right) f(x)\,dx \\
&= \frac{1}{nh}\int K^2(u) f(x_0 + uh)\,du \leq \frac{\|f\|_\infty \int K^2(u)\,du}{nh}
\end{aligned}$$

Majoration du biais : calculs (1)

- Pour tout $h > 0$, soit

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$$

alors

$$\hat{f}_h(x_0) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x_0).$$

- De plus,

$$\begin{aligned} f_h(x_0) \quad &:= \quad \mathbb{E}[\hat{f}_h(x_0)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} K_h\left(X_i - x_0\right)\right] = \mathbb{E}\left[K_h\left(X_1 - x_0\right)\right] \\ &= \quad \int K_h(x - x_0) f(x)\, dx = K_h * f(x_0) \end{aligned}$$

où $*$ désigne le produit de convolution.

- Enfin, par changement de variable :

$$f_h(x_0) = \int K(u) f(x_0 + hu)\, du$$

Majoration du biais : calculs (2)

$$f_h(x_0)-f(x_0) = \int K(u)f(x_0+hu)\,du - f(x_0) = \int K(u)[f(x_0+hu)-f(x_0)]\,du$$

- Si $f \in \mathcal{C}^{1*}$ et $\int |uK(u)|\,du < \infty$,

$$f(x_0+hu) = f(x_0) + huf'(x_1) \quad \text{avec} \quad x_1 \in [x_0, x_0+hu]$$

$$\Rightarrow |f_h(x_0) - f(x_0)| = \left|\int huK(u)f(x_1)\,du\right| \leq \|f'\|_\infty h \int |uK(u)|\,du = C_0 h$$

- Si $f \in \mathcal{C}^{2*}$, $\int uK(u)\,du = 0$ et $\int u^2K(u)\,du < \infty$

$$f(x_0+hu) = f(x_0)+huf'(x_0)+\frac{h^2u^2}{2}f''(x_2) \quad \text{avec} \quad x_2 \in [x_0, x_0+hu]$$

$$\begin{aligned}\Rightarrow |f_h(x_0) - f(x_0)| &= \left|hf(x_0)\int uK(u)\,du + \frac{h^2}{2}\int u^2K(u)f''(x_2)\,du\right| \\ &\leq \frac{\|f''\|_\infty h^2}{2}\int u^2|K(u)|\,du = C_1 h^2\end{aligned}$$

Majoration du biais : calculs (3)

- Généralisation : Si $f \in \mathcal{C}^{\ell *}$ et K est un noyau d'ordre ℓ i.e.

$$\begin{cases} \int u^j K(u) du = 0 \quad \forall j = 1, \dots, \ell - 1 \\ \int |u^\ell K(u)| du < \infty \end{cases}$$

alors

$$\Big(\mathbb{E}[\hat{f}_h(x_0)] - f(x_0) \Big)^2 \leq \kappa_0 h^{2\ell}$$

- Pour tout $\ell \in \mathbb{N}^*$, on sait construire des noyaux d'ordre ℓ à partir des polynomes de Legendre (base de polynomes orthonormée pour le produit scalaire L^2).

Compromis biais-variance

- Soit $f \in \mathcal{C}^{\ell *}$ et K un noyau d'ordre ℓ, alors

$$\mathbb{E}\left[(\hat{f}_h - f)^2(x_0)\right] \leq \left\{\kappa_0 h^{2\ell} + \frac{\kappa_1}{nh}\right\}$$

avec

$$\begin{cases} \text{Biais :} & \kappa_0 h^{2\ell} \nearrow \text{ when } h \nearrow \\ \text{Variance :} & \kappa_1/nh \searrow \text{ when } h \nearrow \end{cases} \tag{3}$$

- Afin que le biais et la variance tendent vers 0, on doit nécessairement choisir une fenêtre h telle que $nh \to +\infty$ et $h \to 0$.
- La fenêtre optimale

$$h^{**} = \arg\min\left\{\kappa_0 h^{2\ell} + \frac{\kappa_1}{nh}\right\} = \kappa n^{-1/(2\ell+1)}$$

dépend de la régularité inconnue de la densité f.

Compromis biais-variance

- Soit $f \in \mathcal{C}^{\ell *}$ et K un noyau d'ordre ℓ, alors

$$\mathbb{E}\left[(\hat{f}_h - f)^2(x_0)\right] \leq \left\{\kappa_0 h^{2\ell} + \frac{\kappa_1}{nh}\right\}$$

avec

$$\begin{cases} \text{Biais :} & \kappa_0 h^{2\ell} \nearrow \text{ when } h \nearrow \\ \text{Variance :} & \kappa_1/nh \searrow \text{ when } h \nearrow \end{cases} \tag{3}$$

- Afin que le biais et la variance tendent vers 0, on doit nécessairement choisir une fenêtre h telle que $nh \to +\infty$ et $h \to 0$.
- La fenêtre optimale

$$h^{**} = \arg\min\left\{\kappa_0 h^{2\ell} + \frac{\kappa_1}{nh}\right\} = \kappa n^{-1/(2\ell+1)}$$

dépend de la régularité inconnue de la densité f.

Compromis biais-variance

- Soit $f \in \mathcal{C}^{\ell *}$ et K un noyau d'ordre ℓ, alors

$$\mathbb{E}\left[(\hat{f}_h - f)^2(x_0)\right] \leq \left\{\kappa_0 h^{2\ell} + \frac{\kappa_1}{nh}\right\}$$

avec

$$\begin{cases} \text{Biais :} & \kappa_0 h^{2\ell} \nearrow \text{ when } h \nearrow \\ \text{Variance :} & \kappa_1/nh \searrow \text{ when } h \nearrow \end{cases} \tag{3}$$

- Afin que le biais et la variance tendent vers 0, on doit nécessairement choisir une fenêtre h telle que $nh \to +\infty$ et $h \to 0$.
- La fenêtre optimale

$$h^{**} = \arg\min\left\{\kappa_0 h^{2\ell} + \frac{\kappa_1}{nh}\right\} = \kappa n^{-1/(2\ell+1)}$$

dépend de la régularité inconnue de la densité f.

Sélection automatique de fenêtre pour le MSE

- La meilleure fenêtre pour le risque MSE est :

$$h^* = \arg\min \left\{ (f_h - f)^2(x_0) + \frac{\kappa_1}{nh} \right\}$$

↪ h^* dépend du point x_0.

- Comme pour l'estimation par projection, il existe une procédure de sélection automatique de fenêtre, dite méthode de Lepski. L'implémentation est plus complexe que la sélection de modèle.

Sélection automatique de fenêtre pour le MSE

- La meilleure fenêtre pour le risque MSE est :

$$h^* = \arg\min \left\{ (f_h - f)^2(x_0) + \frac{\kappa_1}{nh} \right\}$$

$\hookrightarrow$ h^* dépend du point x_0.

- Comme pour l'estimation par projection, il existe une procédure de sélection automatique de fenêtre, dite méthode de Lepski. L'implémentation est plus complexe que la sélection de modèle.

Sélection automatique de fenêtre pour le MSE

- La meilleure fenêtre pour le risque MSE est :

$$h^* = \arg\min \left\{ (f_h - f)^2(x_0) + \frac{\kappa_1}{nh} \right\}$$

↪ h^* dépend du point x_0.

- Comme pour l'estimation par projection, il existe une procédure de sélection automatique de fenêtre, dite méthode de Lepski. L'implémentation est plus complexe que la sélection de modèle.

Risque intégré (MISE) des estimateurs par noyaux

$$\text{MISE} = \|f_h - f\|^2 + \mathbb{E}\|\hat{f}_h - f_h\|_2^2$$

Résultat

Si $f \in \mathcal{C}^{\ell *}$ et K un noyau d'ordre ℓ, on montre que :

$$\begin{cases} \|f_h - f\|^2 & \leq & \kappa_0 h^{2\ell} \\ \mathbb{E}\left[\|\hat{f}_h - f_h\|_2^2\right] & \leq & \kappa_1 \frac{1}{nh} R(K) \quad \text{avec} \quad R(K) = \int K^2 \end{cases}$$

Risque intégré (MISE) des estimateurs par noyaux

$$\text{MISE} = \|f_h - f\|^2 + \mathbb{E}\|\hat{f}_h - f_h\|_2^2$$

Résultat

Si $f \in \mathcal{C}^{\ell *}$ et K un noyau d'ordre ℓ, on montre que :

$$\begin{cases} \|f_h - f\|^2 & \leq & \kappa_0 h^{2\ell} \\ \mathbb{E}\left[\|\hat{f}_h - f_h\|_2^2\right] & \leq & \kappa_1 \frac{1}{nh} R(K) \quad \text{avec} \quad R(K) = \int K^2 \end{cases}$$

MISE asymptotique : autre méthode de choix de h

- Si $f \in \mathcal{C}^2$, par des développements de Taylor,

$$\begin{cases} \|f_h - f\|^2 & = & \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'') + o(h^2) \\ \mathbb{E}\left[\|\hat{f}_h - f_h\|_2^2\right] & = & \frac{1}{nh}R(K) + o\left(\frac{1}{nh}\right) \end{cases}$$

- Ainsi, le MISE est asymptotiquement équivalent à :

$$AMISE = \frac{1}{nh}R(K) + \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'')$$

sous la condition $nh \to +\infty$ et $h \to 0$.

- La fenêtre optimal pour l'AMISE est

$$h_{AMISE} = \left[\frac{R(K)}{n\left(\int u^2 K^2(u)\,du\right)^2 R(f'')}\right]^{1/5}$$

$\hookrightarrow$ h_{AMISE} dépend de f (inconnue !)

MISE asymptotique : autre méthode de choix de h

- Si $f \in \mathcal{C}^2$, par des développements de Taylor,

$$\begin{cases} \|f_h - f\|^2 & = & \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'') + \mathsf{o}(h^2) \\ \mathbb{E}\left[\|\hat{f}_h - f_h\|_2^2\right] & = & \frac{1}{nh} R(K) + \mathsf{o}\left(\frac{1}{nh}\right) \end{cases}$$

- Ainsi, le MISE est asymptotiquement équivalent à :

$$AMISE = \frac{1}{nh} R(K) + \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'')$$

sous la condition $nh \to +\infty$ et $h \to 0$.

- La fenêtre optimal pour l'AMISE est

$$h_{AMISE} = \left[\frac{R(K)}{n\left(\int u^2 K^2(u)\,du\right)^2 R(f'')}\right]^{1/5}$$

$\hookrightarrow$ h_{AMISE} dépend de f (inconnue !)

MISE asymptotique : autre méthode de choix de h

- Si $f \in \mathcal{C}^2$, par des développements de Taylor,

$$\begin{cases} \|f_h - f\|^2 & = & \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'') + \mathsf{o}(h^2) \\ \mathbb{E}\left[\|\hat{f}_h - f_h\|_2^2\right] & = & \frac{1}{nh} R(K) + \mathsf{o}\left(\frac{1}{nh}\right) \end{cases}$$

- Ainsi, le MISE est asymptotiquement équivalent à :

$$AMISE = \frac{1}{nh} R(K) + \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'')$$

sous la condition $nh \to +\infty$ et $h \to 0$.

- La fenêtre optimal pour l'AMISE est

$$h_{AMISE} = \left[\frac{R(K)}{n\left(\int u^2 K^2(u)\,du\right)^2 R(f'')}\right]^{1/5}$$

$\hookrightarrow$ h_{AMISE} dépend de f (inconnue !)

MISE asymptotique : autre méthode de choix de h

- Si $f \in \mathcal{C}^2$, par des développements de Taylor,

$$\begin{cases} \|f_h - f\|^2 & = & \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'') + \mathsf{o}(h^2) \\ \mathbb{E}\left[\|\hat{f}_h - f_h\|_2^2\right] & = & \frac{1}{nh} R(K) + \mathsf{o}\left(\frac{1}{nh}\right) \end{cases}$$

- Ainsi, le MISE est asymptotiquement équivalent à :

$$AMISE = \frac{1}{nh} R(K) + \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'')$$

sous la condition $nh \to +\infty$ et $h \to 0$.

- La fenêtre optimal pour l'AMISE est

$$h_{AMISE} = \left[\frac{R(K)}{n\left(\int u^2 K^2(u)\,du\right)^2 R(f'')}\right]^{1/5}$$

$\hookrightarrow$ h_{AMISE} dépend de f (inconnue !)

MISE asymptotique : autre méthode de choix de h

- Si $f \in \mathcal{C}^2$, par des développements de Taylor,

$$\begin{cases} \|f_h - f\|^2 & = & \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'') + \mathsf{o}(h^2) \\ \mathbb{E}\left[\|\hat{f}_h - f_h\|_2^2\right] & = & \frac{1}{nh} R(K) + \mathsf{o}\left(\frac{1}{nh}\right) \end{cases}$$

- Ainsi, le MISE est asymptotiquement équivalent à :

$$AMISE = \frac{1}{nh} R(K) + \frac{h^4}{4}\left(\int u^2 K^2(u)\,du\right)^2 R(f'')$$

sous la condition $nh \to +\infty$ et $h \to 0$.

- La fenêtre optimal pour l'AMISE est

$$h_{AMISE} = \left[\frac{R(K)}{n\left(\int u^2 K^2(u)\,du\right)^2 R(f'')}\right]^{1/5}$$

$\hookrightarrow$ h_{AMISE} dépend de f (inconnue !)

Sélection de fenêtre globale basé sur l'AMISE

$$h_{AMISE} = \left[\frac{R(K)}{n \left(\int u^2 K^2(u) du \right)^2 R(f'')} \right]^{1/5}$$

Des critères de sélection de fenêtre globale sont construits en calculant $R(f'')$ sur une classe de fonction paramétriques, où le paramètre peut être estimé à partir des données.

- Règle de Silverman : h minimisant le AMISE pour une densité gaussienne (par défaut dans `density`).

$$\hat{h}_n = 0.9 \min(\text{sd}(X), \text{IQR}(X))$$

- Règle de Scott : Pour une variance fixée, la densité f telle que $R(f'')$ soit minimale est $\beta(4,4)$. La fenêtre optimale correspondante est :

$$\hat{h}_n = 1.144 \text{sd}(X) n^{-1/5}$$

↪ "Oversmoothing" : la fenêtre sélectionnée est un majorant de la fenêtre optimale.

Sélection de fenêtre globale basé sur l'AMISE

$$h_{AMISE} = \left[\frac{R(K)}{n \left(\int u^2 K^2(u) du \right)^2 R(f'')} \right]^{1/5}$$

Des critères de sélection de fenêtre globale sont construits en calculant $R(f'')$ sur une classe de fonction paramétriques, où le paramètre peut être estimé à partir des données.

- Règle de Silverman : h minimisant le AMISE pour une densité gaussienne (par défaut dans `density`).

$$\hat{h}_n = 0.9 \min(\text{sd}(X), \text{IQR}(X))$$

- Règle de Scott : Pour une variance fixée, la densité f telle que $R(f'')$ soit minimale est $\beta(4,4)$. La fenêtre optimale correspondante est :

$$\hat{h}_n = 1.144 \text{sd}(X) n^{-1/5}$$

↪ "Oversmoothing" : la fenêtre sélectionnée est un majorant de la fenêtre optimale.

Sélection de fenêtre globale basé sur l'AMISE

$$h_{AMISE} = \left[\frac{R(K)}{n \left(\int u^2 K^2(u) du \right)^2 R(f'')} \right]^{1/5}$$

Des critères de sélection de fenêtre globale sont construits en calculant $R(f'')$ sur une classe de fonction paramétriques, où le paramètre peut être estimé à partir des données.

- Règle de Silverman : h minimisant le AMISE pour une densité gaussienne (par défaut dans `density`).

$$\hat{h}_n = 0.9 \min(\text{sd}(X), \text{IQR}(X))$$

- Règle de Scott : Pour une variance fixée, la densité f telle que $R(f'')$ soit minimale est $\mathcal{B}(4,4)$. La fenêtre optimale correspondante est :

$$\hat{h}_n = 1.144 \text{sd}(X) n^{-1/5}$$

↪ "Oversmoothing" : la fenêtre sélectionnée est un majorant de la fenêtre optimale.

Sélection de fenêtre globale basé sur l'AMISE

$$h_{AMISE} = \left[\frac{R(K)}{n \left(\int u^2 K^2(u) du \right)^2 R(f'')} \right]^{1/5}$$

Des critères de sélection de fenêtre globale sont construits en calculant $R(f'')$ sur une classe de fonction paramétriques, où le paramètre peut être estimé à partir des données.

- Règle de Silverman : h minimisant le AMISE pour une densité gaussienne (par défaut dans `density`).

$$\hat{h}_n = 0.9 \min(\text{sd}(X), \text{IQR}(X))$$

- Règle de Scott : Pour une variance fixée, la densité f telle que $R(f'')$ soit minimale est $\mathcal{B}(4,4)$. La fenêtre optimale correspondante est :

$$\hat{h}_n = 1.144 \text{sd}(X) n^{-1/5}$$

↪ "Oversmoothing" : la fenêtre sélectionnée est un majorant de la fenêtre optimale.

Conclusion sur le choix de fenêtres

- Lepski : sélection de fenêtre en chaque point :
  - Adaptation à la régularité locale
  - Résultats théoriques solides (inégalités oracles et adaptativité)
  - Complexe à implémenter
- Règles basées sur le AMISE
  - Simples à implémenter
  - Pas d'adaptation locale
  - Heuristique basée sur une classe de fonctions paramétriques : pas de résultats théoriques non paramétriques
- Nombreuses autres méthodes dans la littérature

Conclusion sur estimation de densité non-paramétrique

- On construit une collection d'estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$:

$$\begin{cases} \text{Estimateurs à noyaux :} & \lambda = h \in \mathbb{R}^+ \\ \text{Estimateurs par projection :} & \lambda = D \in 1 : D_{max} \end{cases}$$

- Pour chaque estimateur, le risque est décomposé en biais variance, qui varient en sens opposé quand λ varie.
- Le paramètre λ optimal appelé oracle dépend de la régularité inconnue de f.
- $\hat{\lambda}$ est sélectionné par minimisation d'un estimateur de la somme biais-variance.
  - Méthodes comportant des résultats théoriques (inégalité oracles, adaptativité sur des classes de régularité)
  - Règles empiriques ("rules of thumb") : plus simples à implémenter mais peu de résultats théoriques.

Conclusion sur estimation de densité non-paramétrique

- On construit une collection d'estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$:

$$\begin{cases} \text{Estimateurs à noyaux :} & \lambda = h \in \mathbb{R}^+ \\ \text{Estimateurs par projection :} & \lambda = D \in 1 : D_{max} \end{cases}$$

- Pour chaque estimateur, le risque est décomposé en biais variance, qui varient en sens opposé quand λ varie.
- Le paramètre λ optimal appelé oracle dépend de la régularité inconnue de f.
- $\hat{\lambda}$ est sélectionné par minimisation d'un estimateur de la somme biais-variance.
  - Méthodes comportant des résultats théoriques (inégalité oracles, adaptativité sur des classes de régularité)
  - Règles empiriques ("rules of thumb") : plus simples à implémenter mais peu de résultats théoriques.

Conclusion sur estimation de densité non-paramétrique

- On construit une collection d'estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$:

$$\begin{cases} \text{Estimateurs à noyaux :} & \lambda = h \in \mathbb{R}^+ \\ \text{Estimateurs par projection :} & \lambda = D \in 1 : D_{max} \end{cases}$$

- Pour chaque estimateur, le risque est décomposé en biais variance, qui varient en sens opposé quand λ varie.
- Le paramètre λ optimal appelé oracle dépend de la régularité inconnue de f.
- $\hat{\lambda}$ est sélectionné par minimisation d'un estimateur de la somme biais-variance.
  - Méthodes comportant des résultats théoriques (inégalité oracles, adaptativité sur des classes de régularité)
  - Règles empiriques ("rules of thumb") : plus simples à implémenter mais peu de résultats théoriques.

Conclusion sur estimation de densité non-paramétrique

- On construit une collection d'estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$:

$$\begin{cases} \text{Estimateurs à noyaux :} & \lambda = h \in \mathbb{R}^+ \\ \text{Estimateurs par projection :} & \lambda = D \in 1 : D_{max} \end{cases}$$

- Pour chaque estimateur, le risque est décomposé en biais variance, qui varient en sens opposé quand λ varie.
- Le paramètre λ optimal appelé oracle dépend de la régularité inconnue de f.
- $\hat{\lambda}$ est sélectionné par minimisation d'un estimateur de la somme biais-variance.
  - Méthodes comportant des résultats théoriques (inégalité oracles, adaptativité sur des classes de régularité)
  - Règles empiriques ("rules of thumb") : plus simples à implémenter mais peu de résultats théoriques.

Retour sur la sélection de modèle pour les estimateurs par projection

- On calcule une collection d'estimateurs $\{\hat{f}_D\}_{D=1,\dots D_{max}}$, et le modèle optimal pour le MISE est :

$$\arg\min_D \mathcal{R}(\hat{f}_D, f) = \arg\min_D \{\|f - f_D\|^2 + \mathbb{E}[\|\hat{f}_D - f_D\|^2]\}$$

- Le terme de variance $\mathbb{E}[\|\hat{f}_D - f_D\|^2]$ est majoré par KD/n.
- On sélectionne un modèle qui satisfait :

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{\|f - f_D\|^2 + \frac{KD}{n}\} + \text{reste}$$

- Commentaires :
  - $\{\ \}$ n'est qu'un majorant du risque : il faut prouver que $\{\ \}$ est du même ordre que $\mathcal{R}(\hat{f}_D, f)$
  - L'approche oracle ne prouve l'optimalité que sur une collection d'estimateurs donnés (ex : histogrammes)
    $\hookrightarrow$ nécéssité d'une définition de l'optimalité plus générale.

Retour sur la sélection de modèle pour les estimateurs par projection

- On calcule une collection d'estimateurs $\{\hat{f}_D\}_{D=1,\dots D_{max}}$, et le modèle optimal pour le MISE est :

$$\arg\min_D \mathcal{R}(\hat{f}_D, f) = \arg\min_D \{\|f - f_D\|^2 + \mathbb{E}[\|\hat{f}_D - f_D\|^2]\}$$

- Le terme de variance $\mathbb{E}[\|\hat{f}_D - f_D\|^2]$ est majoré par KD/n.
- On sélectionne un modèle qui satisfait :

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{\|f - f_D\|^2 + \frac{KD}{n}\} + \text{reste}$$

- Commentaires :
  - $\{\ \}$ n'est qu'un majorant du risque : il faut prouver que $\{\ \}$ est du même ordre que $\mathcal{R}(\hat{f}_D, f)$
  - L'approche oracle ne prouve l'optimalité que sur une collection d'estimateurs donnés (ex : histogrammes)
    $\hookrightarrow$ nécéssité d'une définition de l'optimalité plus générale.

Retour sur la sélection de modèle pour les estimateurs par projection

- On calcule une collection d'estimateurs $\{\hat{f}_D\}_{D=1,\dots D_{max}}$, et le modèle optimal pour le MISE est :

$$\arg\min_D \mathcal{R}(\hat{f}_D, f) = \arg\min_D \{\|f - f_D\|^2 + \mathbb{E}[\|\hat{f}_D - f_D\|^2]\}$$

- Le terme de variance $\mathbb{E}[\|\hat{f}_D - f_D\|^2]$ est majoré par KD/n.
- On sélectionne un modèle qui satisfait :

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{\|f - f_D\|^2 + \frac{KD}{n}\} + \text{reste}$$

- Commentaires :
  - $\{\ \}$ n'est qu'un majorant du risque : il faut prouver que $\{\ \}$ est du même ordre que $\mathcal{R}(\hat{f}_D, f)$
  - L'approche oracle ne prouve l'optimalité que sur une collection d'estimateurs donnés (ex : histogrammes)
    $\hookrightarrow$ nécéssité d'une définition de l'optimalité plus générale.

Retour sur la sélection de modèle pour les estimateurs par projection

- On calcule une collection d'estimateurs $\{\hat{f}_D\}_{D=1,\dots D_{max}}$, et le modèle optimal pour le MISE est :

$$\arg\min_D \mathcal{R}(\hat{f}_D, f) = \arg\min_D \{\|f - f_D\|^2 + \mathbb{E}[\|\hat{f}_D - f_D\|^2]\}$$

- Le terme de variance $\mathbb{E}[\|\hat{f}_D - f_D\|^2]$ est majoré par KD/n.
- On sélectionne un modèle qui satisfait :

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{\|f - f_D\|^2 + \frac{KD}{n}\} + \text{reste}$$

- Commentaires :
  - $\{\ \}$ n'est qu'un majorant du risque : il faut prouver que $\{\ \}$ est du même ordre que $\mathcal{R}(\hat{f}_D, f)$
  - L'approche oracle ne prouve l'optimalité que sur une collection d'estimateurs donnés (ex : histogrammes)
    $\hookrightarrow$ nécéssité d'une définition de l'optimalité plus générale.

Retour sur la sélection de modèle pour les estimateurs par projection

- On calcule une collection d'estimateurs $\{\hat{f}_D\}_{D=1,\dots D_{max}}$, et le modèle optimal pour le MISE est :

$$\arg\min_D \mathcal{R}(\hat{f}_D, f) = \arg\min_D \{\|f - f_D\|^2 + \mathbb{E}[\|\hat{f}_D - f_D\|^2]\}$$

- Le terme de variance $\mathbb{E}[\|\hat{f}_D - f_D\|^2]$ est majoré par KD/n.
- On sélectionne un modèle qui satisfait :

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{\|f - f_D\|^2 + \frac{KD}{n}\} + \text{reste}$$

- Commentaires :
  - $\{\ \}$ n'est qu'un majorant du risque : il faut prouver que $\{\ \}$ est du même ordre que $\mathcal{R}(\hat{f}_D, f)$
  - L'approche oracle ne prouve l'optimalité que sur une collection d'estimateurs donnés (ex : histogrammes)
    $\hookrightarrow$ nécéssité d'une définition de l'optimalité plus générale.

Retour sur la sélection de modèle pour les estimateurs par projection

- On calcule une collection d'estimateurs $\{\hat{f}_D\}_{D=1,\dots D_{max}}$, et le modèle optimal pour le MISE est :

$$\arg\min_D \mathcal{R}(\hat{f}_D, f) = \arg\min_D \{\|f - f_D\|^2 + \mathbb{E}[\|\hat{f}_D - f_D\|^2]\}$$

- Le terme de variance $\mathbb{E}[\|\hat{f}_D - f_D\|^2]$ est majoré par KD/n.
- On sélectionne un modèle qui satisfait :

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{\|f - f_D\|^2 + \frac{KD}{n}\} + \text{reste}$$

- Commentaires :
  - $\{\ \}$ n'est qu'un majorant du risque : il faut prouver que $\{\ \}$ est du même ordre que $\mathcal{R}(\hat{f}_D, f)$
  - L'approche oracle ne prouve l'optimalité que sur une collection d'estimateurs donnés (ex : histogrammes)
    $\hookrightarrow$ nécéssité d'une définition de l'optimalité plus générale.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f\,[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$a r_n \leq \inf_{\hat{f}_n} \sup_{f \in \mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq A r_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

↪ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f\,[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$a r_n \leq \inf_{\hat{f}_n} \sup_{f \in \mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq A r_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

↪ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f\,[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$a r_n \leq \inf_{\hat{f}_n} \sup_{f \in \mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq A r_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

↪ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f\,[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$a r_n \leq \inf_{\hat{f}_n} \sup_{f \in \mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq A r_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

↪ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$ar_n \leq \inf_{\hat{f}_n} \sup_{f\in\mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq Ar_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

↪ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$ar_n \leq \inf_{\hat{f}_n} \sup_{f\in\mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq Ar_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

↪ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$ar_n \leq \inf_{\hat{f}_n} \sup_{f\in\mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq Ar_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

$\hookrightarrow$ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

Approche minimax en estimation de densité

- Soit une procédure de tirage d'un échantillon de taille n : $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{F}$.
- Soit f une fonction qui dépend de la loi $\mathcal{F}$: f densité de X_i
- Soit $\mathcal{R}$ une fonction de risque : MISE
- Soit $\mathcal{S}$ une classe de régularité pour f. Exple :

$$\mathcal{S}^\alpha = \{f[0,1] \to \mathbb{R}, \int f = 1, f \in \mathcal{H}(\alpha, L)\} \quad \text{avec} \quad 0 < \alpha \leq 1$$

- On dit que la suite $(r_n)_{n\in\mathbb{N}}$ est la vitesse minimax sur l'espace $\mathcal{S}^\alpha$ pour le risque $\mathcal{R}_n$ si il existe $a, A > 0$ tels que :

$$ar_n \leq \inf_{\hat{f}_n} \sup_{f\in\mathcal{S}^\alpha} \mathcal{R}(\hat{f}_n, f) \leq Ar_n$$

où l'inf est pris sur tous les estimateurs possibles de f.

↪ r_n est la vitesse de convergence du meilleur estimateur possible, pour la pire fonction de $\mathcal{S}^\alpha$.

- (r_n) est définie à constante près.

- Proposition : En estimation de densité et pour le MISE :
  - la vitesse minimax sur les espaces de Holder $\mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
  - la vitesse minimax sur les espaces de Sobolev $\mathcal{S}(\alpha, L)$ avec $\alpha > 0$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
- Soit $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour les histogrammes réguliers. Si $f \in \mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$, alors
$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{C_0 D^{-2\alpha} + \frac{KD}{n}\} = C_1 n^{-2\alpha/(2\alpha+1)}$$

↪ On dit que l'estimateur $\hat{f}_{\hat{D}}$ est minimax sur les classes de Holder pour le MISE.

- De même, $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour la base trigonométrique. Si $f \in \mathcal{S}(\alpha, L)$, alors $\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C_1 n^{-2\alpha/(2\alpha+1)}$. Ainsi $\hat{f}_{\hat{D}}$ est minimax sur les classes de Sobolev.

- Proposition : En estimation de densité et pour le MISE :
  - la vitesse minimax sur les espaces de Holder $\mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
  - la vitesse minimax sur les espaces de Sobolev $\mathcal{S}(\alpha, L)$ avec $\alpha > 0$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
- Soit $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour les histogrammes réguliers. Si $f \in \mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$, alors
$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{C_0 D^{-2\alpha} + \frac{KD}{n}\} = C_1 n^{-2\alpha/(2\alpha+1)}$$

↪ On dit que l'estimateur $\hat{f}_{\hat{D}}$ est minimax sur les classes de Holder pour le MISE.

- De même, $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour la base trigonométrique. Si $f \in \mathcal{S}(\alpha, L)$, alors $\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C_1 n^{-2\alpha/(2\alpha+1)}$. Ainsi $\hat{f}_{\hat{D}}$ est minimax sur les classes de Sobolev.

- Proposition : En estimation de densité et pour le MISE :
  - la vitesse minimax sur les espaces de Holder $\mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
  - la vitesse minimax sur les espaces de Sobolev $\mathcal{S}(\alpha, L)$ avec $\alpha > 0$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
- Soit $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour les histogrammes réguliers. Si $f \in \mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$, alors
$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{C_0 D^{-2\alpha} + \frac{KD}{n}\} = C_1 n^{-2\alpha/(2\alpha+1)}$$

↪ On dit que l'estimateur $\hat{f}_{\hat{D}}$ est minimax sur les classes de Holder pour le MISE.

- De même, $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour la base trigonométrique. Si $f \in \mathcal{S}(\alpha, L)$, alors $\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C_1 n^{-2\alpha/(2\alpha+1)}$. Ainsi $\hat{f}_{\hat{D}}$ est minimax sur les classes de Sobolev.

- Proposition : En estimation de densité et pour le MISE :
  - la vitesse minimax sur les espaces de Holder $\mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
  - la vitesse minimax sur les espaces de Sobolev $\mathcal{S}(\alpha, L)$ avec $\alpha > 0$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
- Soit $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour les histogrammes réguliers. Si $f \in \mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$, alors

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{C_0 D^{-2\alpha} + \frac{KD}{n}\} = C_1 n^{-2\alpha/(2\alpha+1)}$$

↪ On dit que l'estimateur $\hat{f}_{\hat{D}}$ est minimax sur les classes de Holder pour le MISE.

- De même, $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour la base trigonométrique. Si $f \in \mathcal{S}(\alpha, L)$, alors $\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C_1 n^{-2\alpha/(2\alpha+1)}$. Ainsi $\hat{f}_{\hat{D}}$ est minimax sur les classes de Sobolev.

- Proposition : En estimation de densité et pour le MISE :
  - la vitesse minimax sur les espaces de Holder $\mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
  - la vitesse minimax sur les espaces de Sobolev $\mathcal{S}(\alpha, L)$ avec $\alpha > 0$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
- Soit $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour les histogrammes réguliers. Si $f \in \mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$, alors

$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{C_0 D^{-2\alpha} + \frac{KD}{n}\} = C_1 n^{-2\alpha/(2\alpha+1)}$$

$\hookrightarrow$ On dit que l'estimateur $\hat{f}_{\hat{D}}$ est minimax sur les classes de Holder pour le MISE.

- De même, $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour la base trigonométrique. Si $f \in \mathcal{S}(\alpha, L)$, alors $\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C_1 n^{-2\alpha/(2\alpha+1)}$. Ainsi $\hat{f}_{\hat{D}}$ est minimax sur les classes de Sobolev.

- Proposition : En estimation de densité et pour le MISE :
  - la vitesse minimax sur les espaces de Holder $\mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
  - la vitesse minimax sur les espaces de Sobolev $\mathcal{S}(\alpha, L)$ avec $\alpha > 0$ est
$$r_n = n^{-2\alpha/(2\alpha+1)}.$$
- Soit $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour les histogrammes réguliers. Si $f \in \mathcal{H}(\alpha, L)$ avec $0 < \alpha < 1$, alors

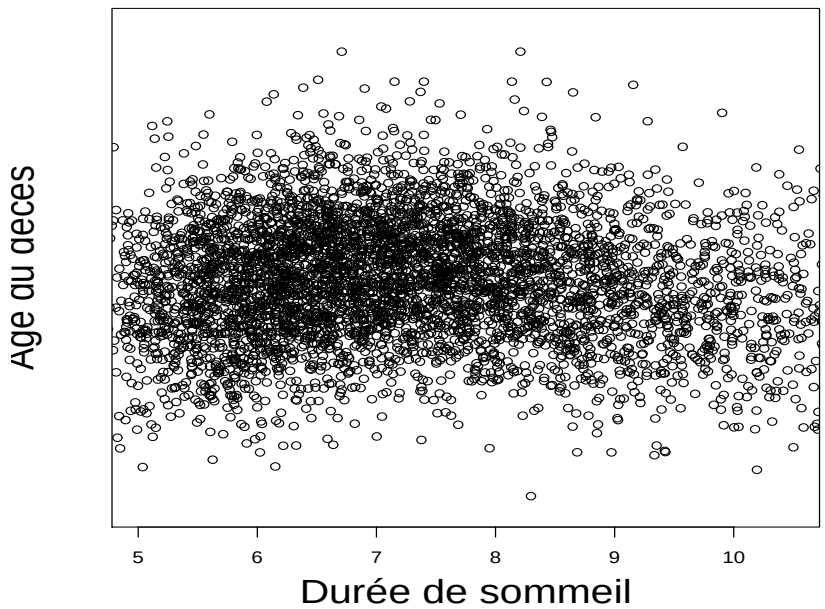
$$\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C \inf_D \{C_0 D^{-2\alpha} + \frac{KD}{n}\} = C_1 n^{-2\alpha/(2\alpha+1)}$$

$\hookrightarrow$ On dit que l'estimateur $\hat{f}_{\hat{D}}$ est minimax sur les classes de Holder pour le MISE.

- De même, $\hat{f}_{\hat{D}}$ l'estimateur de sélection de modèles pour la base trigonométrique. Si $f \in \mathcal{S}(\alpha, L)$, alors $\mathbb{E}[\|\hat{f}_{\hat{D}} - f\|^2] \leq C_1 n^{-2\alpha/(2\alpha+1)}$. Ainsi $\hat{f}_{\hat{D}}$ est minimax sur les classes de Sobolev.

Exemple 1

Pour $n = 2000$ personnes décédées, on regarde le lien entre l'âge décès X_i et le nombre d'heures moyen de sommeil à l'âge adulte Y_i.

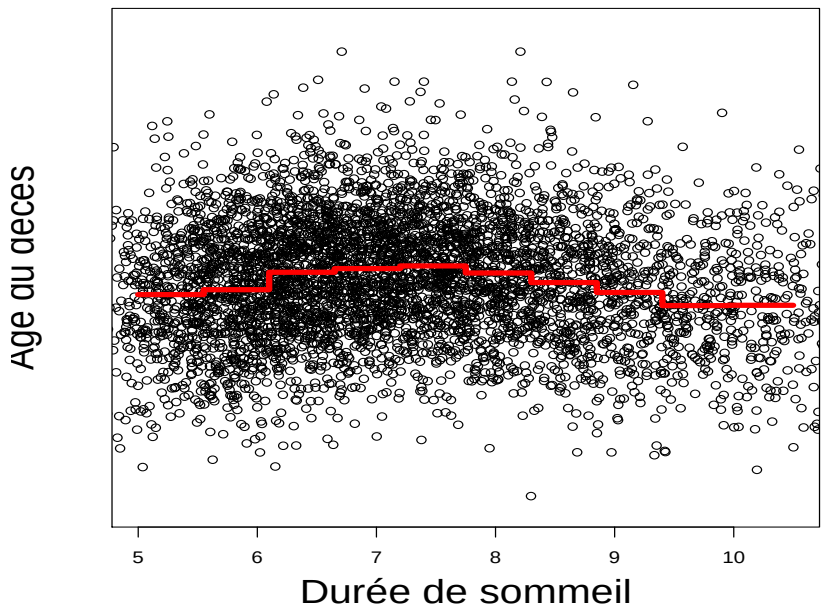


- Estimateur par histogramme : Sur chaque intervalle, on calcule la moyenne des Y_i.
- Constatation : L'espérance de vie est maximale pour les personnes dormant 7/8 heures :

$$\mathbb{E}[Y_i|X_i]$$

Exemple 1

Pour $n = 2000$ personnes décédées, on regarde le lien entre l'âge décès X_i et le nombre d'heures moyen de sommeil à l'âge adulte Y_i.

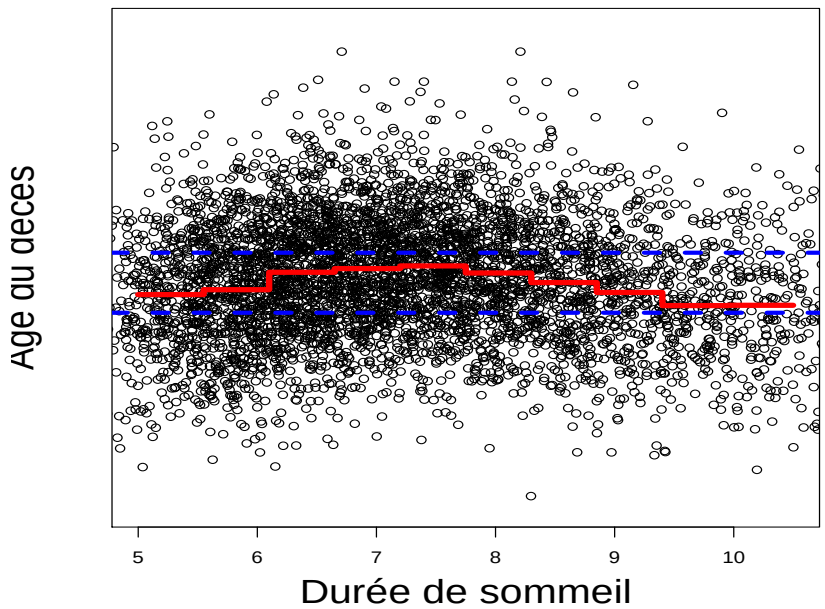


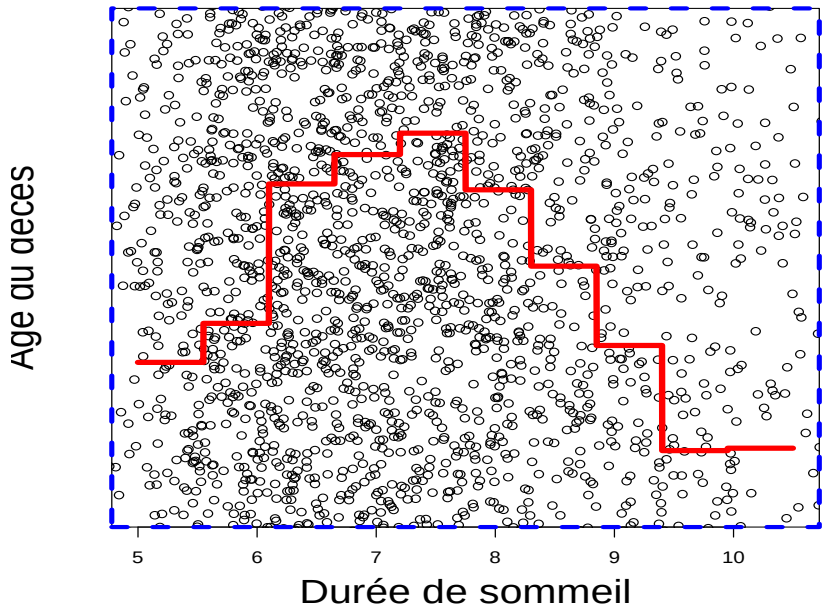
- Estimateur par histogramme : Sur chaque intervalle, on calcule la moyenne des Y_i.
- Constatation : L'espérance de vie est maximale pour les personnes dormant 7/8 heures :

$$\mathbb{E}[Y_i|X_i]$$

Exemple 1

Pour $n = 2000$ personnes décédées, on regarde le lien entre l'âge décès X_i et le nombre d'heures moyen de sommeil à l'âge adulte Y_i.



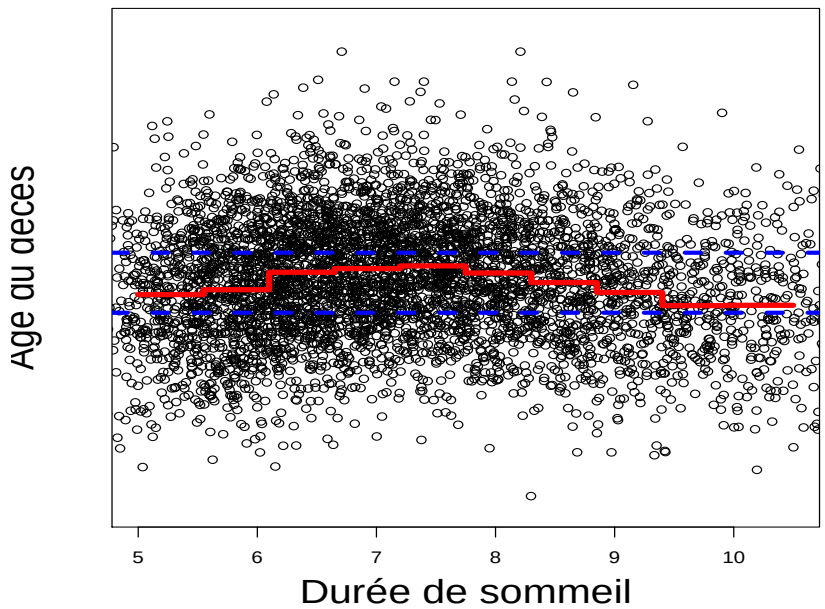


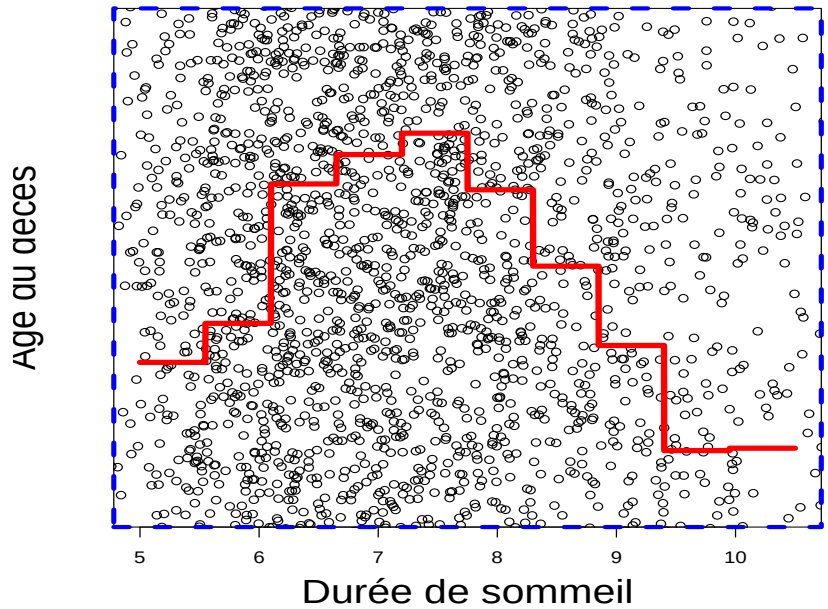
- Estimateur par histogramme : Sur chaque intervalle, on calcule la moyenne des Y_i.
- Constatation : L'espérance de vie est maximale pour les personnes dormant 7/8 heures :

$$\mathbb{E}[Y_i|X_i]$$

Exemple 1

Pour $n = 2000$ personnes décédées, on regarde le lien entre l'âge décès X_i et le nombre d'heures moyen de sommeil à l'âge adulte Y_i.



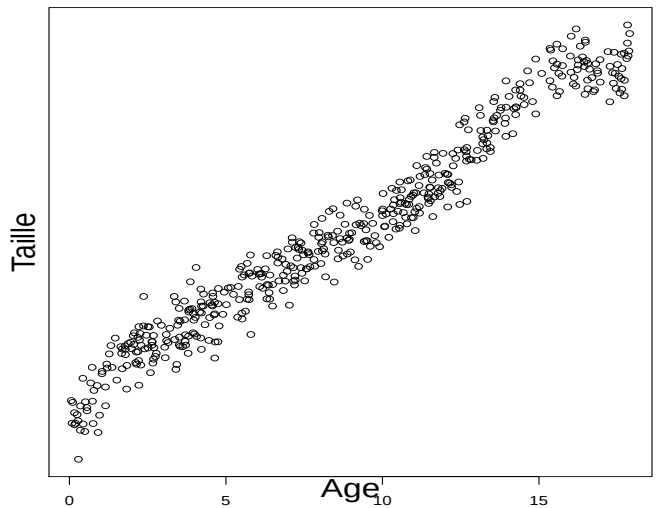


- Estimateur par histogramme : Sur chaque intervalle, on calcule la moyenne des Y_i.
- Constatation : L'espérance de vie est maximale pour les personnes dormant 7/8 heures :

$$\mathbb{E}[\, Y_i | X_i]$$

Exemple 2

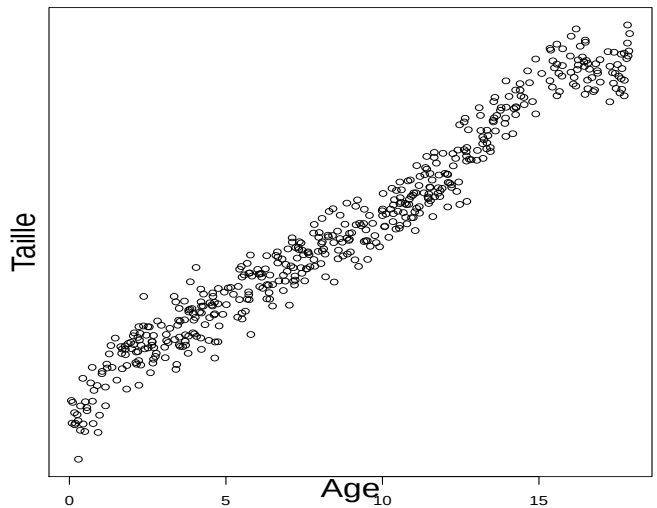
On mesure la taille Y_i et l'âge X_i de 300 filles dans un certain groupe ethnique.



- Constatations : La taille augmente particulièrement avant 2 ans et entre 12 et 15 ans, puis stagne au delà.
- Prédiction : Peut-on prédire la taille d'une fille de ce groupe ethnique pour un âge donné ?

Exemple 2

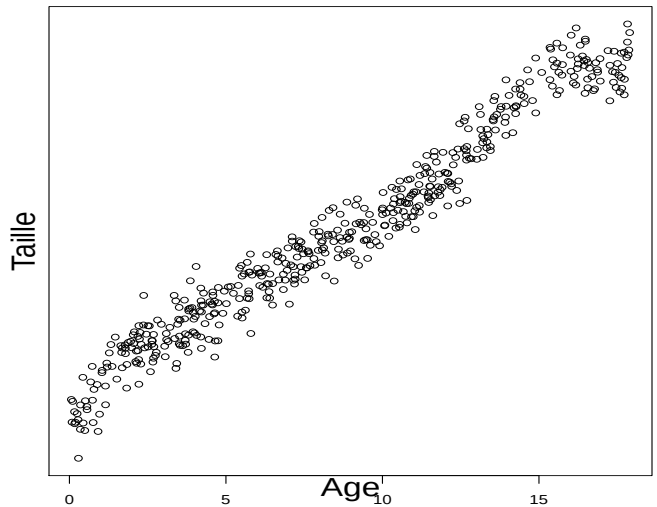
On mesure la taille Y_i et l'âge X_i de 300 filles dans un certain groupe ethnique.



- Constatations : La taille augmente particulièrement avant 2 ans et entre 12 et 15 ans, puis stagne au delà.
- Prédiction : Peut-on prédire la taille d'une fille de ce groupe ethnique pour un âge donné ?

Exemple 2

On mesure la taille Y_i et l'âge X_i de 300 filles dans un certain groupe ethnique.



- Constatations : La taille augmente particulièrement avant 2 ans et entre 12 et 15 ans, puis stagne au delà.
- Prédiction : Peut-on prédire la taille d'une fille de ce groupe ethnique pour un âge donné ?

Contexte de la régression

- Contexte général de la régression : Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. de couples de v.a.r. On appelle

$$r(x) = \mathbb{E}(Y|X = x)$$

la fonction de régression de Y sur X, et $\{\varepsilon_i\}_{i=1,\ldots,n}$ les erreurs définies par $\varepsilon_i = Y_i - r(X_i)$.

↪ **Exple :** Soit

$$\begin{cases} X \sim \mathcal{U}([0,1]) & \text{taille d'une tumeur cancéreuse} \\ Y|X \sim \mathcal{B}(X^2) & \text{succès d'un traitement donné (oui/non)} \end{cases}$$

alors $r(x) = \mathbb{E}[\mathcal{B}(x^2)] = x^2$.

- Contexte restreint : régression avec bruit additif. Dans une partir du cours, nous allons nous restreindre au modèle

$$Y_i = r(X_i) + \varepsilon_i \quad \text{avec} \quad \epsilon_i \perp X_i, \quad \mathbb{E}[\epsilon_i] = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2 < \infty$$

  - $r(x)$ correspond à la valeur moyenne de Y quand $X = x$
  - $\{\varepsilon_i\}$: variations individuelles.

Contexte de la régression

- Contexte général de la régression : Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. de couples de v.a.r. On appelle

$$r(x) = \mathbb{E}(Y|X = x)$$

la fonction de régression de Y sur X, et $\{\varepsilon_i\}_{i=1,\ldots,n}$ les erreurs définies par $\varepsilon_i = Y_i - r(X_i)$.
$\hookrightarrow$ **Exple :** Soit

$$\begin{cases} X \sim \mathcal{U}([0,1]) & \text{taille d'une tumeur cancéreuse} \\ Y|X \sim \mathcal{B}(X^2) & \text{succès d'un traitement donné (oui/non)} \end{cases}$$

alors $r(x) = \mathbb{E}[\mathcal{B}(x^2)] = x^2$.

- Contexte restreint : régression avec bruit additif. Dans une partir du cours, nous allons nous restreindre au modèle

$$Y_i = r(X_i) + \varepsilon_i \quad \text{avec} \quad \epsilon_i \perp X_i, \quad \mathbb{E}[\epsilon_i] = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2 < \infty$$

  - $r(x)$ correspond à la valeur moyenne de Y quand $X = x$
  - $\{\varepsilon_i\}$: variations individuelles.

Contexte de la régression

- Contexte général de la régression : Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. de couples de v.a.r. On appelle

$$r(x) = \mathbb{E}(Y|X = x)$$

la fonction de régression de Y sur X, et $\{\varepsilon_i\}_{i=1,\ldots,n}$ les erreurs définies par $\varepsilon_i = Y_i - r(X_i)$.
$\hookrightarrow$ **Exple :** Soit

$$\begin{cases} X \sim \mathcal{U}([0,1]) & \text{taille d'une tumeur cancéreuse} \\ Y|X \sim \mathcal{B}(X^2) & \text{succès d'un traitement donné (oui/non)} \end{cases}$$

alors $r(x) = \mathbb{E}[\mathcal{B}(x^2)] = x^2$.

- Contexte restreint : régression avec bruit additif. Dans une partir du cours, nous allons nous restreindre au modèle

$$Y_i = r(X_i) + \varepsilon_i \quad \text{avec} \quad \epsilon_i \perp X_i, \quad \mathbb{E}[\epsilon_i] = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2 < \infty$$

  - $r(x)$ correspond à la valeur moyenne de Y quand $X = x$
  - $\{\varepsilon_i\}$: variations individuelles.

Regression avec bruit additif : fix/random design

On appelle "dispositif expérimental" ou "design" les observations $(X_1, \dots, X_n)$. Il existe deux situations :

- Effets fixes ou fix-design : $X_1, \dots, X_n$ fixé par l'utilisateur

  Ex : X_i=dosage de médicament ingéré et Y_i=taux de la molécule active dans le sang 1h après ingestion : les dosages $(X_1, \dots, X_n)$ sont fixés par l'expérimentateur.

- Effets aléatoires ou random design : $X_1, \dots, X_n$ échantillon i.i.d. de densité f_X.

  Ex : X_i=salaire moyen et Y_i = âge du décès. On sélectionne aléatoirement n individus parmi la population.

$\hookrightarrow$ Dans ce cours, nous allons considérer la regression en random design, mais les méthodes d'estimation, et une partie des résultats, sont directement transposables au fix design. Néanmoins, les résultats en fix-design font souvent des hypothèses supplémentaires (exple $(X_1, \dots, X_n)$ équidistants).

Regression avec bruit additif : fix/random design

On appelle "dispositif expérimental" ou "design" les observations $(X_1, \dots, X_n)$. Il existe deux situations :

- Effets fixes ou fix-design : $X_1, \dots, X_n$ fixé par l'utilisateur

  Ex : X_i=dosage de médicament ingéré et Y_i=taux de la molécule active dans le sang 1h après ingestion : les dosages $(X_1, \dots, X_n)$ sont fixés par l'expérimentateur.

- Effets aléatoires ou random design : $X_1, \dots, X_n$ échantillon i.i.d. de densité f_X.

  Ex : X_i=salaire moyen et Y_i = âge du décès. On sélectionne aléatoirement n individus parmi la population.

↪ Dans ce cours, nous allons considérer la regression en random design, mais les méthodes d'estimation, et une partie des résultats, sont directement transposables au fix design. Néanmoins, les résultats en fix-design font souvent des hypothèses supplémentaires (exple $(X_1, \dots, X_n)$ équidistants).

Regression avec bruit additif : fix/random design

On appelle "dispositif expérimental" ou "design" les observations $(X_1, \dots, X_n)$. Il existe deux situations :

- Effets fixes ou fix-design : $X_1, \dots, X_n$ fixé par l'utilisateur

  Ex : X_i=dosage de médicament ingéré et Y_i=taux de la molécule active dans le sang 1h après ingestion : les dosages $(X_1, \dots, X_n)$ sont fixés par l'expérimentateur.

- Effets aléatoires ou random design : $X_1, \dots, X_n$ échantillon i.i.d. de densité f_X.

  Ex : X_i=salaire moyen et Y_i = âge du décès. On sélectionne aléatoirement n individus parmi la population.

↪ Dans ce cours, nous allons considérer la regression en random design, mais les méthodes d'estimation, et une partie des résultats, sont directement transposables au fix design. Néanmoins, les résultats en fix-design font souvent des hypothèses supplémentaires (exple $(X_1, \dots, X_n)$ équidistants).

Regression avec bruit additif : fix/random design

On appelle "dispositif expérimental" ou "design" les observations $(X_1, \ldots, X_n)$. Il existe deux situations :

- Effets fixes ou fix-design : $X_1, \ldots, X_n$ fixé par l'utilisateur

  Ex : X_i=dosage de médicament ingéré et Y_i=taux de la molécule active dans le sang 1h après ingestion : les dosages $(X_1, \ldots, X_n)$ sont fixés par l'expérimentateur.
- Effets aléatoires ou random design : $X_1, \ldots, X_n$ échantillon i.i.d. de densité f_X.

  Ex : X_i=salaire moyen et Y_i = âge du décès. On sélectionne aléatoirement n individus parmi la population.

$\hookrightarrow$ Dans ce cours, nous allons considérer la regression en random design, mais les méthodes d'estimation, et une partie des résultats, sont directement transposables au fix design. Néanmoins, les résultats en fix-design font souvent des hypothèses supplémentaires (exple $(X_1, \ldots, X_n)$ équidistants).

Objectifs de l'estimation de régression

Description : informations sur la fonction de régression :

- Variations, mode, etc

Prédiction

- A partir d'un échantillon d'observation $\{(X_i, Y_i)\}_{i=1,\dots,n}$, on calcule un estimateur $\hat{r}$ de la fonction de régression $r(x) = \mathbb{E}[Y|X = x]$

  $\hookrightarrow$ On dit qu'on apprend r à partir de l'échantillon d'apprentissage $\{(X_i, Y_i)\}_{i=1,\dots,n}$.

- Pour une observation i' indépendante de l'échantillon $\{(X_i, Y_i)\}_{i=1,\dots,n}$, pour laquelle on observe uniquement $X_{i'}$, on prédit $Y_{i'}$ par :

$$\hat{Y}_{i'} = \hat{r}(X_{i'})$$

  $\hookrightarrow$ Dans l'exemple 2, on voudrait prédire la taille d'un enfant à 10 ans.

Objectifs de l'estimation de régression

Description : informations sur la fonction de régression :

- Variations, mode, etc

Prédiction

- A partir d'un échantillon d'observation $\{(X_i, Y_i)\}_{i=1,\dots,n}$, on calcule un estimateur $\hat{r}$ de la fonction de régression $r(x) = \mathbb{E}[Y|X = x]$

  $\hookrightarrow$ On dit qu'on apprend r à partir de l'échantillon d'apprentissage $\{(X_i, Y_i)\}_{i=1,\dots,n}$.

- Pour une observation i' indépendante de l'échantillon $\{(X_i, Y_i)\}_{i=1,\dots,n}$, pour laquelle on observe uniquement $X_{i'}$, on prédit $Y_{i'}$ par :

$$\hat{Y}_{i'} = \hat{r}(X_{i'})$$

  $\hookrightarrow$ Dans l'exemple 2, on voudrait prédire la taille d'un enfant à 10 ans.

Objectifs de l'estimation de régression

Description : informations sur la fonction de régression :

- Variations, mode, etc

Prédiction

- A partir d'un échantillon d'observation $\{(X_i, Y_i)\}_{i=1,\dots,n}$, on calcule un estimateur $\hat{r}$ de la fonction de régression $r(x) = \mathbb{E}[Y|X = x]$

  $\hookrightarrow$ On dit qu'on apprend r à partir de l'échantillon d'apprentissage $\{(X_i, Y_i)\}_{i=1,\dots,n}$.

- Pour une observation i' indépendante de l'échantillon $\{(X_i, Y_i)\}_{i=1,\dots,n}$, pour laquelle on observe uniquement $X_{i'}$, on prédit $Y_{i'}$ par :

$$\hat{Y}_{i'} = \hat{r}(X_{i'})$$

  $\hookrightarrow$ Dans l'exemple 2, on voudrait prédire la taille d'un enfant à 10 ans.

Objectifs de l'estimation de régression

Description : informations sur la fonction de régression :

- Variations, mode, etc

Prédiction

- A partir d'un échantillon d'observation $\{(X_i, Y_i)\}_{i=1,\dots,n}$, on calcule un estimateur $\hat{r}$ de la fonction de régression $r(x) = \mathbb{E}[Y|X = x]$

  $\hookrightarrow$ On dit qu'on apprend r à partir de l'échantillon d'apprentissage $\{(X_i, Y_i)\}_{i=1,\dots,n}$.

- Pour une observation i' indépendante de l'échantillon $\{(X_i, Y_i)\}_{i=1,\dots,n}$, pour laquelle on observe uniquement $X_{i'}$, on prédit $Y_{i'}$ par :

$$\hat{Y}_{i'} = \hat{r}(X_{i'})$$

  $\hookrightarrow$ Dans l'exemple 2, on voudrait prédire la taille d'un enfant à 10 ans.

Objectifs de l'estimation de régression

Description : informations sur la fonction de régression :

- Variations, mode, etc

Prédiction

- A partir d'un échantillon d'observation $\{(X_i, Y_i)\}_{i=1,\dots,n}$, on calcule un estimateur $\hat{r}$ de la fonction de régression $r(x) = \mathbb{E}[Y|X = x]$

  $\hookrightarrow$ On dit qu'on apprend r à partir de l'échantillon d'apprentissage $\{(X_i, Y_i)\}_{i=1,\dots,n}$.

- Pour une observation i' indépendante de l'échantillon $\{(X_i, Y_i)\}_{i=1,\dots,n}$, pour laquelle on observe uniquement $X_{i'}$, on prédit $Y_{i'}$ par :

$$\hat{Y}_{i'} = \hat{r}(X_{i'})$$

  $\hookrightarrow$ Dans l'exemple 2, on voudrait prédire la taille d'un enfant à 10 ans.

Régression non-paramétrique

- Si on n'a pas d'idée a priori sur la forme de la fonction de régression r, on va considérer un estimateur non-paramétrique.
- On veut estimer r en faisant le moins d'hypothèses possibles
- Techniques similaires à celles de la partie Estimation de densité.
  - Noyaux : plus complexe qu'en densité *(présentation rapide)*
  - Estimateurs des moindres carrés : estimateur de type projection mais avec la norme L^2 pondérée par la densité f_X du design. *(présentation détaillée)*
- Autres méthodes : k plus proches voisins, estimateurs par polynomes locaux.

Contexte et notations

Contexte

Soit $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. tq

$$Y_i = r(X_i) + \varepsilon_i$$

avec

- $\epsilon_i \perp X_i$ et $\mathbb{E}[\varepsilon_i] = 0$.
- X_i est à valeur dans $I \subset \mathbb{R}$
- $r \in L^2(I)$.

  Par la suite, on supposera également que X_i admet une densité $f_X > 0$ sur I.

Notations

- On note $\mathbf{X} = (X_i)_{i=1,\dots,n}$, $\mathbf{Y} = (Y_i)_{i=1,\dots,n} \in \mathbb{R}^n$
- Soit $\|.\|_{\ell^2}$ la norme ℓ^2 sur $\mathbb{R}^n$ et $\langle ., . \rangle_{\ell^2}$ le p.s. associé.

Contexte et notations

Contexte

Soit $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. tq

$$Y_i = r(X_i) + \varepsilon_i$$

avec

- $\epsilon_i \perp X_i$ et $\mathbb{E}[\varepsilon_i] = 0$.
- X_i est à valeur dans $I \subset \mathbb{R}$
- $r \in L^2(I)$.

  Par la suite, on supposera également que X_i admet une densité $f_X > 0$ sur I.

Notations

- On note $\mathbf{X} = (X_i)_{i=1,\dots,n}$, $\mathbf{Y} = (Y_i)_{i=1,\dots,n} \in \mathbb{R}^n$
- Soit $\|.\|_{\ell^2}$ la norme ℓ^2 sur $\mathbb{R}^n$ et $\langle ., . \rangle_{\ell^2}$ le p.s. associé.

Contexte et notations

Contexte

Soit $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. tq

$$Y_i = r(X_i) + \varepsilon_i$$

avec

- $\epsilon_i \perp X_i$ et $\mathbb{E}[\varepsilon_i] = 0$.
- X_i est à valeur dans $I \subset \mathbb{R}$
- $r \in L^2(I)$.

  Par la suite, on supposera également que X_i admet une densité $f_X > 0$ sur I.

Notations

- On note $\mathbf{X} = (X_i)_{i=1,\dots,n}$, $\mathbf{Y} = (Y_i)_{i=1,\dots,n} \in \mathbb{R}^n$
- Soit $\|.\|_{\ell^2}$ la norme ℓ^2 sur $\mathbb{R}^n$ et $\langle ., . \rangle_{\ell^2}$ le p.s. associé.

Définition de l'estimateur des moindres carrés

- Idée : trouver une fonction $\hat{r}$ tq $\hat{r}(X_i)$ est proche de Y_i.
- Contraste des moindre carrés : on va choisir un fonction $t : I \to \mathbb{R}$ qui minimise :

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2 = \frac{1}{n}\|\mathbf{Y} - t(\mathbf{X})\|^2_{\ell^2}$$

sur un espace de fonctions.

- Justification Pour tout $t : I \to \mathbb{R}$,

$$\begin{aligned}\gamma(t) = \mathbb{E}\left[\gamma_n(t)\right] &= \mathbb{E}[(Y_1 - t(X_1))^2] = \mathbb{E}[((r-t)(X_1) + \varepsilon_1)^2] \\ &= \mathbb{E}[\epsilon_1^2] + \mathbb{E}[(r-t)^2(X_1)] + 2\mathbb{E}[\epsilon_1]\mathbb{E}[(r-t)(X_1) \\ &= \sigma^2 + \mathbb{E}[(r-t)^2(X_1)]\end{aligned}$$

donc $\gamma(.)$ est minimal pour $t = r$ (mais γ est inconnu).

Définition de l'estimateur des moindres carrés

- Idée : trouver une fonction $\hat{r}$ tq $\hat{r}(X_i)$ est proche de Y_i.
- Contraste des moindre carrés : on va choisir un fonction $t : I \to \mathbb{R}$ qui minimise :

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2 = \frac{1}{n}\|\mathbf{Y} - t(\mathbf{X})\|^2_{\ell^2}$$

sur un espace de fonctions.

- Justification Pour tout $t : I \to \mathbb{R}$,

$$\begin{aligned}\gamma(t) = \mathbb{E}\left[\gamma_n(t)\right] &= \mathbb{E}[(Y_1 - t(X_1))^2] = \mathbb{E}[((r-t)(X_1) + \varepsilon_1)^2] \\ &= \mathbb{E}[\epsilon_1^2] + \mathbb{E}[(r-t)^2(X_1)] + 2\mathbb{E}[\epsilon_1]\mathbb{E}[(r-t)(X_1) \\ &= \sigma^2 + \mathbb{E}[(r-t)^2(X_1)]\end{aligned}$$

donc $\gamma(.)$ est minimal pour $t = r$ (mais γ est inconnu).

Définition de l'estimateur des moindres carrés

- Idée : trouver une fonction $\hat{r}$ tq $\hat{r}(X_i)$ est proche de Y_i.
- Contraste des moindre carrés : on va choisir un fonction $t : I \to \mathbb{R}$ qui minimise :

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2 = \frac{1}{n}\|\mathbf{Y} - t(\mathbf{X})\|_{\ell^2}^2$$

sur un espace de fonctions.

- Justification Pour tout $t : I \to \mathbb{R}$,

$$\begin{aligned}\gamma(t) = \mathbb{E}\left[\gamma_n(t)\right] &= \mathbb{E}[(Y_1 - t(X_1))^2] = \mathbb{E}[((r-t)(X_1) + \varepsilon_1)^2] \\ &= \mathbb{E}[\epsilon_1^2] + \mathbb{E}[(r-t)^2(X_1)] + 2\mathbb{E}[\epsilon_1]\mathbb{E}[(r-t)(X_1) \\ &= \sigma^2 + \mathbb{E}[(r-t)^2(X_1)]\end{aligned}$$

donc $\gamma(.)$ est minimal pour $t = r$ (mais γ est inconnu).

- Soit $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$ un espace d'approximation, un estimateur des moindres carrés sur S_D est :

$$\hat{r}_D \in \arg\min_{t \in S_D} \gamma_n(t)$$

$\hookrightarrow$ Sous certaines conditions, $\hat{r}_D$ est unique : on parle alors de l'estimateur des moindres carrés

- Soit $S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$ un espace d'approximation, un estimateur des moindres carrés sur S_D est :

$$\hat{r}_D \in \arg\min_{t \in S_D} \gamma_n(t)$$

$\hookrightarrow$ Sous certaines conditions, $\hat{r}_D$ est unique : on parle alors de l'estimateur des moindres carrés

Notations

Soit $\mathbf{X} = (X_i)_{i=1,\dots,n}$ avec $X_i \in I$.

- Pour tout $t : I \to \mathbb{R}$, soit $t(\mathbf{X}) = (t(X_1), \dots, t(X_n)) \in \mathbb{R}^n$.
- Soit $S_D = \text{vect}\,(\phi_1, \dots, \phi_D)$ un espace d'approximation, on note $S_D(\mathbf{X})$ le s.e.v. de $\mathbb{R}^n$:

$$S_D(\mathbf{X}) \quad = \quad \{t(\mathbf{X}), t \in S_D\} = \text{vect}\,(\phi_1(\mathbf{X}), \dots, \phi_D(\mathbf{X}))$$

- Soit $\mathbb{X} = (\phi_j(X_i))_{i=1,\dots,n, j=1,\dots,D} \in \mathcal{M}_{n,D}(\mathbb{R})$, alors $S_D(\mathbf{X})$ est l'espace engendré par les colonnes de $\mathbb{X}$.

Conséquences

- Pour tout $t \in S_D$, $t(.) = \sum_{j=1}^{D} \theta_j \phi_j(.)$ avec $\theta \in \mathbb{R}^D$:

$$t(\mathbf{X}) = \mathbb{X}\theta$$

- Comme $\gamma_n(t) = (1/n)\|\mathbf{Y} - t(\mathbf{X})\|_{\ell^2}^2$,

$$\hat{r}_D(.) = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j(.) \in \arg\min_{t \in S_D} \gamma_n(t) \quad \Leftrightarrow \quad \hat{\theta}^D \in \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$$

.

- De plus, soit $\Pi_{S_D(\mathbf{X})}$ le projecteur orthog. de $\mathbb{R}^n$ sur $S_D(\mathbf{X})$:

$$\hat{r}_D \in \arg\min_{t \in S_D} \|\mathbf{Y} - t(\mathbf{X})\|_{\ell^2}^2$$

$$\Leftrightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in S_D(\mathbf{X})} \|\mathbf{Y} - U\|_{\ell^2}^2 = \Pi_{S_D(\mathbf{X})}\mathbf{Y}$$

Conséquences

- Pour tout $t \in S_D$, $t(.) = \sum_{j=1}^{D} \theta_j \phi_j(.)$ avec $\theta \in \mathbb{R}^D$:

$$t(\mathbf{X}) = \mathbb{X}\theta$$

- Comme $\gamma_n(t) = (1/n)\|\mathbf{Y} - t(\mathbf{X})\|_{\ell^2}^2$,

$$\hat{r}_D(.) = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j(.) \in \arg\min_{t \in S_D} \gamma_n(t) \quad \Leftrightarrow \quad \hat{\theta}^D \in \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$$

.

- De plus, soit $\Pi_{S_D(\mathbf{X})}$ le projecteur orthog. de $\mathbb{R}^n$ sur $S_D(\mathbf{X})$:

$$\hat{r}_D \in \arg\min_{t \in S_D} \|\mathbf{Y} - t(\mathbf{X})\|_{\ell^2}^2$$

$$\Leftrightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in S_D(\mathbf{X})} \|\mathbf{Y} - U\|_{\ell^2}^2 = \Pi_{S_D(\mathbf{X})}\mathbf{Y}$$

Conséquences

- Pour tout $t \in S_D$, $t(.) = \sum_{j=1}^{D} \theta_j \phi_j(.)$ avec $\theta \in \mathbb{R}^D$:

$$t(\mathbf{X}) = \mathbb{X}\theta$$

- Comme $\gamma_n(t) = (1/n)\|\mathbf{Y} - t(\mathbf{X})\|_{\ell^2}^2$,

$$\hat{r}_D(.) = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j(.) \in \arg\min_{t \in S_D} \gamma_n(t) \quad \Leftrightarrow \quad \hat{\theta}^D \in \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$$

.

- De plus, soit $\Pi_{S_D(\mathbf{X})}$ le projecteur orthog. de $\mathbb{R}^n$ sur $S_D(\mathbf{X})$:

$$\hat{r}_D \in \arg\min_{t \in S_D} \|\mathbf{Y} - t(\mathbf{X})\|_{\ell^2}^2$$

$$\Leftrightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in S_D(\mathbf{X})} \|\mathbf{Y} - U\|_{\ell^2}^2 = \Pi_{S_D(\mathbf{X})}\mathbf{Y}$$

Parallèle avec la régression linéaire

- Régression linéaire multiple :
  - Soient $(x_i^0, Y_i)_{i=1,\dots,n} \in \mathbb{R}^D \times \mathbb{R}$ tq :

$$Y_i = \sum_{j=1}^{D} \theta_j^0 x_{i,j}^0 + \varepsilon_i \quad \text{avec} \quad \{\varepsilon_i\} \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$

  - Estimation :

$$\hat{\theta}^0 = \arg\min_{\theta \in \mathbb{R}^D} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{D} \theta_j x_{i,j}^0 \right)^2 = \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}^0 \theta\|_{\ell^2}^2$$

avec $\mathbb{X}^0 = \left(x_{i,j}^0\right)_{i=1,\dots,n, j=1,\dots,D}$.

- Analogie avec l'estimateur des MC

$$\hat{\theta}^D = \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$$

avec $\mathbb{X} = \left(\phi_j(X_i)\right)_{i=1,\dots,n, j=1,\dots,D}$.

Parallèle avec la régression linéaire

- Régression linéaire multiple :
  - Soient $(x_i^0, Y_i)_{i=1,\dots,n} \in \mathbb{R}^D \times \mathbb{R}$ tq :

$$Y_i = \sum_{j=1}^{D} \theta_j^0 x_{i,j}^0 + \varepsilon_i \quad \text{avec} \quad \{\varepsilon_i\} \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$

  - Estimation :

$$\hat{\theta}^0 = \arg\min_{\theta \in \mathbb{R}^D} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{D} \theta_j x_{i,j}^0 \right)^2 = \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}^0 \theta\|_{\ell^2}^2$$

    avec $\mathbb{X}^0 = \left(x_{i,j}^0\right)_{i=1,\dots,n, j=1,\dots,D}$.

- Analogie avec l'estimateur des MC

$$\hat{\theta}^D = \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$$

avec $\mathbb{X} = \left(\phi_j(X_i)\right)_{i=1,\dots,n, j=1,\dots,D}$.

Parallèle avec la régression linéaire

- Régression linéaire multiple :
  - Soient $(x_i^0, Y_i)_{i=1,\dots,n} \in \mathbb{R}^D \times \mathbb{R}$ tq :

$$Y_i = \sum_{j=1}^{D} \theta_j^0 x_{i,j}^0 + \varepsilon_i \quad \text{avec} \quad \{\varepsilon_i\} \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$

  - Estimation :

$$\hat{\theta}^0 = \arg\min_{\theta \in \mathbb{R}^D} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{D} \theta_j x_{i,j}^0 \right)^2 = \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}^0 \theta\|_{\ell^2}^2$$

    avec $\mathbb{X}^0 = \left(x_{i,j}^0\right)_{i=1,\dots,n, j=1,\dots,D}$.

- Analogie avec l'estimateur des MC

$$\hat{\theta}^D = \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$$

avec $\mathbb{X} = \left(\phi_j(X_i)\right)_{i=1,\dots,n, j=1,\dots,D}$.

Calcul de l'estimateur des MC (I)

- $\theta \to \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$ fonction convexe, donc $\theta \in \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$ ssi pour tout j' :

$$\frac{\partial}{\partial \theta_{j'}} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2 = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{D} \theta_j \mathbb{X}_{i,j} \right) \mathbb{X}_{i,j'} = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} Y_i \mathbb{X}_{i,j'} = \sum_{j=1}^{D} \theta_j \left(\sum_{i=1}^{n} \mathbb{X}_{i,j} \mathbb{X}_{i,j'} \right), \quad \forall j' = 1, \ldots, D$$

- En écriture matricielle :

$$\hat{\theta}^D \in \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2 \quad \Leftrightarrow \quad \mathbb{X}^t \mathbb{X} \hat{\theta}^D = \mathbb{X}^t Y$$

Calcul de l'estimateur des MC (I)

- $\theta \to \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$ fonction convexe, donc $\theta \in \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2$ ssi pour tout j' :

$$\frac{\partial}{\partial \theta_{j'}} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2 = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{D} \theta_j \mathbb{X}_{i,j} \right) \mathbb{X}_{i,j'} = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} Y_i \mathbb{X}_{i,j'} = \sum_{j=1}^{D} \theta_j \left(\sum_{i=1}^{n} \mathbb{X}_{i,j} \mathbb{X}_{i,j'} \right), \quad \forall j' = 1, \ldots, D$$

- En écriture matricielle :

$$\hat{\theta}^D \in \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \mathbb{X}\theta\|_{\ell^2}^2 \quad \Leftrightarrow \quad \mathbb{X}^t \mathbb{X} \hat{\theta}^D = \mathbb{X}^t Y$$

- $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\, Y$ admet une unique solution $\hat{\theta}^D$ ssi $\mathbb{X}^t\mathbb{X}$ est inversible. Dans ce cas, on dit que l'estimateur des MC est identifiable.
- De plus,

$$\mathbb{X}^t\mathbb{X} = \left(\langle \phi_j(\mathbf{X}), \phi_{j'}(\mathbf{X})\rangle_{\ell^2}\right)_{j,j'=1,\dots,D}$$

- Rappel : *soit $\mathbf{x} = (x_1, \dots, x_k)$ éléments d'un e.v. E et $\langle \cdot, \cdot \rangle$ un p.s. sur E. Alors le rang de la famille $\mathbf{x}$ est égal au rang de sa matrice de Gram $\left(\langle x_i, x_j\rangle\right)_{i,j=1,\dots,k}$.*

  D'où

$$\text{rang}\left(\mathbb{X}^t\mathbb{X}\right) = \text{rang}\left(S_D(\mathbf{X})\right)$$

- Ainsi, l'estimateur des MC est identifiable ssi $\text{rang}\left(S_D(\mathbf{X})\right) = D$.

- $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\, Y$ admet une unique solution $\hat{\theta}^D$ ssi $\mathbb{X}^t\mathbb{X}$ est inversible. Dans ce cas, on dit que l'estimateur des MC est identifiable.
- De plus,

$$\mathbb{X}^t\mathbb{X} = \left(\langle \phi_j(\mathbf{X}), \phi_{j'}(\mathbf{X})\rangle_{\ell^2}\right)_{j,j'=1,\dots,D}$$

- Rappel : *soit $\mathbf{x} = (x_1, \dots, x_k)$ éléments d'un e.v. E et $\langle \cdot, \cdot \rangle$ un p.s. sur E. Alors le rang de la famille $\mathbf{x}$ est égal au rang de sa matrice de Gram $(\langle x_i, x_j \rangle)_{i,j=1,\dots,k}$.*

  D'où

$$\text{rang}\left(\mathbb{X}^t\mathbb{X}\right) = \text{rang}\left(S_D(\mathbf{X})\right)$$

- Ainsi, l'estimateur des MC est identifiable ssi $\text{rang}\left(S_D(\mathbf{X})\right) = D$.

- $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\, Y$ admet une unique solution $\hat{\theta}^D$ ssi $\mathbb{X}^t\mathbb{X}$ est inversible. Dans ce cas, on dit que l'estimateur des MC est identifiable.
- De plus,
$$\mathbb{X}^t\mathbb{X} = \left(\langle \phi_j(\mathbf{X}), \phi_{j'}(\mathbf{X})\rangle_{\ell^2}\right)_{j,j'=1,\dots,D}$$
- Rappel : *soit* $\mathbf{x} = (x_1, \dots, x_k)$ *éléments d'un e.v.* E *et* $\langle \cdot, \cdot \rangle$ *un p.s. sur* E*. Alors le rang de la famille* $\mathbf{x}$ *est égal au rang de sa matrice de Gram* $\left(\langle x_i, x_j\rangle\right)_{i,j=1,\dots,k}$.

  D'où
$$\text{rang}\left(\mathbb{X}^t\mathbb{X}\right) = \text{rang}\left(S_D(\mathbf{X})\right)$$
- Ainsi, l'estimateur des MC est identifiable ssi $\text{rang}\left(S_D(\mathbf{X})\right) = D$.

- $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\, Y$ admet une unique solution $\hat{\theta}^D$ ssi $\mathbb{X}^t\mathbb{X}$ est inversible. Dans ce cas, on dit que l'estimateur des MC est identifiable.
- De plus,
$$\mathbb{X}^t\mathbb{X} = \left(\langle \phi_j(\mathbf{X}), \phi_{j'}(\mathbf{X})\rangle_{\ell^2}\right)_{j,j'=1,\dots,D}$$
- Rappel : *soit* $\mathbf{x} = (x_1, \dots, x_k)$ *éléments d'un e.v.* E *et* $\langle \cdot, \cdot \rangle$ *un p.s. sur* E*. Alors le rang de la famille* $\mathbf{x}$ *est égal au rang de sa matrice de Gram* $\left(\langle x_i, x_j\rangle\right)_{i,j=1,\dots,k}$.

  D'où
$$\text{rang}\left(\mathbb{X}^t\mathbb{X}\right) = \text{rang}\left(S_D(\mathbf{X})\right)$$
- Ainsi, l'estimateur des MC est identifiable ssi $\text{rang}\left(S_D(\mathbf{X})\right) = D$.

- $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\, Y$ admet une unique solution $\hat{\theta}^D$ ssi $\mathbb{X}^t\mathbb{X}$ est inversible. Dans ce cas, on dit que l'estimateur des MC est identifiable.
- De plus,
$$\mathbb{X}^t\mathbb{X} = \left(\langle \phi_j(\mathbf{X}), \phi_{j'}(\mathbf{X})\rangle_{\ell^2}\right)_{j,j'=1,\dots,D}$$
- Rappel : *soit* $\mathbf{x} = (x_1, \dots, x_k)$ *éléments d'un e.v.* E *et* $\langle \cdot, \cdot \rangle$ *un p.s. sur* E*. Alors le rang de la famille* $\mathbf{x}$ *est égal au rang de sa matrice de Gram* $\left(\langle x_i, x_j\rangle\right)_{i,j=1,\dots,k}$.

  D'où
$$\text{rang}\left(\mathbb{X}^t\mathbb{X}\right) = \text{rang}\left(S_D(\mathbf{X})\right)$$
- Ainsi, l'estimateur des MC est identifiable ssi $\text{rang}\left(S_D(\mathbf{X})\right) = D$.

Cas identifiable : $\dim(S_D(\mathbf{X})) = D$

Rappel : $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\,Y$.

- $\text{rang}(\mathbb{X}^t\mathbb{X}) = D$ donc $\mathbb{X}^t\mathbb{X}$ est inversible, d'où $\hat{r}_D = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j$ avec

$$\hat{\theta}^D = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$$

- De plus, $\hat{r}_D(\mathbf{X}) = \mathbb{X}\hat{\theta}^D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$
- Remarque : soit $M_D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t$, alors $M_D M_D = M_D$, donc M_D est bien la matrice d'un projecteur.
- Remarque 2 : comme $S_D(\mathbf{X})$ sev de $\mathbb{R}^n$, l'estimateur des MC ne peut pas être identifiable si $D > n$.

Cas identifiable : $\dim(S_D(\mathbf{X})) = D$

Rappel : $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\,Y$.

- rang$(\mathbb{X}^t\mathbb{X}) = D$ donc $\mathbb{X}^t\mathbb{X}$ est inversible, d'où $\hat{r}_D = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j$ avec

$$\hat{\theta}^D = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$$

- De plus, $\hat{r}_D(\mathbf{X}) = \mathbb{X}\hat{\theta}^D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$
- Remarque : soit $M_D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t$, alors $M_D M_D = M_D$, donc M_D est bien la matrice d'un projecteur.
- Remarque 2 : comme $S_D(\mathbf{X})$ sev de $\mathbb{R}^n$, l'estimateur des MC ne peut pas être identifiable si $D > n$.

Cas identifiable : $\dim(S_D(\mathbf{X})) = D$

Rappel : $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\,Y$.

- rang$(\mathbb{X}^t\mathbb{X}) = D$ donc $\mathbb{X}^t\mathbb{X}$ est inversible, d'où $\hat{r}_D = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j$ avec

$$\hat{\theta}^D = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$$

- De plus, $\hat{r}_D(\mathbf{X}) = \mathbb{X}\hat{\theta}^D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$
- Remarque : soit $M_D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t$, alors $M_D M_D = M_D$, donc M_D est bien la matrice d'un projecteur.
- Remarque 2 : comme $S_D(\mathbf{X})$ sev de $\mathbb{R}^n$, l'estimateur des MC ne peut pas être identifiable si $D > n$.

Cas identifiable : $\dim(S_D(\mathbf{X})) = D$

Rappel : $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\,Y$.

- rang$(\mathbb{X}^t\mathbb{X}) = D$ donc $\mathbb{X}^t\mathbb{X}$ est inversible, d'où $\hat{r}_D = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j$ avec

$$\hat{\theta}^D = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$$

- De plus, $\hat{r}_D(\mathbf{X}) = \mathbb{X}\hat{\theta}^D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\,Y$
- Remarque : soit $M_D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t$, alors $M_D M_D = M_D$, donc M_D est bien la matrice d'un projecteur.
- Remarque 2 : comme $S_D(\mathbf{X})$ sev de $\mathbb{R}^n$, l'estimateur des MC ne peut pas être identifiable si $D > n$.

Cas identifiable : $\dim(S_D(\mathbf{X})) = D$

Rappel : $\mathbb{X}^t\mathbb{X}\hat{\theta}^D = \mathbb{X}^t\, Y$.

- $\text{rang}(\mathbb{X}^t\mathbb{X}) = D$ donc $\mathbb{X}^t\mathbb{X}$ est inversible, d'où $\hat{r}_D = \sum_{j=1}^{D} \hat{\theta}_j^D \phi_j$ avec

$$\hat{\theta}^D = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\, Y$$

- De plus, $\hat{r}_D(\mathbf{X}) = \mathbb{X}\hat{\theta}^D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\, Y$
- Remarque : soit $M_D = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t$, alors $M_D M_D = M_D$, donc M_D est bien la matrice d'un projecteur.
- Remarque 2 : comme $S_D(\mathbf{X})$ sev de $\mathbb{R}^n$, l'estimateur des MC ne peut pas être identifiable si $D > n$.

Cas non identifiable : $\dim(S_D(\mathbf{X})) < D$

- L'identifiabilité ou non de l'estimateur des MC dépend de l'échantillon $\mathbf{X}$.
- Il y a une solution théorique dans le cas non-identifiable, mais ce n'est pas toujours la meilleure en pratique.
- En pratique : si $\dim(S_D(\mathbf{X})) < D$, l'espace S_D est mal adapté

  $\hookrightarrow$ situation qui se présente quand on calcule $\{\hat{r}_D, D = 1, \ldots, D_{max}\}$ (ex : sélection de modèle).

Cas non-identifiable : solution théorique

Supposons que $\dim(S_D(\mathbf{X})) = D' < D$.

- **Résultat d'algèbre linéaire.** *Soit* $(u_1, \ldots, u_k)$ *une famille d'un e.v. telle que dim* $(vect(u_1, \ldots, u_k))) = k' < k$, *alors il existe* $J \subset \{1, \ldots, k\}$ *tel que :*
  - $\#J = k'$
  - $vect(u_j, j \in J) = vect(u_1, \ldots, u_k)$.
- Soit $J \subset \{1, \ldots, D\}$ tel que $\#J = D'$ et

$$S_D(\mathbf{X}) = \text{vect}\{\phi_j(\mathbf{X}), j \in J\}.$$

- Soit $\tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.)$ avec

$$\tilde{\mathbb{X}} = \left(\sum_{i=1}^{n} \phi_j(X_i)\phi_{j'}(X_i) \right)_{j,j' \in J}, \quad \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t \tilde{\mathbb{X}})^{-1} \tilde{\mathbb{X}}^t \mathbf{Y}$$

- Remarque : cela revient à fixer $\hat{\theta}_j^D = 0, \quad \forall j \notin J'$.

Cas non-identifiable : solution théorique

Supposons que $\dim(S_D(\mathbf{X})) = D' < D$.

- **Résultat d'algèbre linéaire.** *Soit $(u_1, \ldots, u_k)$ une famille d'un e.v. telle que $\dim(\text{vect}(u_1, \ldots, u_k))) = k' < k$, alors il existe $J \subset \{1, \ldots, k\}$ tel que :*
  - $\#J = k'$
  - $\text{vect}(u_j, j \in J) = \text{vect}(u_1, \ldots, u_k)$.
- Soit $J \subset \{1, \ldots, D\}$ tel que $\#J = D'$ et

$$S_D(\mathbf{X}) = \text{vect}\{\phi_j(\mathbf{X}), j \in J\}.$$

- Soit $\tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.)$ avec

$$\tilde{\mathbb{X}} = \left(\sum_{i=1}^{n} \phi_j(X_i)\phi_{j'}(X_i) \right)_{j,j' \in J}, \quad \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t \tilde{\mathbb{X}})^{-1} \tilde{\mathbb{X}}^t \mathbf{Y}$$

- Remarque : cela revient à fixer $\hat{\theta}_j^D = 0, \quad \forall j \notin J'$.

Cas non-identifiable : solution théorique

Supposons que $\dim(S_D(\mathbf{X})) = D' < D$.

- **Résultat d'algèbre linéaire.** *Soit* $(u_1, \ldots, u_k)$ *une famille d'un e.v. telle que dim* $(vect(u_1, \ldots, u_k))) = k' < k$, *alors il existe* $J \subset \{1, \ldots, k\}$ *tel que :*
  - $\#J = k'$
  - $vect(u_j, j \in J) = vect(u_1, \ldots, u_k)$.
- Soit $J \subset \{1, \ldots, D\}$ tel que $\#J = D'$ et

$$S_D(\mathbf{X}) = \text{vect}\{\phi_j(\mathbf{X}), j \in J\}.$$

- Soit $\tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.)$ avec

$$\tilde{\mathbb{X}} = \left(\sum_{i=1}^{n} \phi_j(X_i)\phi_{j'}(X_i) \right)_{j,j' \in J}, \quad \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t \tilde{\mathbb{X}})^{-1} \tilde{\mathbb{X}}^t \mathbf{Y}$$

- Remarque : cela revient à fixer $\hat{\theta}_j^D = 0, \quad \forall j \notin J'$.

Cas non-identifiable : solution théorique

Supposons que $\dim(S_D(\mathbf{X})) = D' < D$.

- **Résultat d'algèbre linéaire.** *Soit* $(u_1, \ldots, u_k)$ *une famille d'un e.v. telle que dim* $(vect(u_1, \ldots, u_k))) = k' < k$, *alors il existe* $J \subset \{1, \ldots, k\}$ *tel que :*
  - $\#J = k'$
  - $vect(u_j, j \in J) = vect(u_1, \ldots, u_k)$.
- Soit $J \subset \{1, \ldots, D\}$ tel que $\#J = D'$ et

$$S_D(\mathbf{X}) = \text{vect}\{\phi_j(\mathbf{X}), j \in J\}.$$

- Soit $\tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.)$ avec

$$\tilde{\mathbb{X}} = \left(\sum_{i=1}^{n} \phi_j(X_i)\phi_{j'}(X_i) \right)_{j,j' \in J}, \quad \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t \tilde{\mathbb{X}})^{-1} \tilde{\mathbb{X}}^t \mathbf{Y}$$

- Remarque : cela revient à fixer $\hat{\theta}_j^D = 0, \quad \forall j \notin J'$.

- $\tilde{r}_D$ est bien un estimateur des MC sur S_D :

$$\tilde{r}_D(\mathbf{X}) = \arg \min_{U \in S_D(\mathbf{X})} \| Y - U \|_{\ell^2}^2 \Leftrightarrow \tilde{r}_D \in \arg \min_{t \in S_D} \gamma_n(t)$$

- Cette solution théorique n'est pas toujours la meilleure

- $\tilde{r}_D$ est bien un estimateur des MC sur S_D :

$$\tilde{r}_D(\mathbf{X}) = \arg \min_{U \in S_D(\mathbf{X})} \| Y - U \|_{\ell^2}^2 \Leftrightarrow \tilde{r}_D \in \arg \min_{t \in S_D} \gamma_n(t)$$

- Cette solution théorique n'est pas toujours la meilleure

Cas non identifiable : histogramme réguliers

- Soit $I = [0, 1]$ et

$$S_D = \text{vect}\{\phi_j = \sqrt{D} 1_{I_j}, j = 1, \ldots, D\} \quad \text{avec} \quad I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$$

- $\mathbb{X}^t\mathbb{X}$ est diagonale et

$$(\mathbb{X}\mathbb{X})_{j,j} = D \times \text{Card}\left\{i = 1, \ldots, n, X_i \in I_j\right\}.$$

- Supposons que $\mathbf{X}$ est tel que $\dim(S_D(\mathbf{X})) = D - 1$, alors $\text{rang}(\mathbb{X}^t\mathbb{X}) = D - 1$ i.e. il existe $j_0 \in \{1, \ldots, D\}$ tq $(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$.

- Solution théorique : Soit $J = \{1, \ldots j_0 - 1, j_0 + 1, \ldots, D\}$,

$$\begin{cases} \tilde{\mathbb{X}} = \mathbb{X}[, J] \\ \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t\tilde{\mathbb{X}})^{-1}\tilde{\mathbb{X}}^t\mathbf{Y} \\ \tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.) \end{cases}$$

Alors

$$\tilde{r}_D(x) = 0, \quad \forall x \in I_{j_0}$$

Cas non identifiable : histogramme réguliers

- Soit $I = [0, 1]$ et

$$S_D = \text{vect}\{\phi_j = \sqrt{D} 1_{I_j}, j = 1, \ldots, D\} \quad \text{avec} \quad I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$$

- $\mathbb{X}^t\mathbb{X}$ est diagonale et

$$(\mathbb{X}\mathbb{X})_{j,j} = D \times \text{Card}\left\{i = 1, \ldots, n, X_i \in I_j\right\}.$$

- Supposons que $\mathbf{X}$ est tel que $\dim(S_D(\mathbf{X})) = D - 1$, alors $\text{rang}(\mathbb{X}^t\mathbb{X}) = D - 1$ i.e. il existe $j_0 \in \{1, \ldots, D\}$ tq $(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$.

- Solution théorique : Soit $J = \{1, \ldots j_0 - 1, j_0 + 1, \ldots, D\}$,

$$\begin{cases} \tilde{\mathbb{X}} = \mathbb{X}[, J] \\ \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t\tilde{\mathbb{X}})^{-1}\tilde{\mathbb{X}}^t\mathbf{Y} \\ \tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.) \end{cases}$$

Alors

$$\tilde{r}_D(x) = 0, \quad \forall x \in I_{j_0}$$

Cas non identifiable : histogramme réguliers

- Soit $I = [0, 1]$ et

$$S_D = \text{vect}\{\phi_j = \sqrt{D} 1_{I_j}, j = 1, \ldots, D\} \quad \text{avec} \quad I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$$

- $\mathbb{X}^t\mathbb{X}$ est diagonale et

$$(\mathbb{X}\mathbb{X})_{j,j} = D \times \text{Card}\left\{i = 1, \ldots, n, X_i \in I_j\right\}.$$

- Supposons que $\mathbf{X}$ est tel que $\dim(S_D(\mathbf{X})) = D - 1$, alors $\text{rang}(\mathbb{X}^t\mathbb{X}) = D - 1$ i.e. il existe $j_0 \in \{1, \ldots, D\}$ tq $(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$.

- Solution théorique : Soit $J = \{1, \ldots j_0 - 1, j_0 + 1, \ldots, D\}$,

$$\begin{cases} \tilde{\mathbb{X}} = \mathbb{X}[, J] \\ \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t\tilde{\mathbb{X}})^{-1}\tilde{\mathbb{X}}^t\mathbf{Y} \\ \tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.) \end{cases}$$

Alors

$$\tilde{r}_D(x) = 0, \quad \forall x \in I_{j_0}$$

Cas non identifiable : histogramme réguliers

- Soit $I = [0, 1]$ et

$$S_D = \text{vect}\{\phi_j = \sqrt{D} 1_{I_j}, j = 1, \ldots, D\} \quad \text{avec} \quad I_j = \left[\frac{j-1}{D}, \frac{j}{D}\right[$$

- $\mathbb{X}^t\mathbb{X}$ est diagonale et

$$(\mathbb{X}\mathbb{X})_{j,j} = D \times \text{Card}\left\{i = 1, \ldots, n, X_i \in I_j\right\}.$$

- Supposons que $\mathbf{X}$ est tel que $\dim(S_D(\mathbf{X})) = D - 1$, alors $\text{rang}(\mathbb{X}^t\mathbb{X}) = D - 1$ i.e. il existe $j_0 \in \{1, \ldots, D\}$ tq $(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$.

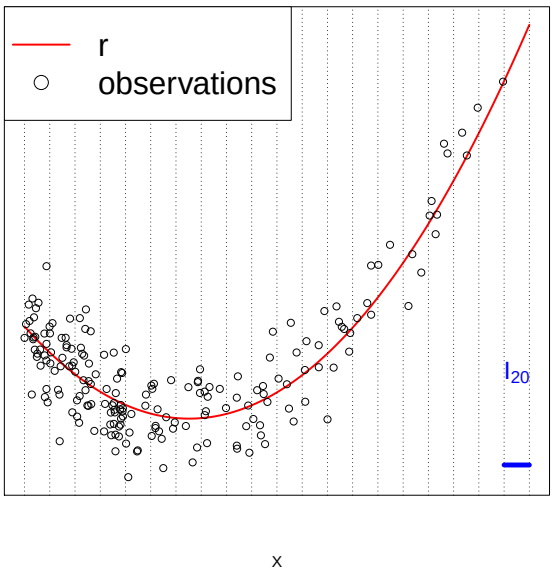
- Solution théorique : Soit $J = \{1, \ldots j_0 - 1, j_0 + 1, \ldots, D\}$,

$$\begin{cases} \tilde{\mathbb{X}} = \mathbb{X}[, J] \\ \tilde{\theta}_j^D = (\tilde{\mathbb{X}}^t\tilde{\mathbb{X}})^{-1}\tilde{\mathbb{X}}^t\mathbf{Y} \\ \tilde{r}_D(.) = \sum_{j \in J} \tilde{\theta}_j^D \phi_j(.) \end{cases}$$

Alors

$$\tilde{r}_D(x) = 0, \quad \forall x \in I_{j_0}$$

Cas non identifiable : histogramme réguliers (2)



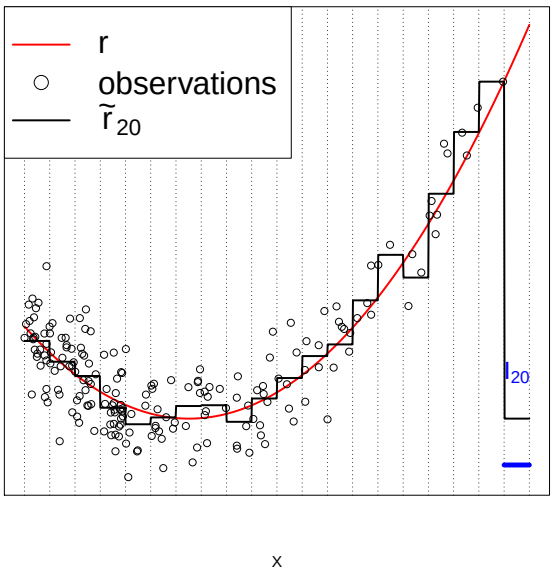
$(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$ donc $\tilde{\theta}_{j_0} = 0$ (avec $j_0 = 20$)

- Interpretation erronée : "r est proche de 0 sur I_{j_0}"
- Interpretation juste : "f_X est proche de 0 sur I_{j_0}", donc nous n'avons pas assez d'observations sur I_{j_0} pour estimer r.

Mais $\tilde{r}_D$ n'est pas l'unique estimateur des MC

- N'importe quelle valeur pour $\tilde{\theta}_{j_0}$ convient.
- Ex : $\tilde{\theta}_{j_0} = \tilde{\theta}_{j_0-1}$

Cas non identifiable : histogramme réguliers (2)



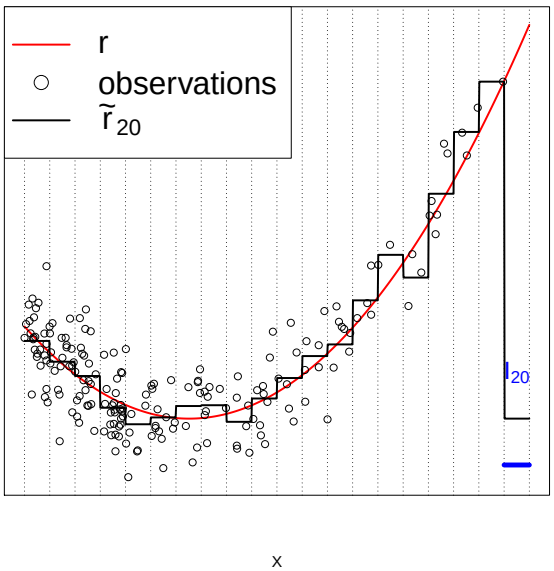
$(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$ donc $\tilde{\theta}_{j_0} = 0$ (avec $j_0 = 20$)

- Interpretation erronée : "r est proche de 0 sur I_{j_0}"
- Interpretation juste : "f_X est proche de 0 sur I_{j_0}", donc nous n'avons pas assez d'observations sur I_{j_0} pour estimer r.

Mais $\tilde{r}_D$ n'est pas l'unique estimateur des MC

- N'importe quelle valeur pour $\tilde{\theta}_{j_0}$ convient.
- Ex : $\tilde{\theta}_{j_0} = \tilde{\theta}_{j_0-1}$

Cas non identifiable : histogramme réguliers (2)



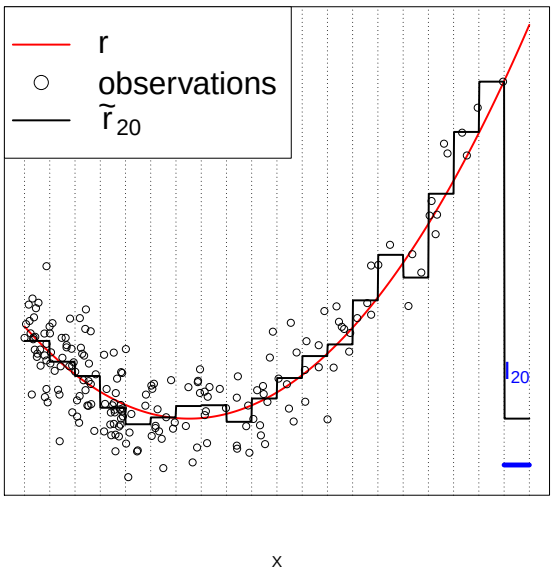
$(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$ donc $\tilde{\theta}_{j_0} = 0$ (avec $j_0 = 20$)

- Interpretation erronée : "r est proche de 0 sur I_{j_0}"
- Interpretation juste : " f_X est proche de 0 sur I_{j_0}", donc nous n'avons pas assez d'observations sur I_{j_0} pour estimer r.

Mais $\tilde{r}_D$ n'est pas l'unique estimateur des MC

- N'importe quelle valeur pour $\tilde{\theta}_{j_0}$ convient.
- Ex : $\tilde{\theta}_{j_0} = \tilde{\theta}_{j_0-1}$

Cas non identifiable : histogramme réguliers (2)



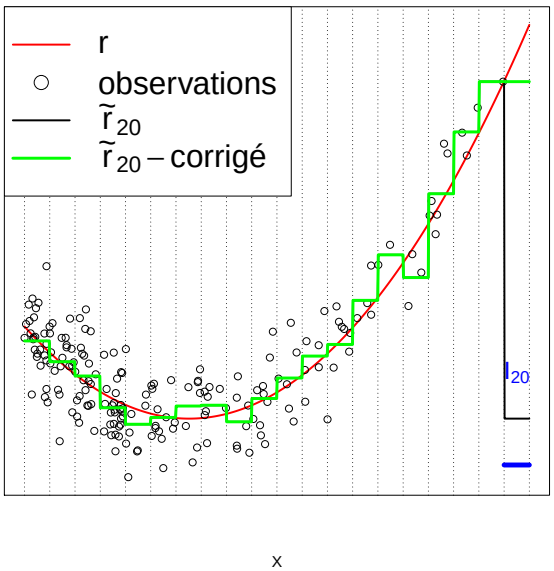
$(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$ donc $\tilde{\theta}_{j_0} = 0$ (avec $j_0 = 20$)

- Interpretation erronée : "r est proche de 0 sur I_{j_0}"
- Interpretation juste : " f_X est proche de 0 sur I_{j_0}", donc nous n'avons pas assez d'observations sur I_{j_0} pour estimer r.

Mais $\tilde{r}_D$ n'est pas l'unique estimateur des MC

- N'importe quelle valeur pour $\tilde{\theta}_{j_0}$ convient.
- Ex : $\tilde{\theta}_{j_0} = \tilde{\theta}_{j_0-1}$

Cas non identifiable : histogramme réguliers (2)

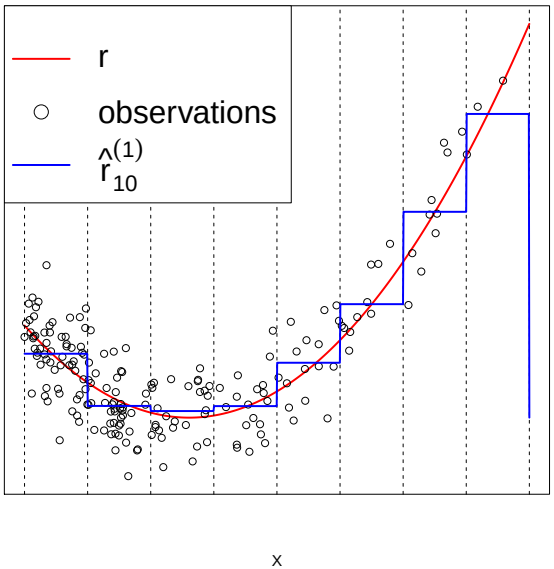


$(\mathbb{X}^t\mathbb{X})_{j_0,j_0} = 0$ donc $\tilde{\theta}_{j_0} = 0$ (avec $j_0 = 20$)

- Interpretation erronée : "r est proche de 0 sur I_{j_0}"
- Interpretation juste : " f_X est proche de 0 sur I_{j_0}", donc nous n'avons pas assez d'observations sur I_{j_0} pour estimer r.

Mais $\tilde{r}_D$ n'est pas l'unique estimateur des MC

- N'importe quelle valeur pour $\tilde{\theta}_{j_0}$ convient.
- Ex : $\tilde{\theta}_{j_0} = \tilde{\theta}_{j_0-1}$



Alternatives

- Prendre un D plus petit
- Considérer des histogrammes irréguliers
- Restreindre l'intervalle d'estimation.
- Considérer une autre base de fonction

Identifiabilité pour les modèles trigonométriques

- Supposons que $I = [a, b]$ et $f_X(x) > 0,\ \forall x \in I$.
- Soit $\{\phi_j\}_{j \in \mathbb{N}^*}$ la base trigonométrique, et S_D le modèle

$$S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$$

- Pour tout $D \leq n$, on montre que

$$\dim(S_D(\mathbf{X})) = D \quad \text{p.s.}$$

  i.e. la probabilité de tirer un échantillon $\mathbf{X}$ tel que $\dim(S_D(\mathbf{X})) < D$ vaut 0.
- Conséquence : pour $D \leq n$, l'estimateur des moindres carrés est unique p.s.

Identifiabilité pour les modèles trigonométriques

- Supposons que $I = [a, b]$ et $f_X(x) > 0,\ \forall x \in I$.
- Soit $\{\phi_j\}_{j \in \mathbb{N}^*}$ la base trigonométrique, et S_D le modèle

$$S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$$

- Pour tout $D \leq n$, on montre que

$$\dim(S_D(\mathbf{X})) = D \quad \text{p.s.}$$

  i.e. la probabilité de tirer un échantillon $\mathbf{X}$ tel que $\dim(S_D(\mathbf{X})) < D$ vaut 0.
- Conséquence : pour $D \leq n$, l'estimateur des moindres carrés est unique p.s.

Identifiabilité pour les modèles trigonométriques

- Supposons que $I = [a, b]$ et $f_X(x) > 0,\ \forall x \in I$.
- Soit $\{\phi_j\}_{j \in \mathbb{N}^*}$ la base trigonométrique, et S_D le modèle

$$S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$$

- Pour tout $D \leq n$, on montre que

$$\dim(S_D(\mathbf{X})) = D \quad \text{p.s.}$$

  i.e. la probabilité de tirer un échantillon $\mathbf{X}$ tel que $\dim(S_D(\mathbf{X})) < D$ vaut 0.

- Conséquence : pour $D \leq n$, l'estimateur des moindres carrés est unique p.s.

Identifiabilité pour les modèles trigonométriques

- Supposons que $I = [a, b]$ et $f_X(x) > 0,\ \forall x \in I$.
- Soit $\{\phi_j\}_{j \in \mathbb{N}^*}$ la base trigonométrique, et S_D le modèle

$$S_D = \text{vect}\{\phi_j, j = 1, \ldots, D\}$$

- Pour tout $D \leq n$, on montre que

$$\dim(S_D(\mathbf{X})) = D \quad \text{p.s.}$$

  i.e. la probabilité de tirer un échantillon $\mathbf{X}$ tel que $\dim(S_D(\mathbf{X})) < D$ vaut 0.
- Conséquence : pour $D \leq n$, l'estimateur des moindres carrés est unique p.s.

L'erreur d'approximation, un risque pertinent ?

- Soit $\hat{r}$ un estimateur de r calculé à partir d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$.
- Idée naïve : Par définition $r(X_i) = \mathbb{E}[Y_i|X_i]$, d'où la valeur prédite pour Y_i connaissant X_i est

$$\hat{Y}_i = \hat{r}(X_i)$$

donc on pourrait considérer l'erreur d'approximation comme critère d'évalution de $\hat{r}$:

$$\text{MCR}(\hat{r}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}(X_i))^2 \quad \text{(Moyenne des Carrés des Résidus)}$$

- Constat : Pour les estimateurs des MC, $\text{MCR}(\hat{r}_D)$ diminue automatiquement quand S_D augmente : :

$$\hat{r}^D = \arg\min_{t \in S_D} \frac{1}{n}\sum_{i=1}^{n}(t(X_i) - Y_i)^2 \quad \Rightarrow \quad MCR(\hat{r}^D) = \min_{t \in S_D} \frac{1}{n}\sum_{i=1}^{n}(t(X_i) - Y_i)^2$$

$$\text{donc} \quad S_D \subset S_{D'} \quad \Rightarrow \quad \text{MCR}(\hat{r}_D) \geq \text{MCR}(\hat{r}_{D'})$$

L'erreur d'approximation, un risque pertinent ?

- Soit $\hat{r}$ un estimateur de r calculé à partir d'un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$.
- Idée naïve : Par définition $r(X_i) = \mathbb{E}[Y_i|X_i]$, d'où la valeur prédite pour Y_i connaissant X_i est

$$\hat{Y}_i = \hat{r}(X_i)$$

donc on pourrait considérer l'erreur d'approximation comme critère d'évalution de $\hat{r}$:

$$\text{MCR}(\hat{r}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}(X_i))^2 \quad \text{(Moyenne des Carrés des Résidus)}$$

- Constat : Pour les estimateurs des MC, $\text{MCR}(\hat{r}_D)$ diminue automatiquement quand S_D augmente : :

$$\hat{r}^D = \arg\min_{t \in S_D} \frac{1}{n}\sum_{i=1}^{n}(t(X_i) - Y_i)^2 \quad \Rightarrow \quad MCR(\hat{r}^D) = \min_{t \in S_D} \frac{1}{n}\sum_{i=1}^{n}(t(X_i) - Y_i)^2$$

$$\text{donc} \quad S_D \subset S_{D'} \quad \Rightarrow \quad \text{MCR}(\hat{r}_D) \geq \text{MCR}(\hat{r}_{D'})$$

L'erreur d'approximation, un risque pertinent ? (2)

- Cas extrême : $D = n$. Considérons la base trigo :

$$\dim(S_D(\mathbf{X})) = D = n \quad \Rightarrow \quad S_D(\mathbf{X}) = \mathbb{R}^n$$
$$\Rightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in \mathbb{R}^n} \|\mathbf{Y} - U\|_{\ell^2}^2 = \mathbf{Y}$$

D'où $MCR(\hat{r}_D) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_D(X_i))^2 = 0$. Or $\hat{r}_D \neq r$.

- Exple (n=15)

Estimateurs MC sur la base trigo pour $D = 6$ et 15.

- $\hat{r}_6$ est un meilleur estimateur que $\hat{r}_{15}$
- $MCR(\hat{r}_6) > MCR(\hat{r}_{15}) = 0$

- Conclusion : MCR n'est pas un indicateur de la qualité d'estimation

↪ considérer MCR($\hat{r}$) comme critère conduit à un overfit.

- D'une manière générale, le modèle qui approche le mieux les données est rarement le meilleur modèle

L'erreur d'approximation, un risque pertinent ? (2)

- Cas extrême : $D = n$. Considérons la base trigo :

$$\dim(S_D(\mathbf{X})) = D = n \quad \Rightarrow \quad S_D(\mathbf{X}) = \mathbb{R}^n$$
$$\Rightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in \mathbb{R}^n} \|\mathbf{Y} - U\|_{\ell^2}^2 = \mathbf{Y}$$

D'où $MCR(\hat{r}_D) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_D(X_i))^2 = 0$. Or $\hat{r}_D \neq r$.

- Exple (n=15)

Estimateurs MC sur la base trigo pour $D = 6$ et 15.

- $\hat{r}_6$ est un meilleur estimateur que $\hat{r}_{15}$
- $MCR(\hat{r}_6) > MCR(\hat{r}_{15}) = 0$

- Conclusion : MCR n'est pas un indicateur de la qualité d'estimation

↪ considérer MCR($\hat{r}$) comme critère conduit à un overfit.

- D'une manière générale, le modèle qui approche le mieux les données est rarement le meilleur modèle

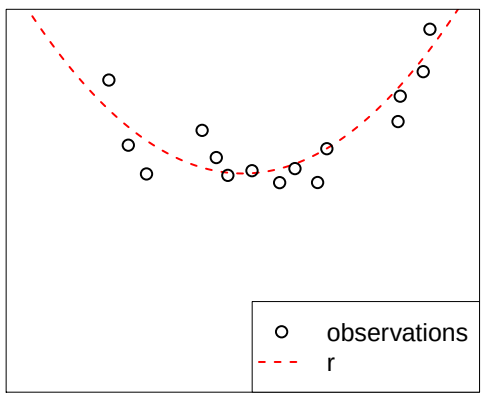
L'erreur d'approximation, un risque pertinent ? (2)

- Cas extrême : $D = n$. Considérons la base trigo :

$$\dim(S_D(\mathbf{X})) = D = n \quad \Rightarrow \quad S_D(\mathbf{X}) = \mathbb{R}^n$$
$$\Rightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in \mathbb{R}^n} \|\mathbf{Y} - U\|^2_{\ell^2} = \mathbf{Y}$$

D'où $MCR(\hat{r}_D) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_D(X_i))^2 = 0$. Or $\hat{r}_D \neq r$.

- Exple (n=15)



Estimateurs MC sur la base trigo pour $D = 6$ et 15.

- $\hat{r}_6$ est un meilleur estimateur que $\hat{r}_{15}$
- $MCR(\hat{r}_6) > MCR(\hat{r}_{15}) = 0$

- Conclusion : MCR n'est pas un indicateur de la qualité d'estimation

↪ considérer MCR($\hat{r}$) comme critère conduit à un overfit.

- D'une manière générale, le modèle qui approche le mieux les données est rarement le meilleur modèle

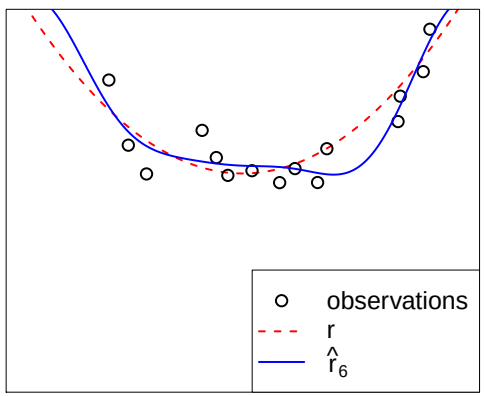
L'erreur d'approximation, un risque pertinent ? (2)

- Cas extrême : $D = n$. Considérons la base trigo :

$$\dim(S_D(\mathbf{X})) = D = n \quad \Rightarrow \quad S_D(\mathbf{X}) = \mathbb{R}^n$$
$$\Rightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in \mathbb{R}^n} \|\mathbf{Y} - U\|^2_{\ell^2} = \mathbf{Y}$$

D'où $MCR(\hat{r}_D) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_D(X_i))^2 = 0$. Or $\hat{r}_D \neq r$.

- Exple (n=15)



Estimateurs MC sur la base trigo pour $D = 6$ et 15.

- $\hat{r}_6$ est un meilleur estimateur que $\hat{r}_{15}$
- $MCR(\hat{r}_6) > MCR(\hat{r}_{15}) = 0$

- Conclusion : MCR n'est pas un indicateur de la qualité d'estimation

↪ considérer MCR($\hat{r}$) comme critère conduit à un overfit.

- D'une manière générale, le modèle qui approche le mieux les données est rarement le meilleur modèle

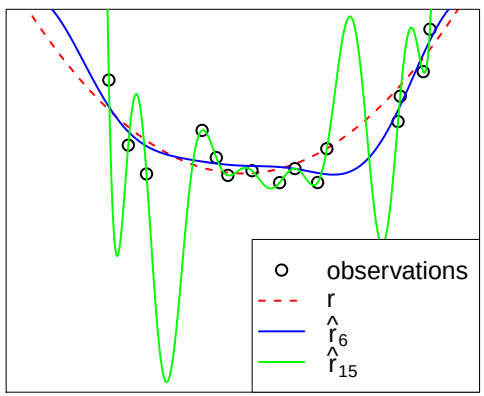
L'erreur d'approximation, un risque pertinent ? (2)

- Cas extrême : $D = n$. Considérons la base trigo :

$$\dim(S_D(\mathbf{X})) = D = n \quad \Rightarrow \quad S_D(\mathbf{X}) = \mathbb{R}^n$$
$$\Rightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in \mathbb{R}^n} \|\mathbf{Y} - U\|^2_{\ell^2} = \mathbf{Y}$$

D'où $MCR(\hat{r}_D) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_D(X_i))^2 = 0$. Or $\hat{r}_D \neq r$.

- Exple (n=15)



Estimateurs MC sur la base trigo pour $D = 6$ et 15.

- $\hat{r}_6$ est un meilleur estimateur que $\hat{r}_{15}$
- $MCR(\hat{r}_6) > MCR(\hat{r}_{15}) = 0$

- Conclusion : MCR n'est pas un indicateur de la qualité d'estimation

↪ considérer MCR($\hat{r}$) comme critère conduit à un overfit.

- D'une manière générale, le modèle qui approche le mieux les données est rarement le meilleur modèle

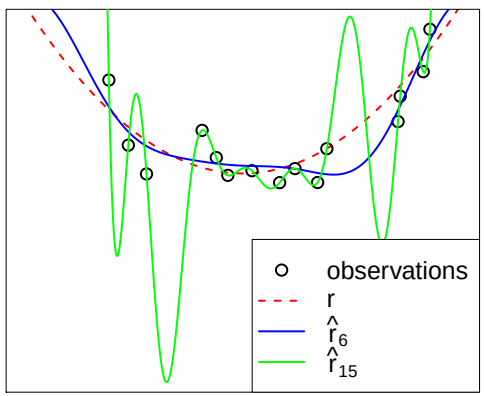
L'erreur d'approximation, un risque pertinent ? (2)

- Cas extrême : $D = n$. Considérons la base trigo :

$$\dim(S_D(\mathbf{X})) = D = n \quad \Rightarrow \quad S_D(\mathbf{X}) = \mathbb{R}^n$$
$$\Rightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in \mathbb{R}^n} \|\mathbf{Y} - U\|^2_{\ell^2} = \mathbf{Y}$$

D'où $MCR(\hat{r}_D) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_D(X_i))^2 = 0$. Or $\hat{r}_D \neq r$.

- Exple (n=15)



Estimateurs MC sur la base trigo pour $D = 6$ et 15.

- $\hat{r}_6$ est un meilleur estimateur que $\hat{r}_{15}$
- $MCR(\hat{r}_6) > MCR(\hat{r}_{15}) = 0$

- Conclusion : MCR n'est pas un indicateur de la qualité d'estimation

↪ considérer MCR($\hat{r}$) comme critère conduit à un overfit.

- D'une manière générale, le modèle qui approche le mieux les données est rarement le meilleur modèle

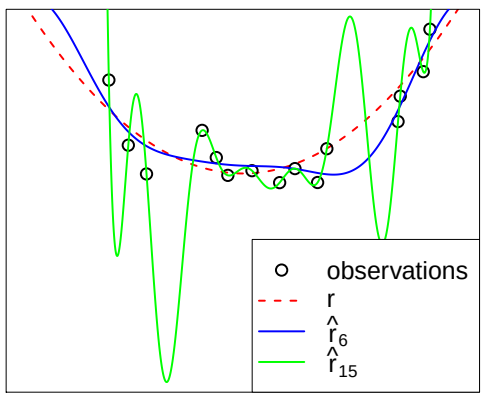
L'erreur d'approximation, un risque pertinent ? (2)

- Cas extrême : $D = n$. Considérons la base trigo :

$$\dim(S_D(\mathbf{X})) = D = n \quad \Rightarrow \quad S_D(\mathbf{X}) = \mathbb{R}^n$$
$$\Rightarrow \quad \hat{r}_D(\mathbf{X}) = \arg\min_{U \in \mathbb{R}^n} \|\mathbf{Y} - U\|^2_{\ell^2} = \mathbf{Y}$$

D'où $MCR(\hat{r}_D) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_D(X_i))^2 = 0$. Or $\hat{r}_D \neq r$.

- Exple (n=15)



Estimateurs MC sur la base trigo pour $D = 6$ et 15.

- $\hat{r}_6$ est un meilleur estimateur que $\hat{r}_{15}$
- $MCR(\hat{r}_6) > MCR(\hat{r}_{15}) = 0$

- Conclusion : MCR n'est pas un indicateur de la qualité d'estimation

↪ considérer MCR($\hat{r}$) comme critère conduit à un overfit.

- D'une manière générale, le modèle qui approche le mieux les données est rarement le meilleur modèle

Risques considérés

Risque empirique

- Pour $\mathbf{X}$ donné, on définit la norme emprique pour tout $t : I \to \mathbb{R}$:

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) \ = \frac{1}{n}\|t(\mathbf{X})\|_{\ell^2}^2$$

- On définit le risque empirique d'un estimateur $\hat{r}$:

$$\mathbb{E}\left[\|\hat{r} - r\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

$\hookrightarrow$ Contrôle l'erreur d'estimation aux points du design.

Risque intégré (MISE)

- On définit le MISE d'un estimateur $\hat{r}$:

$$\mathbb{E}\left[\|\hat{r} - r\|_{f_X}^2\right] \quad \text{où} \quad \|t\|_{f_X}^2 = \int t(x)^2 f_X(x)\,dx = \mathbb{E}\left[\|t\|_n^2\right]$$

$\hookrightarrow$ Le MISE contrôle plus fortement l'erreur d'estimation aux points où f_X est grand.

Risques considérés

Risque empirique

- Pour $\mathbf{X}$ donné, on définit la norme emprique pour tout $t : I \to \mathbb{R}$:

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) \ = \frac{1}{n}\|t(\mathbf{X})\|_{\ell^2}^2$$

- On définit le risque empirique d'un estimateur $\hat{r}$:

$$\mathbb{E}\left[\|\hat{r} - r\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

$\hookrightarrow$ Contrôle l'erreur d'estimation aux points du design.

Risque intégré (MISE)

- On définit le MISE d'un estimateur $\hat{r}$:

$$\mathbb{E}\left[\|\hat{r} - r\|_{f_X}^2\right] \quad \text{où} \quad \|t\|_{f_X}^2 = \int t(x)^2 f_X(x)\,dx = \mathbb{E}\left[\|t\|_n^2\right]$$

$\hookrightarrow$ Le MISE contrôle plus fortement l'erreur d'estimation aux points où f_X est grand.

Risques considérés

Risque empirique

- Pour $\mathbf{X}$ donné, on définit la norme emprique pour tout $t : I \to \mathbb{R}$:
$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) \ = \frac{1}{n}\|t(\mathbf{X})\|_{\ell^2}^2$$
- On définit le risque empirique d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

$\hookrightarrow$ Contrôle l'erreur d'estimation aux points du design.

Risque intégré (MISE)

- On définit le MISE d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_{f_X}^2\right] \quad \text{où} \quad \|t\|_{f_X}^2 = \int t(x)^2 f_X(x)\,dx = \mathbb{E}\left[\|t\|_n^2\right]$$

$\hookrightarrow$ Le MISE contrôle plus fortement l'erreur d'estimation aux points où f_X est grand.

Risques considérés

Risque empirique

- Pour $\mathbf{X}$ donné, on définit la norme emprique pour tout $t : I \to \mathbb{R}$:
$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) \ = \frac{1}{n}\|t(\mathbf{X})\|_{\ell^2}^2$$
- On définit le risque empirique d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

$\hookrightarrow$ Contrôle l'erreur d'estimation aux points du design.

Risque intégré (MISE)

- On définit le MISE d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_{f_X}^2\right] \quad \text{où} \quad \|t\|_{f_X}^2 = \int t(x)^2 f_X(x)\,dx = \mathbb{E}\left[\|t\|_n^2\right]$$

$\hookrightarrow$ Le MISE contrôle plus fortement l'erreur d'estimation aux points où f_X est grand.

Risques considérés

Risque empirique

- Pour $\mathbf{X}$ donné, on définit la norme emprique pour tout $t : I \to \mathbb{R}$:
$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) \ = \frac{1}{n}\|t(\mathbf{X})\|_{\ell^2}^2$$
- On définit le risque empirique d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

$\hookrightarrow$ Contrôle l'erreur d'estimation aux points du design.

Risque intégré (MISE)

- On définit le MISE d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_{f_X}^2\right] \quad \text{où} \quad \|t\|_{f_X}^2 = \int t(x)^2 f_X(x)\,dx = \mathbb{E}\left[\|t\|_n^2\right]$$

$\hookrightarrow$ Le MISE contrôle plus fortement l'erreur d'estimation aux points où f_X est grand.

Risques considérés

Risque empirique

- Pour $\mathbf{X}$ donné, on définit la norme emprique pour tout $t : I \to \mathbb{R}$:
$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) \ = \frac{1}{n}\|t(\mathbf{X})\|_{\ell^2}^2$$
- On définit le risque empirique d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

$\hookrightarrow$ Contrôle l'erreur d'estimation aux points du design.

Risque intégré (MISE)

- On définit le MISE d'un estimateur $\hat{r}$:
$$\mathbb{E}\left[\|\hat{r} - r\|_{f_X}^2\right] \quad \text{où} \quad \|t\|_{f_X}^2 = \int t(x)^2 f_X(x)\,dx = \mathbb{E}\left[\|t\|_n^2\right]$$

$\hookrightarrow$ Le MISE contrôle plus fortement l'erreur d'estimation aux points où f_X est grand.

Majoration du risque empirique

- On va d'abord étudier pour tout $\mathbf{X}$

$$\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}\right]$$

- Ensuite, comme $\mathbb{E}\left[\mathbb{E}[U|V]\right] = \mathbb{E}[U]$:

$$\mathbb{E}\left[\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}\right]\right] = \mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

- **Remarque :** En fait, la majoration de

$$\frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}\right] = \mathbb{E}\left[\|\hat{r} - r\|_n^2 | \mathbf{X}\right]$$

fournit des résultats en fix-design.

Majoration du risque empirique

- On va d'abord étudier pour tout $\mathbf{X}$

$$\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}\right]$$

- Ensuite, comme $\mathbb{E}\left[\mathbb{E}[U|V]\right] = \mathbb{E}[U]$:

$$\mathbb{E}\left[\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}\right]\right] = \mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

- **Remarque :** En fait, la majoration de

$$\frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}\right] = \mathbb{E}\left[\|\hat{r} - r\|_n^2 | \mathbf{X}\right]$$

fournit des résultats en fix-design.

Majoration du risque empirique

- On va d'abord étudier pour tout $\mathbf{X}$

$$\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 |\mathbf{X}\right]$$

- Ensuite, comme $\mathbb{E}\left[\mathbb{E}[U\,|\,V]\right] = \mathbb{E}[U]$:

$$\mathbb{E}\left[\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 |\mathbf{X}\right]\right] = \mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2\right]$$

- **Remarque :** En fait, la majoration de

$$\frac{1}{n}\mathbb{E}\left[\|\hat{r}(\mathbf{X}) - r(\mathbf{X})\|_{\ell^2}^2 |\mathbf{X}\right] = \mathbb{E}\left[\|\hat{r} - r\|_n^2 |\mathbf{X}\right]$$

fournit des résultats en fix-design.

Majoration du risque empirique (2)

Soit $\mathbf{X}$ fixé.

- Soit $r_D(\mathbf{X}) = \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))$, alors d'après Pythagore :

$$\|r(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 = \|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 + \|r_D(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2$$

$$\begin{aligned}
\mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}] &= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\mathbf{Y})|\mathbf{X}] \\
&= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(r(\mathbf{X}) + \varepsilon)|\mathbf{X}] \\
&= \Pi_{S_D(\mathbf{X})}(r(\mathbf{X})) + \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\varepsilon)|\mathbf{X}] \\
&= r_D(\mathbf{X}) + \Pi_{S_D(\mathbf{X})}\mathbb{E}[\varepsilon] \quad \text{par linéarité des projecteurs} \\
&= r_D(\mathbf{X})
\end{aligned}$$

- **Rq** En définissant $r_D(\mathbf{X}) = \mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}]$ et en utilisant une décomposition biais-variance avec double produit nul on retrouve le résultat.

Majoration du risque empirique (2)

Soit $\mathbf{X}$ fixé.

- Soit $r_D(\mathbf{X}) = \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))$, alors d'après Pythagore :

$$\mathbb{E}[\|r(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 + \mathbb{E}[\|r_D(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}]$$

$$\begin{aligned}
\mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}] &= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\mathbf{Y})|\mathbf{X}] \\
&= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(r(\mathbf{X}) + \varepsilon)|\mathbf{X}] \\
&= \Pi_{S_D(\mathbf{X})}(r(\mathbf{X})) + \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\varepsilon)|\mathbf{X}] \\
&= r_D(\mathbf{X}) + \Pi_{S_D(\mathbf{X})}\mathbb{E}[\varepsilon] \quad \text{par linéarité des projecteurs} \\
&= r_D(\mathbf{X})
\end{aligned}$$

- **Rq** En définissant $r_D(\mathbf{X}) = \mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}]$ et en utilisant une décomposition biais-variance avec double produit nul on retrouve le résultat.

Majoration du risque empirique (2)

Soit $\mathbf{X}$ fixé.

- Soit $r_D(\mathbf{X}) = \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))$, alors d'après Pythagore :

$$\mathbb{E}[\|r(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 + \mathbb{E}[\|r_D(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}]$$

$$\begin{aligned}
\mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}] &= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\mathbf{Y})|\mathbf{X}] \\
&= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(r(\mathbf{X}) + \varepsilon)|\mathbf{X}] \\
&= \Pi_{S_D(\mathbf{X})}(r(\mathbf{X})) + \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\varepsilon)|\mathbf{X}] \\
&= r_D(\mathbf{X}) + \Pi_{S_D(\mathbf{X})}\mathbb{E}[\varepsilon] \quad \text{par linéarité des projecteurs} \\
&= r_D(\mathbf{X})
\end{aligned}$$

- **Rq** En définissant $r_D(\mathbf{X}) = \mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}]$ et en utilisant une décomposition biais-variance avec double produit nul on retrouve le résultat.

Majoration du risque empirique (2)

Soit $\mathbf{X}$ fixé.

- Soit $r_D(\mathbf{X}) = \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))$, alors d'après Pythagore :

$$\mathbb{E}[\|r(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \underbrace{\|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2}_{\text{n*biais}} + \underbrace{\mathbb{E}[\|r_D(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}]}_{\text{n*variance}}$$

$$\begin{aligned}
\mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}] &= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\mathbf{Y})|\mathbf{X}] \\
&= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(r(\mathbf{X}) + \varepsilon)|\mathbf{X}] \\
&= \Pi_{S_D(\mathbf{X})}(r(\mathbf{X})) + \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\varepsilon)|\mathbf{X}] \\
&= r_D(\mathbf{X}) + \Pi_{S_D(\mathbf{X})}\mathbb{E}[\varepsilon] \quad \text{par linéarité des projecteurs} \\
&= r_D(\mathbf{X})
\end{aligned}$$

- **Rq** En définissant $r_D(\mathbf{X}) = \mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}]$ et en utilisant une décomposition biais-variance avec double produit nul on retrouve le résultat.

Majoration du risque empirique (2)

Soit $\mathbf{X}$ fixé.

- Soit $r_D(\mathbf{X}) = \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))$, alors d'après Pythagore :

$$\mathbb{E}[\|r(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \underbrace{\|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2}_{\text{n*biais}} + \underbrace{\mathbb{E}[\|r_D(\mathbf{X}) - \hat{r}_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}]}_{\text{n*variance}}$$

$$\begin{aligned}
\mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}] &= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\mathbf{Y})|\mathbf{X}] \\
&= \mathbb{E}[\Pi_{S_D(\mathbf{X})}(r(\mathbf{X}) + \varepsilon)|\mathbf{X}] \\
&= \Pi_{S_D(\mathbf{X})}(r(\mathbf{X})) + \mathbb{E}[\Pi_{S_D(\mathbf{X})}(\varepsilon)|\mathbf{X}] \\
&= r_D(\mathbf{X}) + \Pi_{S_D(\mathbf{X})}\mathbb{E}[\varepsilon] \quad \text{par linéarité des projecteurs} \\
&= r_D(\mathbf{X})
\end{aligned}$$

- **Rq** En définissant $r_D(\mathbf{X}) = \mathbb{E}[\hat{r}_D(\mathbf{X})|\mathbf{X}]$ et en utilisant une décomposition biais-variance avec double produit nul on retrouve le résultat.

Biais pour le risque empirique

- Par définition de $r_D(\mathbf{X})$:

$$\begin{aligned}
\|r - r_D\|_n^2 &= \frac{1}{n}\|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 \\
&= \frac{1}{n}\|r(\mathbf{X}) - \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))\|_{\ell^2}^2 \\
&= \inf_{U \in S_D(\mathbf{X})} \frac{1}{n}\|r(\mathbf{X}) - U\|_{\ell^2}^2 = \inf_{t \in S_D} \|r - t\|_n^2
\end{aligned}$$

- En intégrant sur $\mathbf{X}$ (comme $\mathbb{E}[\inf(.)] \leq \inf(\mathbb{E}[.])$)

$$\text{Biais} = \mathbb{E}[\|r - r_D\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2$$

- Si $f_X(x) > 0, \quad \forall x \in I$, alors $\|.\|_{f_X}$ est une norme et le biais est égal à la distance de r à l'espace S_D pour la norme $\|.\|_{f_X}$
- Si les modèles $\{S_D\}$ sont emboités i.e. $S_1 \subset S_2 \subset \ldots$, alors le biais $\searrow$ quand D $\nearrow$.

$\hookrightarrow$ Exple : modèles trigonométriques.

Biais pour le risque empirique

- Par définition de $r_D(\mathbf{X})$:

$$\begin{aligned}\|r - r_D\|_n^2 &= \frac{1}{n}\|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 \\ &= \frac{1}{n}\|r(\mathbf{X}) - \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))\|_{\ell^2}^2 \\ &= \inf_{U \in S_D(\mathbf{X})} \frac{1}{n}\|r(\mathbf{X}) - U\|_{\ell^2}^2 = \inf_{t \in S_D} \|r - t\|_n^2\end{aligned}$$

- En intégrant sur $\mathbf{X}$ (comme $\mathbb{E}[\inf(.)] \leq \inf(\mathbb{E}[.])$)

$$\text{Biais} = \mathbb{E}[\|r - r_D\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2$$

- Si $f_X(x) > 0, \quad \forall x \in I$, alors $\|.\|_{f_X}$ est une norme et le biais est égal à la distance de r à l'espace S_D pour la norme $\|.\|_{f_X}$
- Si les modèles $\{S_D\}$ sont emboités i.e. $S_1 \subset S_2 \subset \ldots$, alors le biais $\searrow$ quand D $\nearrow$.

$\hookrightarrow$ Exple : modèles trigonométriques.

Biais pour le risque empirique

- Par définition de $r_D(\mathbf{X})$:

$$\begin{aligned}\|r - r_D\|_n^2 &= \frac{1}{n}\|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 \\ &= \frac{1}{n}\|r(\mathbf{X}) - \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))\|_{\ell^2}^2 \\ &= \inf_{U \in S_D(\mathbf{X})} \frac{1}{n}\|r(\mathbf{X}) - U\|_{\ell^2}^2 = \inf_{t \in S_D} \|r - t\|_n^2\end{aligned}$$

- En intégrant sur $\mathbf{X}$ (comme $\mathbb{E}[\inf(.)] \leq \inf(\mathbb{E}[.])$)

$$\text{Biais} = \mathbb{E}[\|r - r_D\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2$$

- Si $f_X(x) > 0, \quad \forall x \in I$, alors $\|.\|_{f_X}$ est une norme et le biais est égal à la distance de r à l'espace S_D pour la norme $\|.\|_{f_X}$
- Si les modèles $\{S_D\}$ sont emboités i.e. $S_1 \subset S_2 \subset \ldots$, alors le biais $\searrow$ quand D $\nearrow$.

↪ Exple : modèles trigonométriques.

Biais pour le risque empirique

- Par définition de $r_D(\mathbf{X})$:

$$\begin{aligned}\|r - r_D\|_n^2 &= \frac{1}{n}\|r(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 \\ &= \frac{1}{n}\|r(\mathbf{X}) - \Pi_{S_D(\mathbf{X})}(r(\mathbf{X}))\|_{\ell^2}^2 \\ &= \inf_{U \in S_D(\mathbf{X})} \frac{1}{n}\|r(\mathbf{X}) - U\|_{\ell^2}^2 = \inf_{t \in S_D} \|r - t\|_n^2\end{aligned}$$

- En intégrant sur $\mathbf{X}$ (comme $\mathbb{E}[\inf(.)] \leq \inf(\mathbb{E}[.])$)

$$\text{Biais} = \mathbb{E}[\|r - r_D\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2$$

- Si $f_X(x) > 0, \quad \forall x \in I$, alors $\|.\|_{f_X}$ est une norme et le biais est égal à la distance de r à l'espace S_D pour la norme $\|.\|_{f_X}$
- Si les modèles $\{S_D\}$ sont emboités i.e. $S_1 \subset S_2 \subset \ldots$, alors le biais $\searrow$ quand D $\nearrow$.

$\hookrightarrow$ Exple : modèles trigonométriques.

Variance pour le risque empirique

- Par définition de $\hat{r}_D$ et de $r_D(\mathbf{X})$:

$$\mathbb{E}[\|\hat{r}_D(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}(\mathbf{Y} - r(\mathbf{X}))\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}]$$

- De plus, $\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 = \varepsilon^t \Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})}\varepsilon$. Or, $\Pi_{S_D(\mathbf{X})}$ est un projecteur donc $\Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})} = \Pi_{S_D(\mathbf{X})}^2 = \Pi_{S_D(\mathbf{X})}$. D'où

$$\mathbb{E}\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}\left[\sum_{i,i'=1}^{n} \varepsilon_i \varepsilon_{i'} (\Pi_{S_D(\mathbf{X})})_{i,i'} | \mathbf{X}\right] = \sum_{i,i'=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i'} \mathbb{E}[\varepsilon_i \varepsilon_{i'}]$$

$$= \sum_{i=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i} \times \sigma^2 = \sigma^2 \mathrm{Tr}(\Pi_{S_D(\mathbf{X})})$$

- **Rappel :** Soit P projecteur orthog. de $\mathbb{R}^n$ sur un sev S, en notant également P sa matrice dans une base b de $\mathbb{R}^n$, alors $\mathrm{Tr}(P) = \dim(\mathrm{S})$ (qq soit b).
- D'où

$$\text{Variance} = \frac{1}{n}\mathbb{E}[\|\hat{r}_D - r_D\|_{\ell^2}^2] \leq \sigma^2 \frac{D}{n}$$

Variance pour le risque empirique

- Par définition de $\hat{r}_D$ et de $r_D(\mathbf{X})$:

$$\mathbb{E}[\|\hat{r}_D(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}(\mathbf{Y} - r(\mathbf{X}))\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}]$$

- De plus, $\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 = \varepsilon^t \Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})}\varepsilon$. Or, $\Pi_{S_D(\mathbf{X})}$ est un projecteur donc $\Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})} = \Pi_{S_D(\mathbf{X})}^2 = \Pi_{S_D(\mathbf{X})}$. D'où

$$\mathbb{E}\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}\left[\sum_{i,i'=1}^{n} \varepsilon_i \varepsilon_{i'} (\Pi_{S_D(\mathbf{X})})_{i,i'} | \mathbf{X}\right] = \sum_{i,i'=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i'} \mathbb{E}[\varepsilon_i \varepsilon_{i'}]$$

$$= \sum_{i=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i} \times \sigma^2 = \sigma^2 \mathrm{Tr}(\Pi_{S_D(\mathbf{X})})$$

- **Rappel :** Soit P projecteur orthog. de $\mathbb{R}^n$ sur un sev S, en notant également P sa matrice dans une base b de $\mathbb{R}^n$, alors $\mathrm{Tr}(P) = \dim(S)$ (qq soit b).
- D'où

$$\text{Variance} = \frac{1}{n}\mathbb{E}[\|\hat{r}_D - r_D\|_{\ell^2}^2] \leq \sigma^2 \frac{D}{n}$$

Variance pour le risque empirique

- Par définition de $\hat{r}_D$ et de $r_D(\mathbf{X})$:

$$\mathbb{E}[\|\hat{r}_D(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}(\mathbf{Y} - r(\mathbf{X}))\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}]$$

- De plus, $\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 = \varepsilon^t \Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})}\varepsilon$. Or, $\Pi_{S_D(\mathbf{X})}$ est un projecteur donc $\Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})} = \Pi_{S_D(\mathbf{X})}^2 = \Pi_{S_D(\mathbf{X})}$. D'où

$$\mathbb{E}\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}\left[\sum_{i,i'=1}^{n} \varepsilon_i \varepsilon_{i'} (\Pi_{S_D(\mathbf{X})})_{i,i'} | \mathbf{X}\right] = \sum_{i,i'=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i'} \mathbb{E}[\varepsilon_i \varepsilon_{i'}]$$

$$= \sum_{i=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i} \times \sigma^2 = \sigma^2 \mathrm{Tr}(\Pi_{S_D(\mathbf{X})})$$

- **Rappel :** Soit P projecteur orthog. de $\mathbb{R}^n$ sur un sev S, en notant également P sa matrice dans une base b de $\mathbb{R}^n$, alors $\mathrm{Tr}(P) = \dim(\mathrm{S})$ (qq soit b).
- D'où

$$\text{Variance} = \frac{1}{n}\mathbb{E}[\|\hat{r}_D - r_D\|_{\ell^2}^2] \leq \sigma^2 \frac{D}{n}$$

Variance pour le risque empirique

- Par définition de $\hat{r}_D$ et de $r_D(\mathbf{X})$:

$$\mathbb{E}[\|\hat{r}_D(\mathbf{X}) - r_D(\mathbf{X})\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}(\mathbf{Y} - r(\mathbf{X}))\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}[\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}]$$

- De plus, $\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 = \varepsilon^t \Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})}\varepsilon$. Or, $\Pi_{S_D(\mathbf{X})}$ est un projecteur donc $\Pi_{S_D(\mathbf{X})}^t \Pi_{S_D(\mathbf{X})} = \Pi_{S_D(\mathbf{X})}^2 = \Pi_{S_D(\mathbf{X})}$. D'où

$$\mathbb{E}\|\Pi_{S_D(\mathbf{X})}\varepsilon\|_{\ell^2}^2 | \mathbf{X}] = \mathbb{E}\left[\sum_{i,i'=1}^{n} \varepsilon_i \varepsilon_{i'} (\Pi_{S_D(\mathbf{X})})_{i,i'} | \mathbf{X}\right] = \sum_{i,i'=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i'} \mathbb{E}[\varepsilon_i \varepsilon_{i'}]$$

$$= \sum_{i=1}^{n} (\Pi_{S_D(\mathbf{X})})_{i,i} \times \sigma^2 = \sigma^2 \mathrm{Tr}(\Pi_{S_D(\mathbf{X})})$$

- **Rappel :** Soit P projecteur orthog. de $\mathbb{R}^n$ sur un sev S, en notant également P sa matrice dans une base b de $\mathbb{R}^n$, alors $\mathrm{Tr}(P) = \dim(\mathrm{S})$ (qq soit b).
- D'où

$$\text{Variance} = \frac{1}{n}\mathbb{E}[\|\hat{r}_D - r_D\|_{\ell^2}^2] \leq \sigma^2 \frac{D}{n}$$

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Majoration du risque empirique : conclusion

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2] \leq \inf_{t \in S_D} \|r - t\|_{f_X}^2 + \sigma^2 \frac{D}{n}$$

- Si $f_X > 0$ sur I, le biais $\inf_{t \in S_D} \|r - t\|_{f_X}^2$ quantifie la distance entre r et S_D pour la norme $\|.\|_{f_X}$.
- Si les modèles sont emboités, le biais diminue avec D
- Si r est de régularité α, le biais est majoré par $cte \times D^{-2\alpha}$
- La variance $\sigma^2 D/n$ augmente avec D.
- Comme en densité, il existe des procédures de sélection de modèles.

Remarque : décomposition biais-variance en fix-design

$$\mathbb{E}[\|\hat{r}_D - r\|_n^2 | \mathbf{X}] \leq \inf_{t \in S_D} \|t - r\|_n^2 + \sigma^2 \frac{D}{n}$$

- Le biais dépend de $\mathbf{X}$: la majoration sur des espaces de régularité nécéssite des hypothèses supplémentaires sur $\mathbf{X}$ (par ex : $(X_1, \ldots, X_n)$ régulièrement espacés).

Passage du risque empirique au MISE

- Pour toute fonction t, $\mathbb{E}[\|t\|_n^2] = \|t\|_{f_X}^2$ mais :

$$\text{MISE} = \mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] \neq \mathbb{E}[\|\hat{r}_D - r\|_n^2]$$

car $\hat{r}_n$ et $\|\cdot\|_n$ dépendent du même échantillon $\mathbf{X}$.

- Heuristique : on montre que pour tout $0 < a < 1$, sur un espace Ω_n de grande probabilité,

$$\|\hat{r}_D - r\|_{f_X}^2 \leq (1+a)\|\hat{r}_D - r\|_n^2$$

- D'où

$$\begin{aligned}\mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] &\leq (1+a)\mathbb{E}[\|\hat{r}_D - r\|_n^2] + \text{reste} \\ &\leq (1+a)\left\{\inf_{t \in S_D} \|t - r\|_{f_X}^2 + \sigma^2 \frac{D}{n}\right\} + \text{reste}\end{aligned}$$

↪ En pratique, la démonstration est complexe !

Passage du risque empirique au MISE

- Pour toute fonction t, $\mathbb{E}[\|t\|_n^2] = \|t\|_{f_X}^2$ mais :

$$\text{MISE} = \mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] \neq \mathbb{E}[\|\hat{r}_D - r\|_n^2]$$

car $\hat{r}_n$ et $\|\cdot\|_n$ dépendent du même échantillon $\mathbf{X}$.

- Heuristique : on montre que pour tout $0 < a < 1$, sur un espace Ω_n de grande probabilité,

$$\|\hat{r}_D - r\|_{f_X}^2 \leq (1+a)\|\hat{r}_D - r\|_n^2$$

- D'où

$$\begin{aligned}\mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] &\leq (1+a)\mathbb{E}[\|\hat{r}_D - r\|_n^2] + \text{reste} \\ &\leq (1+a)\left\{\inf_{t \in S_D} \|t - r\|_{f_X}^2 + \sigma^2 \frac{D}{n}\right\} + \text{reste}\end{aligned}$$

↪ En pratique, la démonstration est complexe !

Passage du risque empirique au MISE

- Pour toute fonction t, $\mathbb{E}[\|t\|_n^2] = \|t\|_{f_X}^2$ mais :

$$\text{MISE} = \mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] \neq \mathbb{E}[\|\hat{r}_D - r\|_n^2]$$

car $\hat{r}_n$ et $\|\cdot\|_n$ dépendent du même échantillon $\mathbf{X}$.

- Heuristique : on montre que pour tout $0 < a < 1$, sur un espace Ω_n de grande probabilité,

$$\|\hat{r}_D - r\|_{f_X}^2 \leq (1+a)\|\hat{r}_D - r\|_n^2$$

- D'où

$$\begin{aligned}
\mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] &\leq (1+a)\mathbb{E}[\|\hat{r}_D - r\|_n^2] + \text{reste} \\
&\leq (1+a)\left\{\inf_{t \in S_D} \|t - r\|_{f_X}^2 + \sigma^2 \frac{D}{n}\right\} + \text{reste}
\end{aligned}$$

↪ En pratique, la démonstration est complexe !

Passage du risque empirique au MISE

- Pour toute fonction t, $\mathbb{E}[\|t\|_n^2] = \|t\|_{f_X}^2$ mais :

$$\text{MISE} = \mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] \neq \mathbb{E}[\|\hat{r}_D - r\|_n^2]$$

car $\hat{r}_n$ et $\|\cdot\|_n$ dépendent du même échantillon $\mathbf{X}$.

- Heuristique : on montre que pour tout $0 < a < 1$, sur un espace Ω_n de grande probabilité,

$$\|\hat{r}_D - r\|_{f_X}^2 \leq (1+a)\|\hat{r}_D - r\|_n^2$$

- D'où

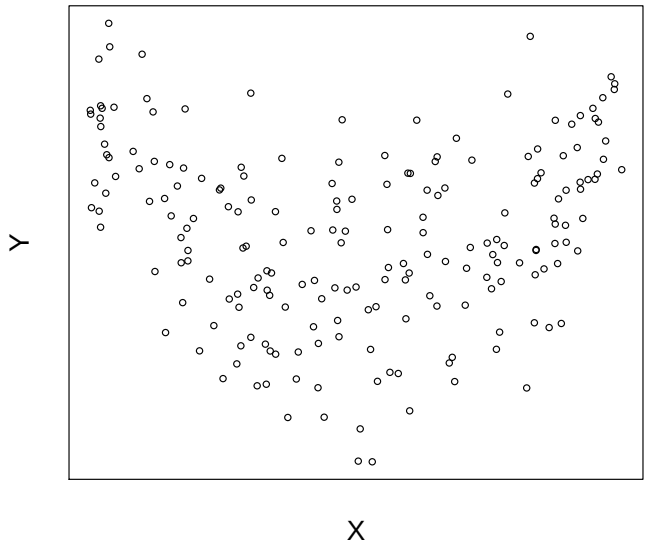
$$\begin{aligned}
\mathbb{E}[\|\hat{r}_D - r\|_{f_X}^2] &\leq (1+a)\mathbb{E}[\|\hat{r}_D - r\|_n^2] + \text{reste} \\
&\leq (1+a)\left\{\inf_{t \in S_D} \|t - r\|_{f_X}^2 + \sigma^2 \frac{D}{n}\right\} + \text{reste}
\end{aligned}$$

↪ En pratique, la démonstration est complexe !

Quelques autres méthodes

- Méthodes par noyaux : Nadaraya-Watson
- Régression par polynômes locaux
- Méthode des k plus proches voisins.

Méthode par noyaux de Nadaraya-Watson



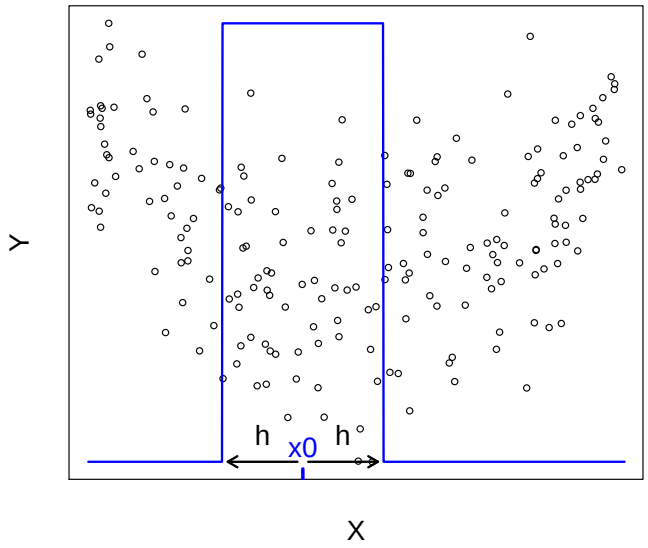
- Noyau rectangulaire

$$\begin{aligned}\hat{r}_h(x_0) &= \text{mean}\{Y_i, X_i \in [x_0-h, x_0+h]\} \\ &= \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in [x_0-h, x_0+h])}{\sum_{i=1}^n \mathbb{1}(X_i \in [x_0-h, x_0+h])}\end{aligned}$$

- Cas général : noyau K

$$\hat{r}_h(x_0) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x_0)}{\sum_{i=1}^n K_h(X_i - x_0)}$$

Méthode par noyaux de Nadaraya-Watson



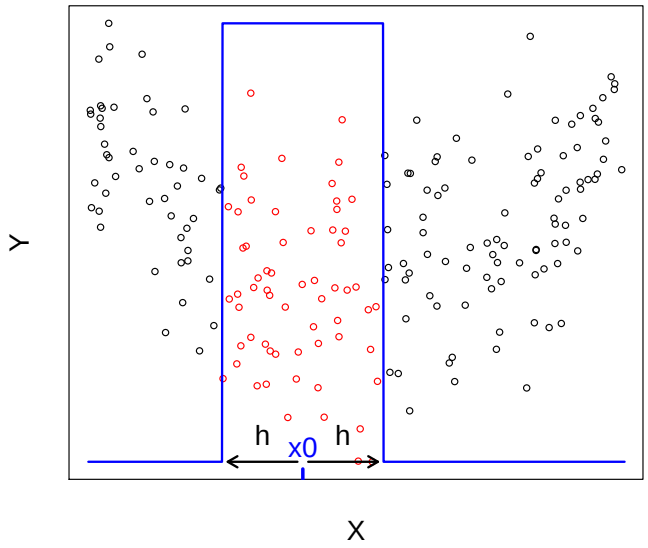
- Noyau rectangulaire

$$\begin{aligned}\hat{r}_h(x_0) &= \text{mean}\{Y_i, X_i \in [x_0-h, x_0+h]\} \\ &= \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in [x_0-h, x_0+h])}{\sum_{i=1}^n \mathbb{1}(X_i \in [x_0-h, x_0+h])}\end{aligned}$$

- Cas général : noyau K

$$\hat{r}_h(x_0) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x_0)}{\sum_{i=1}^n K_h(X_i - x_0)}$$

Méthode par noyaux de Nadaraya-Watson



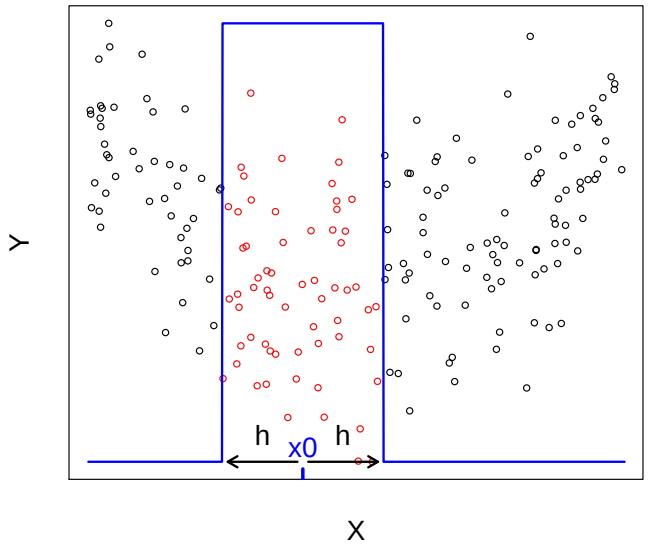
- Noyau rectangulaire

$$\begin{aligned}\hat{r}_h(x_0) &= \text{mean}\{Y_i, X_i \in [x_0-h, x_0+h]\} \\ &= \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in [x_0-h, x_0+h])}{\sum_{i=1}^n \mathbb{1}(X_i \in [x_0-h, x_0+h])}\end{aligned}$$

- Cas général : noyau K

$$\hat{r}_h(x_0) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x_0)}{\sum_{i=1}^n K_h(X_i - x_0)}$$

Méthode par noyaux de Nadaraya-Watson



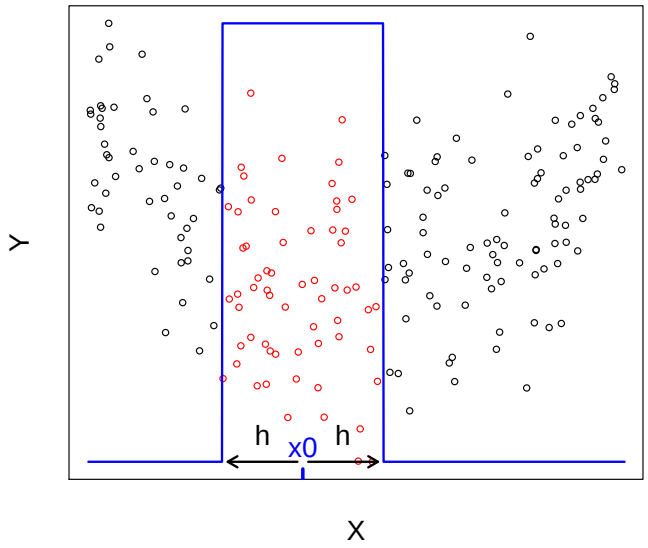
- Noyau rectangulaire

$$\begin{aligned}\hat{r}_h(x_0) &= \text{mean}\{\, Y_i, X_i \in [x_0 - h, x_0 + h]\} \\ &= \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in [x_0 - h, x_0 + h])}{\sum_{i=1}^n \mathbb{1}(X_i \in [x_0 - h, x_0 + h])}\end{aligned}$$

- Cas général : noyau K

$$\hat{r}_h(x_0) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x_0)}{\sum_{i=1}^n K_h(X_i - x_0)}$$

Méthode par noyaux de Nadaraya-Watson



- Noyau rectangulaire

$$\begin{aligned}\hat{r}_h(x_0) &= \text{mean}\{Y_i, X_i \in [x_0-h, x_0+h]\} \\ &= \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in [x_0-h, x_0+h])/2n*h}{\sum_{i=1}^n \mathbb{1}(X_i \in [x_0-h, x_0+h])/2n*h}\end{aligned}$$

- Cas général : noyau K

$$\hat{r}_h(x_0) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x_0)}{\sum_{i=1}^n K_h(X_i - x_0)}$$

Méthode par noyaux de Nadaraya-Watson

- Soit $f_{(X,Y)}$ la densité du couple (X, Y), alors la densité de Y sachant X vaut :

$$f_{Y|X}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

- D'où ,

$$r(x) = \mathbb{E}[Y|X = x] = \int y \frac{f_{(X,Y)}(x, y)}{f_X(x)} dy = \frac{1}{f_X(x)} \int y f_{(X,Y)}(x, y) dy$$

- En utilisant l'estimation de densité par noyaux

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)$$

$$\hat{f}_{(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x) K_h(Y_i - y)$$

Méthode par noyaux de Nadaraya-Watson

- Soit $f_{(X,Y)}$ la densité du couple (X, Y), alors la densité de Y sachant X vaut :

$$f_{Y|X}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

- D'où ,

$$r(x) = \mathbb{E}[Y|X = x] = \int y \frac{f_{(X,Y)}(x, y)}{f_X(x)} dy = \frac{1}{f_X(x)} \int y f_{(X,Y)}(x, y) dy$$

- En utilisant l'estimation de densité par noyaux

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)$$

$$\hat{f}_{(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x) K_h(Y_i - y)$$

Méthode par noyaux de Nadaraya-Watson

- Soit $f_{(X,Y)}$ la densité du couple (X, Y), alors la densité de Y sachant X vaut :

$$f_{Y|X}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

- D'où ,

$$r(x) = \mathbb{E}[Y|X = x] = \int y \frac{f_{(X,Y)}(x, y)}{f_X(x)} dy = \frac{1}{f_X(x)} \int y f_{(X,Y)}(x, y) dy$$

- En utilisant l'estimation de densité par noyaux

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)$$

$$\hat{f}_{(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x) K_h(Y_i - y)$$

Méthode par noyaux de Nadaraya-Watson

- Soit $f_{(X,Y)}$ la densité du couple (X, Y), alors la densité de Y sachant X vaut :

$$f_{Y|X}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

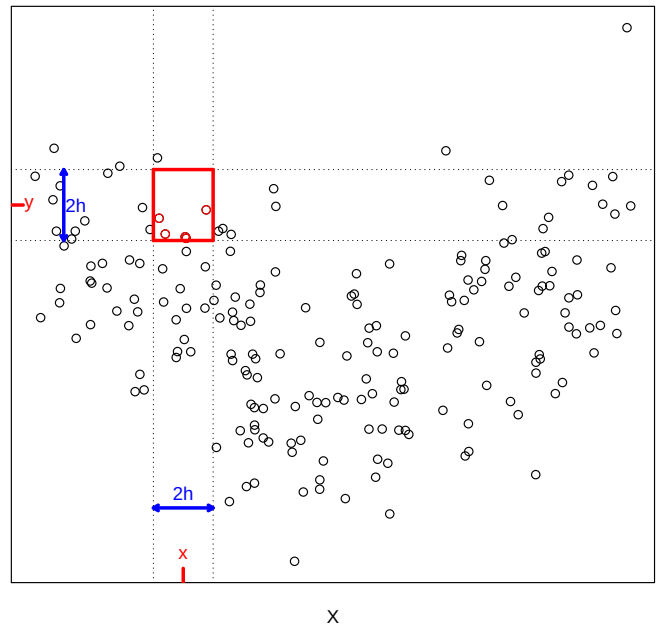
- D'où ,

$$r(x) = \mathbb{E}[Y|X = x] = \int y \frac{f_{(X,Y)}(x, y)}{f_X(x)} dy = \frac{1}{f_X(x)} \int y f_{(X,Y)}(x, y) dy$$

- En utilisant l'estimation de densité par noyaux

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)$$

$$\hat{f}_{(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x) K_h(Y_i - y)$$



Méthode par noyaux de Nadaraya-Watson

- Soit $f_{(X,Y)}$ la densité du couple (X, Y), alors la densité de Y sachant X vaut :

$$f_{Y|X}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

- D'où ,

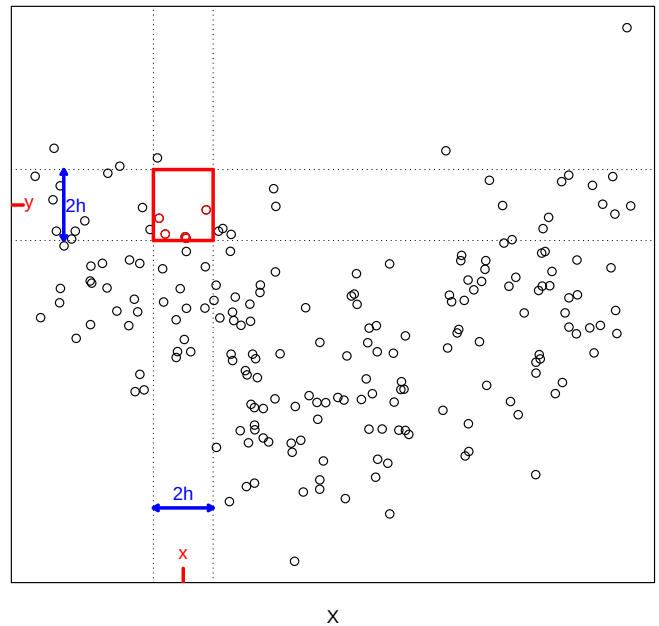
$$r(x) = \mathbb{E}[Y|X = x] = \int y \frac{f_{(X,Y)}(x, y)}{f_X(x)} dy = \frac{1}{f_X(x)} \int y f_{(X,Y)}(x, y) dy$$

- En utilisant l'estimation de densité par noyaux

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)$$

$$\hat{f}_{(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x) K_h(Y_i - y)$$

$\hookrightarrow$ exple : $\hat{f}_{(X,Y)}(x, y) = \frac{4}{2nh}$



- En remplaçant f_X et $f_{(X,Y)}$ par leurs estimateurs, on obtient

$$\begin{aligned}\hat{r}_h(x) &= \frac{1}{\hat{f}_X(x)} \int y \hat{f}_{(X,Y)}(x,y)\,dy \\ &= \frac{1}{\sum_{i=1}^n K_h(X_i - x)} \sum_{i=1}^n K_h(X_i - x) \int y K_h(Y_i - y)\,dy\end{aligned}$$

- De plus, par définition des noyaux $\int uK(u)\,du = 0$, d'où

$$\int y K_h(Y_i - y)\,dy = \int K(u)(Y_i + hu)\,du = Y_i$$

- Ainsi

$$\hat{r}_h(x) = \frac{\sum_{i=1}^n K_h(x - X_i)\, Y_i}{\sum_{i=1}^n K_h(x - X_i)}$$

- Fonction `npreg` dans le package `R np`.

- En remplaçant f_X et $f_{(X,Y)}$ par leurs estimateurs, on obtient

$$\begin{aligned}\hat{r}_h(x) &= \frac{1}{\hat{f}_X(x)} \int y \hat{f}_{(X,Y)}(x,y)\,dy \\ &= \frac{1}{\sum_{i=1}^n K_h(X_i - x)} \sum_{i=1}^n K_h(X_i - x) \int y K_h(Y_i - y)\,dy\end{aligned}$$

- De plus, par définition des noyaux $\int uK(u)\,du = 0$, d'où

$$\int y K_h(Y_i - y)\,dy = \int K(u)(Y_i + hu)\,du = Y_i$$

- Ainsi

$$\hat{r}_h(x) = \frac{\sum_{i=1}^n K_h(x - X_i)\, Y_i}{\sum_{i=1}^n K_h(x - X_i)}$$

- Fonction `npreg` dans le package `R np`.

- En remplaçant f_X et $f_{(X,Y)}$ par leurs estimateurs, on obtient

$$\begin{aligned}\hat{r}_h(x) &= \frac{1}{\hat{f}_X(x)} \int y \hat{f}_{(X,Y)}(x,y)\,dy \\ &= \frac{1}{\sum_{i=1}^n K_h(X_i - x)} \sum_{i=1}^n K_h(X_i - x) \int y K_h(Y_i - y)\,dy\end{aligned}$$

- De plus, par définition des noyaux $\int uK(u)\,du = 0$, d'où

$$\int y K_h(Y_i - y)\,dy = \int K(u)(Y_i + hu)\,du = Y_i$$

- Ainsi

$$\hat{r}_h(x) = \frac{\sum_{i=1}^n K_h(x - X_i)\, Y_i}{\sum_{i=1}^n K_h(x - X_i)}$$

- Fonction `npreg` dans le package `R np`.

- En remplaçant f_X et $f_{(X,Y)}$ par leurs estimateurs, on obtient

$$\begin{aligned}\hat{r}_h(x) &= \frac{1}{\hat{f}_X(x)} \int y \hat{f}_{(X,Y)}(x,y)\,dy \\ &= \frac{1}{\sum_{i=1}^n K_h(X_i - x)} \sum_{i=1}^n K_h(X_i - x) \int y K_h(Y_i - y)\,dy\end{aligned}$$

- De plus, par définition des noyaux $\int uK(u)\,du = 0$, d'où

$$\int y K_h(Y_i - y)\,dy = \int K(u)(Y_i + hu)\,du = Y_i$$

- Ainsi

$$\hat{r}_h(x) = \frac{\sum_{i=1}^n K_h(x - X_i)\, Y_i}{\sum_{i=1}^n K_h(x - X_i)}$$

- Fonction `npreg` dans le package `R np`.

Méthode des k plus proches voisins

- Principe Soit k un entier fixé. Pour tout $x \in I$, on regarde les k observations $(X_{i_1}, \ldots, X_{i_k})$ les plus proches de x, et

$$\hat{r}_k(x) = \frac{1}{k} \sum_{i=i_1,\ldots,i_k} Y_i$$

.

- Cette méthode est très générale et s'applique pour X_i à valeurs dans un espace quelconque (exple $\mathbb{R}^N$) sur lequel on a défini une distance.
- Cette méthode est similaire au noyau rectangulaire mais avec une fenêtre variable (dépendant de la densité des observations autour de x)
- Comme pour les noyaux, le biais augmente et la variance diminue avec k.
- Différentes règles empiriques existent pour le choix de k.

Méthode des k plus proches voisins

- Principe Soit k un entier fixé. Pour tout $x \in I$, on regarde les k observations $(X_{i_1}, \ldots, X_{i_k})$ les plus proches de x, et

$$\hat{r}_k(x) = \frac{1}{k} \sum_{i=i_1,\ldots,i_k} Y_i$$

.

- Cette méthode est très générale et s'applique pour X_i à valeurs dans un espace quelconque (exple $\mathbb{R}^N$) sur lequel on a défini une distance.
- Cette méthode est similaire au noyau rectangulaire mais avec une fenêtre variable (dépendant de la densité des observations autour de x)
- Comme pour les noyaux, le biais augmente et la variance diminue avec k.
- Différentes règles empiriques existent pour le choix de k.

Méthode des k plus proches voisins

- Principe Soit k un entier fixé. Pour tout $x \in I$, on regarde les k observations $(X_{i_1}, \ldots, X_{i_k})$ les plus proches de x, et

$$\hat{r}_k(x) = \frac{1}{k} \sum_{i=i_1,\ldots,i_k} Y_i$$

.

- Cette méthode est très générale et s'applique pour X_i à valeurs dans un espace quelconque (exple $\mathbb{R}^N$) sur lequel on a défini une distance.
- Cette méthode est similaire au noyau rectangulaire mais avec une fenêtre variable (dépendant de la densité des observations autour de x)
- Comme pour les noyaux, le biais augmente et la variance diminue avec k.
- Différentes règles empiriques existent pour le choix de k.

Méthode des k plus proches voisins

- Principe Soit k un entier fixé. Pour tout $x \in I$, on regarde les k observations $(X_{i_1}, \ldots, X_{i_k})$ les plus proches de x, et

$$\hat{r}_k(x) = \frac{1}{k} \sum_{i=i_1,\ldots,i_k} Y_i$$

.

- Cette méthode est très générale et s'applique pour X_i à valeurs dans un espace quelconque (exple $\mathbb{R}^N$) sur lequel on a défini une distance.
- Cette méthode est similaire au noyau rectangulaire mais avec une fenêtre variable (dépendant de la densité des observations autour de x)
- Comme pour les noyaux, le biais augmente et la variance diminue avec k.
- Différentes règles empiriques existent pour le choix de k.

Méthode des k plus proches voisins

- Principe Soit k un entier fixé. Pour tout $x \in I$, on regarde les k observations $(X_{i_1}, \ldots, X_{i_k})$ les plus proches de x, et

$$\hat{r}_k(x) = \frac{1}{k} \sum_{i=i_1,\ldots,i_k} Y_i$$

.

- Cette méthode est très générale et s'applique pour X_i à valeurs dans un espace quelconque (exple $\mathbb{R}^N$) sur lequel on a défini une distance.
- Cette méthode est similaire au noyau rectangulaire mais avec une fenêtre variable (dépendant de la densité des observations autour de x)
- Comme pour les noyaux, le biais augmente et la variance diminue avec k.
- Différentes règles empiriques existent pour le choix de k.

Régression par polynômes locaux

Principe général : Soit $x_0 \in I$, $r(x_0)$ est estimée par régréssion des MC sur une base de polynômes avec des poids attribués aux observations X_i selon leur distance à x_0.

- Pour un degré r donné on considère l'espace des polynomes de degré $\leq r$:

$$S_r = \text{vect}\left\{\psi_j : x \to (x - x_0)^j, j = 0, \ldots, r\right\}$$

- On se donne une fonction de poids $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ décroissante.
- $\hat{r}(x_0) = \sum_{j=0}^{r} \hat{\theta}_j \psi_j(x_0)$ avec

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \omega(|X_i - x_0|) \left(Y_i - \sum_{j=0}^{r} \hat{\theta}_j \psi_j(X_i) \right)^2$$

$\hookrightarrow$ $\hat{\theta}$ dépend du point x_0 considéré.

Régression par polynômes locaux

Principe général : Soit $x_0 \in I$, $r(x_0)$ est estimée par régréssion des MC sur une base de polynômes avec des poids attribués aux observations X_i selon leur distance à x_0.

- Pour un degré r donné on considère l'espace des polynomes de degré $\leq r$:

$$S_r = \text{vect}\left\{\psi_j : x \to (x - x_0)^j, j = 0, \ldots, r\right\}$$

- On se donne une fonction de poids $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ décroissante.
- $\hat{r}(x_0) = \sum_{j=0}^{r} \hat{\theta}_j \psi_j(x_0)$ avec

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \omega(|X_i - x_0|) \left(Y_i - \sum_{j=0}^{r} \hat{\theta}_j \psi_j(X_i) \right)^2$$

$\hookrightarrow$ $\hat{\theta}$ dépend du point x_0 considéré.

Régression par polynômes locaux

Principe général : Soit $x_0 \in I$, $r(x_0)$ est estimée par régréssion des MC sur une base de polynômes avec des poids attribués aux observations X_i selon leur distance à x_0.

- Pour un degré r donné on considère l'espace des polynomes de degré $\leq r$:

$$S_r = \text{vect}\left\{\psi_j : x \to (x - x_0)^j, j = 0, \ldots, r\right\}$$

- On se donne une fonction de poids $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ décroissante.
- $\hat{r}(x_0) = \sum_{j=0}^{r} \hat{\theta}_j \psi_j(x_0)$ avec

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \omega(|X_i - x_0|) \left(Y_i - \sum_{j=0}^{r} \hat{\theta}_j \psi_j(X_i) \right)^2$$

$\hookrightarrow$ $\hat{\theta}$ dépend du point x_0 considéré.

Régression par polynômes locaux

Principe général : Soit $x_0 \in I$, $r(x_0)$ est estimée par régréssion des MC sur une base de polynômes avec des poids attribués aux observations X_i selon leur distance à x_0.

- Pour un degré r donné on considère l'espace des polynomes de degré $\leq r$:

$$S_r = \text{vect}\left\{\psi_j : x \to (x - x_0)^j, j = 0, \ldots, r\right\}$$

- On se donne une fonction de poids $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ décroissante.
- $\hat{r}(x_0) = \sum_{j=0}^{r} \hat{\theta}_j \psi_j(x_0)$ avec

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \omega(|X_i - x_0|) \left(Y_i - \sum_{j=0}^{r} \hat{\theta}_j \psi_j(X_i) \right)^2$$

$\hookrightarrow$ $\hat{\theta}$ dépend du point x_0 considéré.

Régression par polynômes locaux

Principe général : Soit $x_0 \in I$, $r(x_0)$ est estimée par régréssion des MC sur une base de polynômes avec des poids attribués aux observations X_i selon leur distance à x_0.

- Pour un degré r donné on considère l'espace des polynomes de degré $\leq r$:

$$S_r = \text{vect}\left\{\psi_j : x \to (x - x_0)^j, j = 0, \ldots, r\right\}$$

- On se donne une fonction de poids $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ décroissante.
- $\hat{r}(x_0) = \sum_{j=0}^{r} \hat{\theta}_j \psi_j(x_0)$ avec

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \omega(|X_i - x_0|) \left(Y_i - \sum_{j=0}^{r} \hat{\theta}_j \psi_j(X_i) \right)^2$$

$\hookrightarrow$ $\hat{\theta}$ dépend du point x_0 considéré.

Régression par polynômes locaux (2)

Commentaires

- Exemples de fonction de poids ω :
  - Poids défini par un noyau : $\omega(|u|) = K_h(u)$
  - Intervalle fixe : $\omega(|u|) = \mathbb{1}_{[-1,1]}$ (en fait, c'est un noyau !)
- Plus la fonction de poids ω est à décroissance rapide, plus le biais est faible et plus la variance est grande.
- La fonction `lowess` dans `R` mixe les méthodes des k plus proches voisins et des polynômes locaux de degré 1.

Régression par polynômes locaux (2)

Commentaires

- Exemples de fonction de poids ω :
  - Poids défini par un noyau : $\omega(|u|) = K_h(u)$
  - Intervalle fixe : $\omega(|u|) = \mathbb{1}_{[-1,1]}$ (en fait, c'est un noyau !)
- Plus la fonction de poids ω est à décroissance rapide, plus le biais est faible et plus la variance est grande.
- La fonction `lowess` dans `R` mixe les méthodes des k plus proches voisins et des polynômes locaux de degré 1.

Régression par polynômes locaux (2)

Commentaires

- Exemples de fonction de poids ω :
  - Poids défini par un noyau : $\omega(|u|) = K_h(u)$
  - Intervalle fixe : $\omega(|u|) = \mathbb{1}_{[-1,1]}$ (en fait, c'est un noyau !)
- Plus la fonction de poids ω est à décroissance rapide, plus le biais est faible et plus la variance est grande.
- La fonction `lowess` dans `R` mixe les méthodes des k plus proches voisins et des polynômes locaux de degré 1.

Régression par polynômes locaux (2)

Commentaires

- Exemples de fonction de poids ω :
  - Poids défini par un noyau : $\omega(|u|) = K_h(u)$
  - Intervalle fixe : $\omega(|u|) = \mathbb{1}_{[-1,1]}$ (en fait, c'est un noyau !)
- Plus la fonction de poids ω est à décroissance rapide, plus le biais est faible et plus la variance est grande.
- La fonction `lowess` dans `R` mixe les méthodes des k plus proches voisins et des polynômes locaux de degré 1.

Conclusion sur les méthodes de régression présentées

- **Méthodes locales et globales**
  - La méthodes des MC est globale : l'estimateur est calculé simultanément en tout point de I.
  - Les autres méthodes présentées sont locales : l'estimateur est théoriquement calculé en chaque point $x \in I$ (en pratique sur une grille de points)
- **Compromis biais-variance** : toutes les méthodes comportent des paramètres de régularisation qui déterminent la répartition de l'erreur en biais et variance.
  - MC : dimension D
  - Méthodes par noyau : fenêtre h
  - Méthode des k plus proches voisins : nombre de voisins k
- **De nombreuses variantes** mixant ces différentes approches sont disponibles.
- **Le critère des MC** est très général et peut-être utilisé dans des contextes variés (paramétriques ou non).

Conclusion sur les méthodes de régression présentées

- **Méthodes locales et globales**
  - La méthodes des MC est globale : l'estimateur est calculé simultanément en tout point de I.
  - Les autres méthodes présentées sont locales : l'estimateur est théoriquement calculé en chaque point $x \in I$ (en pratique sur une grille de points)
- **Compromis biais-variance** : toutes les méthodes comportent des paramètres de régularisation qui déterminent la répartition de l'erreur en biais et variance.
  - MC : dimension D
  - Méthodes par noyau : fenêtre h
  - Méthode des k plus proches voisins : nombre de voisins k
- **De nombreuses variantes** mixant ces différentes approches sont disponibles.
- **Le critère des MC** est très général et peut-être utilisé dans des contextes variés (paramétriques ou non).

Conclusion sur les méthodes de régression présentées

- **Méthodes locales et globales**
  - La méthodes des MC est globale : l'estimateur est calculé simultanément en tout point de I.
  - Les autres méthodes présentées sont locales : l'estimateur est théoriquement calculé en chaque point $x \in I$ (en pratique sur une grille de points)
- **Compromis biais-variance** : toutes les méthodes comportent des paramètres de régularisation qui déterminent la répartition de l'erreur en biais et variance.
  - MC : dimension D
  - Méthodes par noyau : fenêtre h
  - Méthode des k plus proches voisins : nombre de voisins k
- **De nombreuses variantes** mixant ces différentes approches sont disponibles.
- **Le critère des MC** est très général et peut-être utilisé dans des contextes variés (paramétriques ou non).

Conclusion sur les méthodes de régression présentées

- **Méthodes locales et globales**
  - La méthodes des MC est globale : l'estimateur est calculé simultanément en tout point de I.
  - Les autres méthodes présentées sont locales : l'estimateur est théoriquement calculé en chaque point $x \in I$ (en pratique sur une grille de points)
- **Compromis biais-variance** : toutes les méthodes comportent des paramètres de régularisation qui déterminent la répartition de l'erreur en biais et variance.
  - MC : dimension D
  - Méthodes par noyau : fenêtre h
  - Méthode des k plus proches voisins : nombre de voisins k
- **De nombreuses variantes** mixant ces différentes approches sont disponibles.
- **Le critère des MC** est très général et peut-être utilisé dans des contextes variés (paramétriques ou non).

La validation-croisée : introduction

- Dans ce cours, nous avons souvent comparé visuellement l'estimateur et sa vraie fonction pour conclure aux performances de l'estimateur.

  $\hookrightarrow$ Point de vue théorique (vraie fonction inconnue en pratique !)

- Nous avons vu des méthodes pour choisir le paramètre de régularisation.

  $\hookrightarrow$ Méthodes spécifiques à un contexte.

- Validation croisée : approche beaucoup plus générale, basée sur les données, qui répond à la question : dans quelle mesure $\hat{r}$ permet de prédire Y pour de futures observations où on ne mesure que X ?

  $\hookrightarrow$ Erreur de prédiction

- Les résultats de cette section sont valables pour un contexte de régression général : $r(x) = \mathbb{E}[Y|X = x]$

La validation-croisée : introduction

- Dans ce cours, nous avons souvent comparé visuellement l'estimateur et sa vraie fonction pour conclure aux performances de l'estimateur.

  $\hookrightarrow$ Point de vue théorique (vraie fonction inconnue en pratique !)

- Nous avons vu des méthodes pour choisir le paramètre de régularisation.

  $\hookrightarrow$ Méthodes spécifiques à un contexte.

- Validation croisée : approche beaucoup plus générale, basée sur les données, qui répond à la question : dans quelle mesure $\hat{r}$ permet de prédire Y pour de futures observations où on ne mesure que X ?

  $\hookrightarrow$ Erreur de prédiction

- Les résultats de cette section sont valables pour un contexte de régression général : $r(x) = \mathbb{E}[Y|X = x]$

La validation-croisée : introduction

- Dans ce cours, nous avons souvent comparé visuellement l'estimateur et sa vraie fonction pour conclure aux performances de l'estimateur.

  $\hookrightarrow$ Point de vue théorique (vraie fonction inconnue en pratique !)
- Nous avons vu des méthodes pour choisir le paramètre de régularisation.

  $\hookrightarrow$ Méthodes spécifiques à un contexte.
- Validation croisée : approche beaucoup plus générale, basée sur les données, qui répond à la question : dans quelle mesure $\hat{r}$ permet de prédire Y pour de futures observations où on ne mesure que X ?

  $\hookrightarrow$ Erreur de prédiction
- Les résultats de cette section sont valables pour un contexte de régression général : $r(x) = \mathbb{E}[Y|X = x]$

La validation-croisée : introduction

- Dans ce cours, nous avons souvent comparé visuellement l'estimateur et sa vraie fonction pour conclure aux performances de l'estimateur.

  $\hookrightarrow$ Point de vue théorique (vraie fonction inconnue en pratique !)
- Nous avons vu des méthodes pour choisir le paramètre de régularisation.

  $\hookrightarrow$ Méthodes spécifiques à un contexte.
- Validation croisée : approche beaucoup plus générale, basée sur les données, qui répond à la question : dans quelle mesure $\hat{r}$ permet de prédire Y pour de futures observations où on ne mesure que X ?

  $\hookrightarrow$ Erreur de prédiction
- Les résultats de cette section sont valables pour un contexte de régression général : $r(x) = \mathbb{E}[Y|X = x]$

La validation-croisée : introduction

- Dans ce cours, nous avons souvent comparé visuellement l'estimateur et sa vraie fonction pour conclure aux performances de l'estimateur.

  $\hookrightarrow$ Point de vue théorique (vraie fonction inconnue en pratique !)
- Nous avons vu des méthodes pour choisir le paramètre de régularisation.

  $\hookrightarrow$ Méthodes spécifiques à un contexte.
- Validation croisée : approche beaucoup plus générale, basée sur les données, qui répond à la question : dans quelle mesure $\hat{r}$ permet de prédire Y pour de futures observations où on ne mesure que X ?

  $\hookrightarrow$ Erreur de prédiction
- Les résultats de cette section sont valables pour un contexte de régression général : $r(x) = \mathbb{E}[Y|X = x]$

La validation-croisée : introduction

- Dans ce cours, nous avons souvent comparé visuellement l'estimateur et sa vraie fonction pour conclure aux performances de l'estimateur.

  $\hookrightarrow$ Point de vue théorique (vraie fonction inconnue en pratique !)
- Nous avons vu des méthodes pour choisir le paramètre de régularisation.

  $\hookrightarrow$ Méthodes spécifiques à un contexte.
- Validation croisée : approche beaucoup plus générale, basée sur les données, qui répond à la question : dans quelle mesure $\hat{r}$ permet de prédire Y pour de futures observations où on ne mesure que X ?

  $\hookrightarrow$ Erreur de prédiction
- Les résultats de cette section sont valables pour un contexte de régression général : $r(x) = \mathbb{E}[Y|X = x]$

La validation-croisée : introduction

- Dans ce cours, nous avons souvent comparé visuellement l'estimateur et sa vraie fonction pour conclure aux performances de l'estimateur.

  $\hookrightarrow$ Point de vue théorique (vraie fonction inconnue en pratique !)
- Nous avons vu des méthodes pour choisir le paramètre de régularisation.

  $\hookrightarrow$ Méthodes spécifiques à un contexte.
- Validation croisée : approche beaucoup plus générale, basée sur les données, qui répond à la question : dans quelle mesure $\hat{r}$ permet de prédire Y pour de futures observations où on ne mesure que X ?

  $\hookrightarrow$ Erreur de prédiction
- Les résultats de cette section sont valables pour un contexte de régression général : $r(x) = \mathbb{E}[Y|X = x]$

Erreur de prédiction

- Soit (X, Y) un couple de variables, $r(x) = \mathbb{E}[Y|X = x]$.
- Soit $\hat{r}$ une procédure d'estimation (ex : MC par hist réguliers à 10 morceaux).
- Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ échantillon i.i.d. de même loi que (X, Y).
- Soit $\hat{r}_n$ l'estimateur calculé selon la procédure $\hat{r}$ à partir de l'échantillon $\mathcal{Z}$.
- Soit $\mathcal{Z}' = (X_{n+1}, Y_{n+1})$ une nouvelle réalisation de (X, Y) où on ne mesure que X.

- Y_{n+1} est prédit par

$$\hat{Y}_{n+1} = \hat{r}_n(X_{n+1})$$

- L'erreur de prédiction pour Y_{n+1} est

$$\text{Err}(\hat{r}_n, \mathcal{Z}') = (Y_{n+1} - \hat{r}_n(X_{n+1}))^2$$

Erreur de prédiction

- Soit (X, Y) un couple de variables, $r(x) = \mathbb{E}[Y|X = x]$.
- Soit $\hat{r}$ une procédure d'estimation (ex : MC par hist réguliers à 10 morceaux).
- Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ échantillon i.i.d. de même loi que (X, Y).
- Soit $\hat{r}_n$ l'estimateur calculé selon la procédure $\hat{r}$ à partir de l'échantillon $\mathcal{Z}$.
- Soit $\mathcal{Z}' = (X_{n+1}, Y_{n+1})$ une nouvelle réalisation de (X, Y) où on ne mesure que X.

- Y_{n+1} est prédit par

$$\hat{Y}_{n+1} = \hat{r}_n(X_{n+1})$$

- L'erreur de prédiction pour Y_{n+1} est

$$\text{Err}(\hat{r}_n, \mathcal{Z}') = (Y_{n+1} - \hat{r}_n(X_{n+1}))^2$$

Erreur de prédiction

- Soit (X, Y) un couple de variables, $r(x) = \mathbb{E}[Y|X = x]$.
- Soit $\hat{r}$ une procédure d'estimation (ex : MC par hist réguliers à 10 morceaux).
- Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ échantillon i.i.d. de même loi que (X, Y).
- Soit $\hat{r}_n$ l'estimateur calculé selon la procédure $\hat{r}$ à partir de l'échantillon $\mathcal{Z}$.
- Soit $\mathcal{Z}' = (X_{n+1}, Y_{n+1})$ une nouvelle réalisation de (X, Y) où on ne mesure que X.

- Y_{n+1} est prédit par

$$\hat{Y}_{n+1} = \hat{r}_n(X_{n+1})$$

- L'erreur de prédiction pour Y_{n+1} est

$$\text{Err}(\hat{r}_n, \mathcal{Z}') = (Y_{n+1} - \hat{r}_n(X_{n+1}))^2$$

Erreur de prédiction

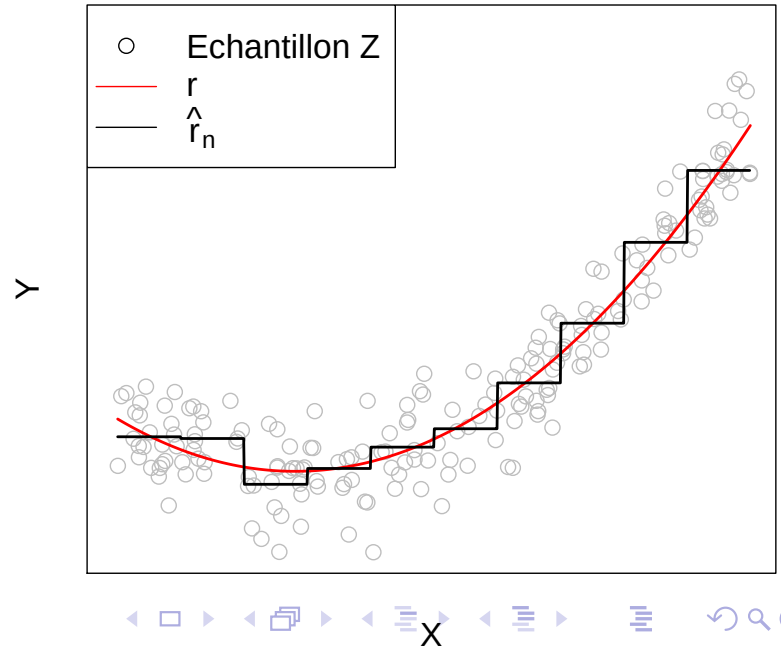
- Soit (X, Y) un couple de variables, $r(x) = \mathbb{E}[Y|X = x]$.
- Soit $\hat{r}$ une procédure d'estimation (ex : MC par hist réguliers à 10 morceaux).
- Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\ldots,n}$ échantillon i.i.d. de même loi que (X, Y).
- Soit $\hat{r}_n$ l'estimateur calculé selon la procédure $\hat{r}$ à partir de l'échantillon $\mathcal{Z}$.
- Soit $\mathcal{Z}' = (X_{n+1}, Y_{n+1})$ une nouvelle réalisation de (X, Y) où on ne mesure que X.

- Y_{n+1} est prédit par

$$\hat{Y}_{n+1} = \hat{r}_n(X_{n+1})$$

- L'erreur de prédiction pour Y_{n+1} est

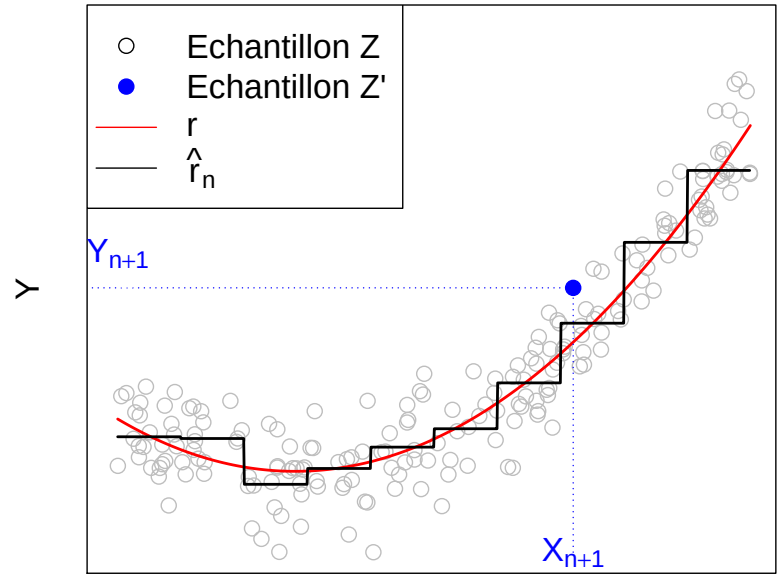
$$\text{Err}(\hat{r}_n, \mathcal{Z}') = (Y_{n+1} - \hat{r}_n(X_{n+1}))^2$$



Erreur de prédiction

- Soit (X, Y) un couple de variables, $r(x) = \mathbb{E}[Y|X = x]$.
- Soit $\hat{r}$ une procédure d'estimation (ex : MC par hist réguliers à 10 morceaux).
- Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\ldots,n}$ échantillon i.i.d. de même loi que (X, Y).
- Soit $\hat{r}_n$ l'estimateur calculé selon la procédure $\hat{r}$ à partir de l'échantillon $\mathcal{Z}$.
- Soit $\mathcal{Z}' = (X_{n+1}, Y_{n+1})$ une nouvelle réalisation de (X, Y) où on ne mesure que X.

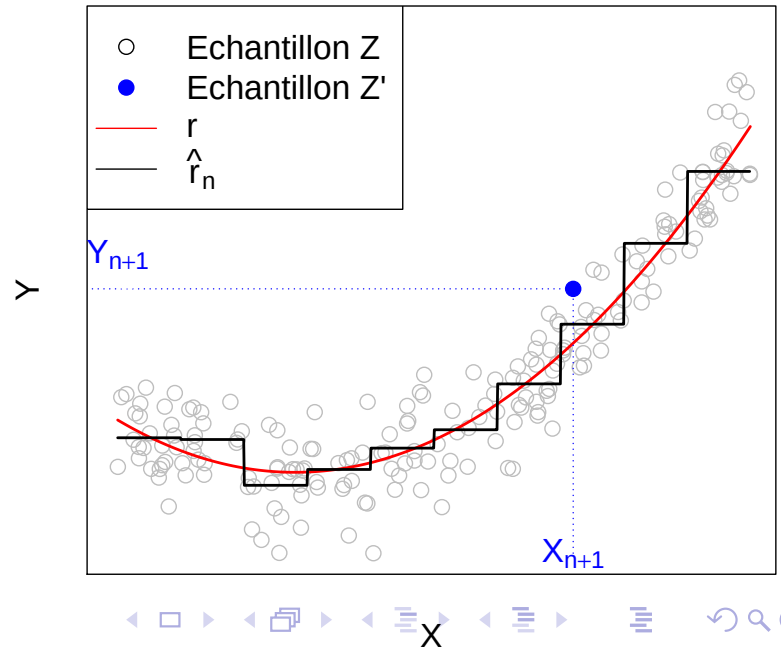
- Y_{n+1} est prédit par
$$\hat{Y}_{n+1} = \hat{r}_n(X_{n+1})$$
- L'erreur de prédiction pour Y_{n+1} est
$$\mathrm{Err}(\hat{r}_n, \mathcal{Z}') = (Y_{n+1} - \hat{r}_n(X_{n+1}))^2$$



Erreur de prédiction

- Soit (X, Y) un couple de variables, $r(x) = \mathbb{E}[Y|X = x]$.
- Soit $\hat{r}$ une procédure d'estimation (ex : MC par hist réguliers à 10 morceaux).
- Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\ldots,n}$ échantillon i.i.d. de même loi que (X, Y).
- Soit $\hat{r}_n$ l'estimateur calculé selon la procédure $\hat{r}$ à partir de l'échantillon $\mathcal{Z}$.
- Soit $\mathcal{Z}' = (X_{n+1}, Y_{n+1})$ une nouvelle réalisation de (X, Y) où on ne mesure que X.

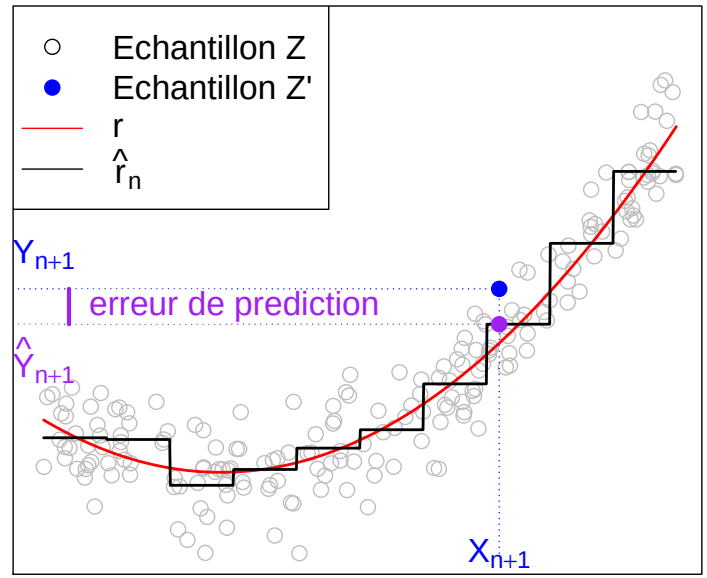
- Y_{n+1} est prédit par
$$\hat{Y}_{n+1} = \hat{r}_n(X_{n+1})$$
- L'erreur de prédiction pour Y_{n+1} est
$$\mathrm{Err}(\hat{r}_n, \mathcal{Z}') = (Y_{n+1} - \hat{r}_n(X_{n+1}))^2$$



Erreur de prédiction

- Soit (X, Y) un couple de variables, $r(x) = \mathbb{E}[Y|X = x]$.
- Soit $\hat{r}$ une procédure d'estimation (ex : MC par hist réguliers à 10 morceaux).
- Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\ldots,n}$ échantillon i.i.d. de même loi que (X, Y).
- Soit $\hat{r}_n$ l'estimateur calculé selon la procédure $\hat{r}$ à partir de l'échantillon $\mathcal{Z}$.
- Soit $\mathcal{Z}' = (X_{n+1}, Y_{n+1})$ une nouvelle réalisation de (X, Y) où on ne mesure que X.

- Y_{n+1} est prédit par
$$\hat{Y}_{n+1} = \hat{r}_n(X_{n+1})$$
- L'erreur de prédiction pour Y_{n+1} est
$$\text{Err}(\hat{r}_n, \mathcal{Z}') = (Y_{n+1} - \hat{r}_n(X_{n+1}))^2$$



Erreur de prédiction moyenne

- On définit l'erreur de prédiction moyenne

$$E_n(\hat{r}) = \mathbb{E}\left[\mathsf{Err}(\hat{r}_n, \mathcal{Z}')\right]$$

  - Ne dépend plus des échantillons particuliers $\mathcal{Z}$ et $\mathcal{Z}'$.
  - Ne dépend que de n, de la procédure $\hat{r}$ et de la loi de (X, Y).

- Cas de la régression additive : $Y_i = r(X_i) + \varepsilon_i$ avec $\epsilon_i \perp X_i$.

$$\begin{aligned}
&\mathbb{E}[\mathsf{Err}(\hat{r}, \mathcal{Z}')|\mathcal{Z}] = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] \\
&= \mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] + \mathbb{E}[\varepsilon_{n+1}^2|\mathcal{Z}] + 2\mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))\varepsilon_{n+1}|\mathcal{Z}] \\
&= \|r - \hat{r}_n\|_{f_X}^2 + \sigma^2 + 0
\end{aligned}$$

  - Notons $\mathsf{MISE}_n = \mathbb{E}[\|r - \hat{r}_n\|_{f_X}^2]$, alors

$$E_n(\hat{r}) = MISE_n + cte$$

- Comment estimer $E_n(\hat{r})$?

Erreur de prédiction moyenne

- On définit l'erreur de prédiction moyenne

$$E_n(\hat{r}) = \mathbb{E}\left[\mathsf{Err}(\hat{r}_n, \mathcal{Z}')\right]$$

  - Ne dépend plus des échantillons particuliers $\mathcal{Z}$ et $\mathcal{Z}'$.
  - Ne dépend que de n, de la procédure $\hat{r}$ et de la loi de (X, Y).

- Cas de la régression additive : $Y_i = r(X_i) + \varepsilon_i$ avec $\epsilon_i \perp X_i$.

$$\begin{aligned}
\mathbb{E}[\mathsf{Err}(\hat{r}, \mathcal{Z}')|\mathcal{Z}] &= \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] \\
&= \mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] + \mathbb{E}[\varepsilon_{n+1}^2|\mathcal{Z}] + 2\mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))\varepsilon_{n+1}|\mathcal{Z}] \\
&= \|r - \hat{r}_n\|_{f_X}^2 + \sigma^2 + 0
\end{aligned}$$

  - Notons $\mathsf{MISE}_n = \mathbb{E}[\|r - \hat{r}_n\|_{f_X}^2]$, alors

$$E_n(\hat{r}) = MISE_n + cte$$

- Comment estimer $E_n(\hat{r})$?

Erreur de prédiction moyenne

- On définit l'erreur de prédiction moyenne

$$E_n(\hat{r}) = \mathbb{E}\left[\mathsf{Err}(\hat{r}_n, \mathcal{Z}')\right]$$

  - Ne dépend plus des échantillons particuliers $\mathcal{Z}$ et $\mathcal{Z}'$.
  - Ne dépend que de n, de la procédure $\hat{r}$ et de la loi de (X, Y).

- Cas de la régression additive : $Y_i = r(X_i) + \varepsilon_i$ avec $\epsilon_i \perp X_i$.

$$\begin{aligned}
\mathbb{E}[\mathsf{Err}(\hat{r}, \mathcal{Z}')|\mathcal{Z}] &= \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] \\
&= \mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] + \mathbb{E}[\varepsilon_{n+1}^2|\mathcal{Z}] + 2\mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))\varepsilon_{n+1}|\mathcal{Z}] \\
&= \|r - \hat{r}_n\|_{f_X}^2 + \sigma^2 + 0
\end{aligned}$$

  - Notons $\mathrm{MISE}_n = \mathbb{E}[\|r - \hat{r}_n\|_{f_X}^2]$, alors

$$E_n(\hat{r}) = MISE_n + cte$$

- Comment estimer $E_n(\hat{r})$?

Erreur de prédiction moyenne

- On définit l'erreur de prédiction moyenne

$$E_n(\hat{r}) = \mathbb{E}\left[\mathsf{Err}(\hat{r}_n, \mathcal{Z}')\right]$$

  - Ne dépend plus des échantillons particuliers $\mathcal{Z}$ et $\mathcal{Z}'$.
  - Ne dépend que de n, de la procédure $\hat{r}$ et de la loi de (X, Y).
- Cas de la régression additive : $Y_i = r(X_i) + \varepsilon_i$ avec $\epsilon_i \perp X_i$.

$$\begin{aligned}
&\mathbb{E}[\mathsf{Err}(\hat{r}, \mathcal{Z}')|\mathcal{Z}] = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] \\
&= \mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] + \mathbb{E}[\varepsilon_{n+1}^2|\mathcal{Z}] + 2\mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))\varepsilon_{n+1}|\mathcal{Z}] \\
&= \|r - \hat{r}_n\|_{f_X}^2 + \sigma^2 + 0
\end{aligned}$$

  - Notons $\mathsf{MISE}_n = \mathbb{E}[\|r - \hat{r}_n\|_{f_X}^2]$, alors

$$E_n(\hat{r}) = MISE_n + cte$$

- Comment estimer $E_n(\hat{r})$?

Erreur de prédiction moyenne

- On définit l'erreur de prédiction moyenne

$$E_n(\hat{r}) = \mathbb{E}\left[\mathsf{Err}(\hat{r}_n, \mathcal{Z}')\right]$$

  - Ne dépend plus des échantillons particuliers $\mathcal{Z}$ et $\mathcal{Z}'$.
  - Ne dépend que de n, de la procédure $\hat{r}$ et de la loi de (X, Y).
- Cas de la régression additive : $Y_i = r(X_i) + \varepsilon_i$ avec $\epsilon_i \perp X_i$.

$$\begin{aligned}
&\mathbb{E}[\mathsf{Err}(\hat{r}, \mathcal{Z}')|\mathcal{Z}] = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] \\
&= \mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] + \mathbb{E}[\varepsilon_{n+1}^2|\mathcal{Z}] + 2\mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))\varepsilon_{n+1}|\mathcal{Z}] \\
&= \|r - \hat{r}_n\|_{f_X}^2 + \sigma^2 + 0
\end{aligned}$$

  - Notons $\mathrm{MISE}_n = \mathbb{E}[\|r - \hat{r}_n\|_{f_X}^2]$, alors

$$E_n(\hat{r}) = MISE_n + cte$$

- Comment estimer $E_n(\hat{r})$?

Erreur de prédiction moyenne

- On définit l'erreur de prédiction moyenne

$$E_n(\hat{r}) = \mathbb{E}\left[\mathsf{Err}(\hat{r}_n, \mathcal{Z}')\right]$$

  - Ne dépend plus des échantillons particuliers $\mathcal{Z}$ et $\mathcal{Z}'$.
  - Ne dépend que de n, de la procédure $\hat{r}$ et de la loi de (X, Y).

- Cas de la régression additive : $Y_i = r(X_i) + \varepsilon_i$ avec $\epsilon_i \perp X_i$.

$$\begin{aligned}
&\mathbb{E}[\mathsf{Err}(\hat{r}, \mathcal{Z}')|\mathcal{Z}] = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] \\
&= \mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] + \mathbb{E}[\varepsilon_{n+1}^2|\mathcal{Z}] + 2\mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))\varepsilon_{n+1}|\mathcal{Z}] \\
&= \|r - \hat{r}_n\|_{f_X}^2 + \sigma^2 + 0
\end{aligned}$$

  - Notons $\mathsf{MISE}_n = \mathbb{E}[\|r - \hat{r}_n\|_{f_X}^2]$, alors

$$E_n(\hat{r}) = MISE_n + cte$$

- Comment estimer $E_n(\hat{r})$?

Erreur de prédiction moyenne

- On définit l'erreur de prédiction moyenne

$$E_n(\hat{r}) = \mathbb{E}\left[\mathsf{Err}(\hat{r}_n, \mathcal{Z}')\right]$$

  - Ne dépend plus des échantillons particuliers $\mathcal{Z}$ et $\mathcal{Z}'$.
  - Ne dépend que de n, de la procédure $\hat{r}$ et de la loi de (X, Y).

- Cas de la régression additive : $Y_i = r(X_i) + \varepsilon_i$ avec $\epsilon_i \perp X_i$.

$$\begin{aligned}
&\mathbb{E}[\mathsf{Err}(\hat{r}, \mathcal{Z}')|\mathcal{Z}] = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] \\
&= \mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))^2|\mathcal{Z}] + \mathbb{E}[\varepsilon_{n+1}^2|\mathcal{Z}] + 2\mathbb{E}[(r(X_{n+1}) - \hat{r}_n(X_{n+1}))\varepsilon_{n+1}|\mathcal{Z}] \\
&= \|r - \hat{r}_n\|_{f_X}^2 + \sigma^2 + 0
\end{aligned}$$

  - Notons $\mathsf{MISE}_n = \mathbb{E}[\|r - \hat{r}_n\|_{f_X}^2]$, alors

$$E_n(\hat{r}) = MISE_n + cte$$

- Comment estimer $E_n(\hat{r})$?

Moyenne des carrés des résidus et erreur de prédiction

- $E_n(\hat{r}) = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2]$ avec $\hat{r}_n$ calculé à partir de $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\dots,n}$.
- Idée naïve : estimer $E_n(\hat{r})$ par la moyenne des carrés des résidus :

$$\text{MCR}(\hat{r}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(X_i) - Y_i)^2$$

$\hookrightarrow$ C'est un estimateur biaisé ! : pour tout $i = 1, \dots, n$

$$\mathbb{E}[(Y_i - \hat{r}_n(X_i))^2 | \hat{r}_n] \neq \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2 | \hat{r}_n]$$

car $\hat{r}_n$ dépend de (X_i, Y_i).

- En particulier, pour les estimateurs des MC, on a vu que $MCR(\hat{r}_n^D)$ diminue automatiquement quand l'espace S_D augmente, et que le MCR vaut 0 si $\dim(S_D(\mathbf{X})) = n$.

Moyenne des carrés des résidus et erreur de prédiction

- $E_n(\hat{r}) = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2]$ avec $\hat{r}_n$ calculé à partir de $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\dots,n}$.
- Idée naïve : estimer $E_n(\hat{r})$ par la moyenne des carrés des résidus :

$$\text{MCR}(\hat{r}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(X_i) - Y_i)^2$$

$\hookrightarrow$ C'est un estimateur biaisé ! : pour tout $i = 1, \dots, n$

$$\mathbb{E}[(Y_i - \hat{r}_n(X_i))^2 | \hat{r}_n] \neq \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2 | \hat{r}_n]$$

car $\hat{r}_n$ dépend de (X_i, Y_i).

- En particulier, pour les estimateurs des MC, on a vu que $MCR(\hat{r}_n^D)$ diminue automatiquement quand l'espace S_D augmente, et que le MCR vaut 0 si $\dim(S_D(\mathbf{X})) = n$.

Moyenne des carrés des résidus et erreur de prédiction

- $E_n(\hat{r}) = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2]$ avec $\hat{r}_n$ calculé à partir de $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\dots,n}$.
- Idée naïve : estimer $E_n(\hat{r})$ par la moyenne des carrés des résidus :

$$\text{MCR}(\hat{r}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(X_i) - Y_i)^2$$

$\hookrightarrow$ C'est un estimateur biaisé ! : pour tout $i = 1, \dots, n$

$$\mathbb{E}[(Y_i - \hat{r}_n(X_i))^2 | \hat{r}_n] \neq \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2 | \hat{r}_n]$$

car $\hat{r}_n$ dépend de (X_i, Y_i).

- En particulier, pour les estimateurs des MC, on a vu que $MCR(\hat{r}_n^D)$ diminue automatiquement quand l'espace S_D augmente, et que le MCR vaut 0 si $\dim(S_D(\mathbf{X})) = n$.

Moyenne des carrés des résidus et erreur de prédiction

- $E_n(\hat{r}) = \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2]$ avec $\hat{r}_n$ calculé à partir de $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\dots,n}$.
- Idée naïve : estimer $E_n(\hat{r})$ par la moyenne des carrés des résidus :

$$\text{MCR}(\hat{r}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(X_i) - Y_i)^2$$

$\hookrightarrow$ C'est un estimateur biaisé ! : pour tout $i = 1, \dots, n$

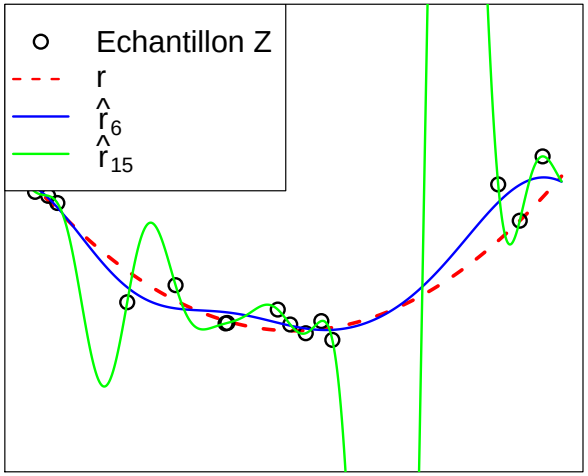
$$\mathbb{E}[(Y_i - \hat{r}_n(X_i))^2 | \hat{r}_n] \neq \mathbb{E}[(Y_{n+1} - \hat{r}_n(X_{n+1}))^2 | \hat{r}_n]$$

car $\hat{r}_n$ dépend de (X_i, Y_i).

- En particulier, pour les estimateurs des MC, on a vu que $MCR(\hat{r}_n^D)$ diminue automatiquement quand l'espace S_D augmente, et que le MCR vaut 0 si $\dim(S_D(\mathbf{X})) = n$.

Moyenne des carrés des résidus et erreur de prédiction (2)

- Exemple



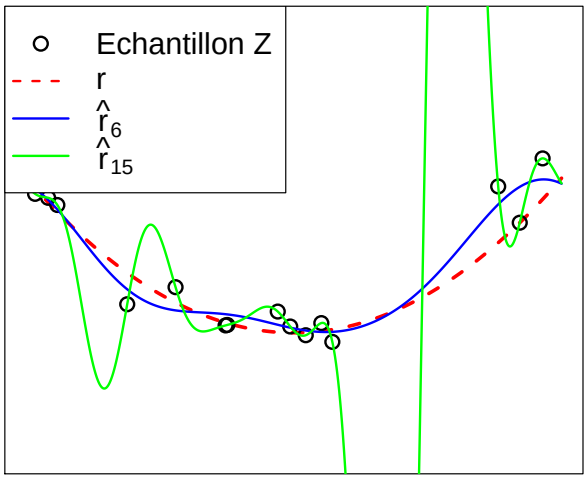
Soit $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\ldots,15}$. On considère deux procédures d'estimation : MC avec modèles trigo pour $D = 6$ et $D = 15$.

- Erreur de prédiction : $\text{MISE}_6 < \text{MISE}_{15} \Rightarrow E_6 < E_{15}$
- MCR : $\text{MCR}(\hat{r}_6) > \text{MCR}(\hat{r}_{15}) = 0$.

- Conclusion : La moyenne des carrés des résidus n'est pas un estimateur pertinent de l'erreur de prédiction
- D'une manière générale, la MCR est d'autant plus grande que le modèle est riche.
- Alternative : l'erreur de prédiction doit être calculée à partir d'un échantillon indépendant de celui utilisé pour l'estimation.

Moyenne des carrés des résidus et erreur de prédiction (2)

- Exemple



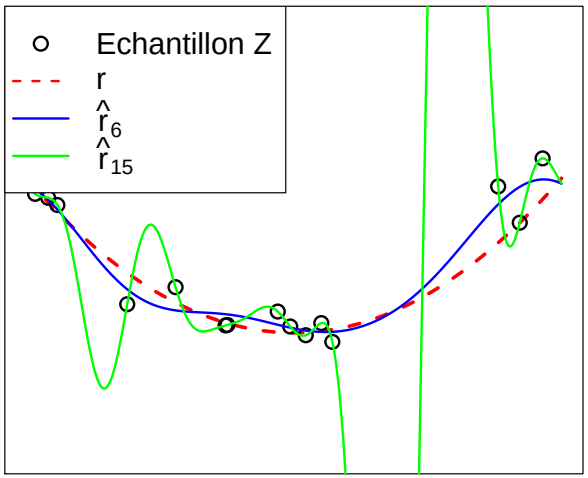
Soit $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\ldots,15}$. On considère deux procédures d'estimation : MC avec modèles trigo pour $D = 6$ et $D = 15$.

- Erreur de prédiction : $\text{MISE}_6 < \text{MISE}_{15}$ $\Rightarrow E_6 < E_{15}$
- MCR : $\text{MCR}(\hat{r}_6) > \text{MCR}(\hat{r}_{15}) = 0$.

- Conclusion : La moyenne des carrés des résidus n'est pas un estimateur pertinent de l'erreur de prédiction
- D'une manière générale, la MCR est d'autant plus grande que le modèle est riche.
- Alternative : l'erreur de prédiction doit être calculée à partir d'un échantillon indépendant de celui utilisé pour l'estimation.

Moyenne des carrés des résidus et erreur de prédiction (2)

- Exemple



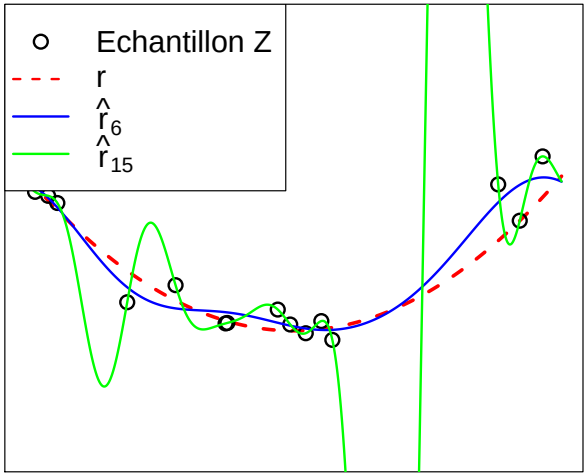
Soit $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\ldots,15}$. On considère deux procédures d'estimation : MC avec modèles trigo pour $D = 6$ et $D = 15$.

- Erreur de prédiction : $\text{MISE}_6 < \text{MISE}_{15}$ $\Rightarrow E_6 < E_{15}$
- MCR : $\text{MCR}(\hat{r}_6) > \text{MCR}(\hat{r}_{15}) = 0$.

- Conclusion : La moyenne des carrés des résidus n'est pas un estimateur pertinent de l'erreur de prédiction
- D'une manière générale, la MCR est d'autant plus grande que le modèle est riche.
- Alternative : l'erreur de prédiction doit être calculée à partir d'un échantillon indépendant de celui utilisé pour l'estimation.

Moyenne des carrés des résidus et erreur de prédiction (2)

- Exemple



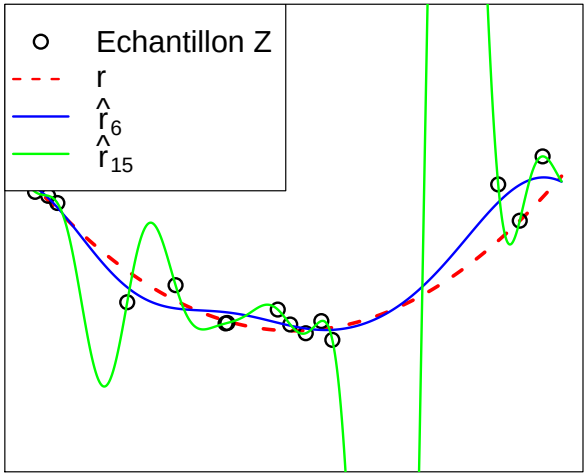
Soit $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\ldots,15}$. On considère deux procédures d'estimation : MC avec modèles trigo pour $D = 6$ et $D = 15$.

- Erreur de prédiction : $\text{MISE}_6 < \text{MISE}_{15}$ $\Rightarrow E_6 < E_{15}$
- MCR : $\text{MCR}(\hat{r}_6) > \text{MCR}(\hat{r}_{15}) = 0$.

- Conclusion : La moyenne des carrés des résidus n'est pas un estimateur pertinent de l'erreur de prédiction
- D'une manière générale, la MCR est d'autant plus grande que le modèle est riche.
- Alternative : l'erreur de prédiction doit être calculée à partir d'un échantillon indépendant de celui utilisé pour l'estimation.

Moyenne des carrés des résidus et erreur de prédiction (2)

- Exemple



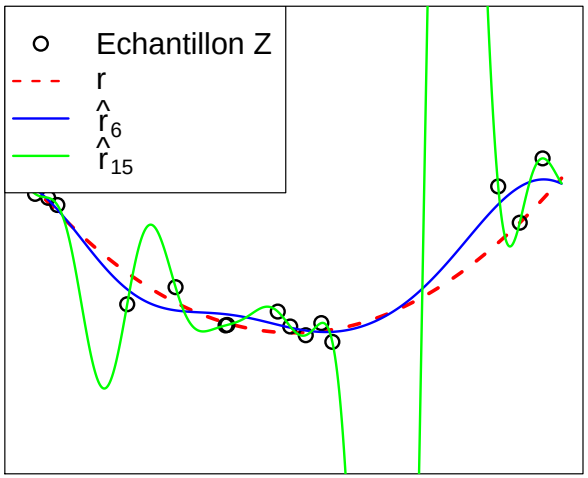
Soit $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\ldots,15}$. On considère deux procédures d'estimation : MC avec modèles trigo pour $D = 6$ et $D = 15$.

- Erreur de prédiction : $\text{MISE}_6 < \text{MISE}_{15}$ $\Rightarrow E_6 < E_{15}$
- MCR : $\text{MCR}(\hat{r}_6) > \text{MCR}(\hat{r}_{15}) = 0$.

- Conclusion : La moyenne des carrés des résidus n'est pas un estimateur pertinent de l'erreur de prédiction
- D'une manière générale, la MCR est d'autant plus grande que le modèle est riche.
- Alternative : l'erreur de prédiction doit être calculée à partir d'un échantillon indépendant de celui utilisé pour l'estimation.

Moyenne des carrés des résidus et erreur de prédiction (2)

- Exemple



Soit $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\ldots,15}$. On considère deux procédures d'estimation : MC avec modèles trigo pour $D = 6$ et $D = 15$.

- Erreur de prédiction : $\text{MISE}_6 < \text{MISE}_{15}$ $\Rightarrow E_6 < E_{15}$
- MCR : $\text{MCR}(\hat{r}_6) > \text{MCR}(\hat{r}_{15}) = 0$.

- Conclusion : La moyenne des carrés des résidus n'est pas un estimateur pertinent de l'erreur de prédiction
- D'une manière générale, la MCR est d'autant plus grande que le modèle est riche.
- Alternative : l'erreur de prédiction doit être calculée à partir d'un échantillon indépendant de celui utilisé pour l'estimation.

Procédure de test-retest

Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ i.i.d. avec $r(x) = \mathbb{E}[Y|X = x]$, et $\hat{r}$ une procédure d'estimation.

- $\mathcal{Z}$ est coupé en deux échantillons indépendants :

$$\begin{cases} \mathcal{Z}^{test} = (X_i, Y_i)_{i=1,\dots,n_0} \\ \mathcal{Z}^{retest} = (X_i, Y_i)_{i=n_0+1,\dots,n} \end{cases}$$

- $\mathcal{Z}^{test}$: calcul de l'estimateur $\hat{r}_{n_0}$
- $\mathcal{Z}^{retest}$: calcul de l'erreur de prédiction

$$\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest}) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} (Y_i - \hat{r}_{n_0}(X_i))^2$$

Comme $\mathcal{Z}^{test}$ and $\mathcal{Z}^{retest}$ sont indépendants :

$$\begin{aligned} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest})] &= \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_i, Y_i))] \\ &= \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_{n_0+1}, Y_{n_0+1}))] = E_{n_0}(\hat{r}) \end{aligned}$$

Procédure de test-retest

Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ i.i.d. avec $r(x) = \mathbb{E}[Y|X = x]$, et $\hat{r}$ une procédure d'estimation.

- $\mathcal{Z}$ est coupé en deux échantillons indépendants :

$$\begin{cases} \mathcal{Z}^{test} = (X_i, Y_i)_{i=1,\dots,n_0} \\ \mathcal{Z}^{retest} = (X_i, Y_i)_{i=n_0+1,\dots,n} \end{cases}$$

- $\mathcal{Z}^{test}$: calcul de l'estimateur $\hat{r}_{n_0}$
- $\mathcal{Z}^{retest}$: calcul de l'erreur de prédiction

$$\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest}) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} (Y_i - \hat{r}_{n_0}(X_i))^2$$

Comme $\mathcal{Z}^{test}$ and $\mathcal{Z}^{retest}$ sont indépendants :

$$\begin{aligned} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest})] &= \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_i, Y_i))] \\ &= \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_{n_0+1}, Y_{n_0+1}))] = E_{n_0}(\hat{r}) \end{aligned}$$

Procédure de test-retest

Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ i.i.d. avec $r(x) = \mathbb{E}[Y|X = x]$, et $\hat{r}$ une procédure d'estimation.

- $\mathcal{Z}$ est coupé en deux échantillons indépendants :

$$\begin{cases} \mathcal{Z}^{test} = (X_i, Y_i)_{i=1,\dots,n_0} \\ \mathcal{Z}^{retest} = (X_i, Y_i)_{i=n_0+1,\dots,n} \end{cases}$$

- $\mathcal{Z}^{test}$: calcul de l'estimateur $\hat{r}_{n_0}$
- $\mathcal{Z}^{retest}$: calcul de l'erreur de prédiction

$$\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest}) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} (Y_i - \hat{r}_{n_0}(X_i))^2$$

Comme $\mathcal{Z}^{test}$ and $\mathcal{Z}^{retest}$ sont indépendants :

$$\begin{aligned} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest})] &= \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_i, Y_i))] \\ &= \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_{n_0+1}, Y_{n_0+1}))] = E_{n_0}(\hat{r}) \end{aligned}$$

Procédure de test-retest

Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ i.i.d. avec $r(x) = \mathbb{E}[Y|X = x]$, et $\hat{r}$ une procédure d'estimation.

- $\mathcal{Z}$ est coupé en deux échantillons indépendants :

$$\begin{cases} \mathcal{Z}^{test} = (X_i, Y_i)_{i=1,\dots,n_0} \\ \mathcal{Z}^{retest} = (X_i, Y_i)_{i=n_0+1,\dots,n} \end{cases}$$

- $\mathcal{Z}^{test}$: calcul de l'estimateur $\hat{r}_{n_0}$
- $\mathcal{Z}^{retest}$: calcul de l'erreur de prédiction

$$\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest}) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} (Y_i - \hat{r}_{n_0}(X_i))^2$$

Comme $\mathcal{Z}^{test}$ and $\mathcal{Z}^{retest}$ sont indépendants :

$$\begin{aligned} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest})] &= \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_i, Y_i))] \\ &= \mathbb{E}[\text{Err}(\hat{r}_{n_0}, (X_{n_0+1}, Y_{n_0+1}))] = E_{n_0}(\hat{r}) \end{aligned}$$

Procédure de test-retest

Soit $\mathcal{Z} = (X_i, Y_i)_{i=1,\dots,n}$ i.i.d. avec $r(x) = \mathbb{E}[Y|X = x]$, et $\hat{r}$ une procédure d'estimation.

- $\mathcal{Z}$ est coupé en deux échantillons indépendants :

$$\begin{cases} \mathcal{Z}^{test} = (X_i, Y_i)_{i=1,\dots,n_0} \\ \mathcal{Z}^{retest} = (X_i, Y_i)_{i=n_0+1,\dots,n} \end{cases}$$

- $\mathcal{Z}^{test}$: calcul de l'estimateur $\hat{r}_{n_0}$
- $\mathcal{Z}^{retest}$: calcul de l'erreur de prédiction

$$\mathsf{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest}) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} (Y_i - \hat{r}_{n_0}(X_i))^2$$

Comme $\mathcal{Z}^{test}$ and $\mathcal{Z}^{retest}$ sont indépendants :

$$\begin{aligned} \mathbb{E}[\mathsf{Err}(\hat{r}_{n_0}, \mathcal{Z}^{retest})] &= \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \mathbb{E}[\mathsf{Err}(\hat{r}_{n_0}, (X_i, Y_i))] \\ &= \mathbb{E}[\mathsf{Err}(\hat{r}_{n_0}, (X_{n_0+1}, Y_{n_0+1}))] = E_{n_0}(\hat{r}) \end{aligned}$$

Procédure de test-retest

- Choix de n_0 :
  - $n - n_0$ petit : l'erreur de prédiction $E_{n_0}(\hat{r})$ n'est pas bien estimée (moyenne empirique basée sur peu d'observations)
  - n_0 petit : l'estimateur $\hat{r}_{n_0}$ n'est pas bien estimé.
- Procédure de test-retest : une partie des données ne sont pas utilisées pour l'estimation $\rightarrow$ perte d'information.
- La validation croisée permet de contourner ce problème.

Procédure de test-retest

- Choix de n_0 :
  - $n - n_0$ petit : l'erreur de prédiction $E_{n_0}(\hat{r})$ n'est pas bien estimée (moyenne empirique basée sur peu d'observations)
  - n_0 petit : l'estimateur $\hat{r}_{n_0}$ n'est pas bien estimé.
- Procédure de test-retest : une partie des données ne sont pas utilisées pour l'estimation $\rightarrow$ perte d'information.
- La validation croisée permet de contourner ce problème.

Procédure de test-retest

- Choix de n_0 :
  - $n - n_0$ petit : l'erreur de prédiction $E_{n_0}(\hat{r})$ n'est pas bien estimée (moyenne empirique basée sur peu d'observations)
  - n_0 petit : l'estimateur $\hat{r}_{n_0}$ n'est pas bien estimé.
- Procédure de test-retest : une partie des données ne sont pas utilisées pour l'estimation $\rightarrow$ perte d'information.
- La validation croisée permet de contourner ce problème.

Procédure de test-retest

- Choix de n_0 :
  - $n - n_0$ petit : l'erreur de prédiction $E_{n_0}(\hat{r})$ n'est pas bien estimée (moyenne empirique basée sur peu d'observations)
  - n_0 petit : l'estimateur $\hat{r}_{n_0}$ n'est pas bien estimé.
- Procédure de test-retest : une partie des données ne sont pas utilisées pour l'estimation $\rightarrow$ perte d'information.
- La validation croisée permet de contourner ce problème.

Procédure de test-retest

- Choix de n_0 :
  - $n - n_0$ petit : l'erreur de prédiction $E_{n_0}(\hat{r})$ n'est pas bien estimée (moyenne empirique basée sur peu d'observations)
  - n_0 petit : l'estimateur $\hat{r}_{n_0}$ n'est pas bien estimé.
- Procédure de test-retest : une partie des données ne sont pas utilisées pour l'estimation $\rightarrow$ perte d'information.
- La validation croisée permet de contourner ce problème.

Erreur de prédiction : validation croisée

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\dots,n}$ un n-échantillon et $\hat{r}$ une procédure d'estimation.
- $\mathcal{Z}$ est partitionné en k sous échantillons de taille égale :

$$\{1, \dots, n\} = I_1 \sqcup \cdots \sqcup I_k.$$

- Pour chaque $\ell = 1 \dots, k$, on définit :
  - l'ensemble d'apprentissage sur lequel l'estimateur est calculé

$$\mathcal{Z}^{train,\ell} = \{(X_i, Y_i), i \notin I_\ell\}$$

  - l'ensemble de validation sur lequel l'erreur de prédiction est calculée.

$$\mathcal{Z}^{val,\ell} = \{(X_i, Y_i), i \in I_\ell\}$$

Erreur de prédiction : validation croisée

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\ldots,n}$ un n-échantillon et $\hat{r}$ une procédure d'estimation.
- $\mathcal{Z}$ est partitionné en k sous échantillons de taille égale :

$$\{1, \ldots, n\} = I_1 \sqcup \cdots \sqcup I_k.$$

- Pour chaque $\ell = 1 \ldots, k$, on définit :
  - l'ensemble d'apprentissage sur lequel l'estimateur est calculé

$$\mathcal{Z}^{train,\ell} = \{(X_i, Y_i), i \notin I_\ell\}$$

  - l'ensemble de validation sur lequel l'erreur de prédiction est calculée.

$$\mathcal{Z}^{val,\ell} = \{(X_i, Y_i), i \in I_\ell\}$$

Plus précisément,

- Pour $\ell = 1 \ldots, k$,
  - $\hat{r}^{<\ell>}_{n-\frac{n}{k}}$: estimateur calculé à partir de $\mathcal{Z}^{train,\ell}$
  - Erreur de prédiction calculée sur $\mathcal{Z}^{val,\ell}$:

$$\text{Err}(\hat{r}^{<\ell>}_{n-\frac{n}{k}}, \mathcal{Z}^{val,\ell}) = \frac{1}{\#I_\ell} \sum_{i \in I_\ell} (Y_i - \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i))^2$$

Alors

$$\mathbb{E}[\text{Err}(\hat{r}^{<\ell>}_{n-\frac{n}{k}}, \mathcal{Z}^{val,\ell})] = E_{n-\frac{n}{k}}(\hat{r})$$

- L'erreur de prédiction par VC est alors définie comme :

$$E^{VC}(\hat{r}) = \frac{1}{k} \sum_{\ell=1}^{k} \text{Err}(\hat{r}^{<\ell>}_{n-\frac{n}{k}}, \mathcal{Z}^{val,\ell})$$

Elle constitue un estimateur sans biais de $E_{n-\frac{n}{k}}(\hat{r})$.

Plus précisément,

- Pour $\ell = 1 \ldots, k$,
  - $\hat{r}^{<\ell>}_{n-\frac{n}{k}}$: estimateur calculé à partir de $\mathcal{Z}^{train,\ell}$
  - Erreur de prédiction calculée sur $\mathcal{Z}^{val,\ell}$:

$$\text{Err}(\hat{r}^{<\ell>}_{n-\frac{n}{k}}, \mathcal{Z}^{val,\ell}) = \frac{1}{\#I_\ell} \sum_{i \in I_\ell} (Y_i - \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i))^2$$

Alors

$$\mathbb{E}[\text{Err}(\hat{r}^{<\ell>}_{n-\frac{n}{k}}, \mathcal{Z}^{val,\ell})] = E_{n-\frac{n}{k}}(\hat{r})$$

- L'erreur de prédiction par VC est alors définie comme :

$$E^{VC}(\hat{r}) = \frac{1}{k} \sum_{\ell=1}^{k} \text{Err}(\hat{r}^{<\ell>}_{n-\frac{n}{k}}, \mathcal{Z}^{val,\ell})$$

Elle constitue un estimateur sans biais de $E_{n-\frac{n}{k}}(\hat{r})$.

Utilisation de la VC : comparaison d'estimateurs.

- La VC peut permettre sous certaines conditions de choisir entre plusieurs estimateurs

↪ **Exple** : Soit $\hat{r}^{kern}$ la procédure de Nadaraya Watson pour une fenêtre fixée et $\hat{r}^{kNN}$ la procédure des k plus proches voisins avec k fixé. On choisira comme meilleur estimateur en termes de prédiction :

$$\arg \min_{\hat{r} \in \{\hat{r}^{kern}, \hat{r}^{kNN}\}} E^{cv}(\hat{r})$$

- Exple où la VC n'est pas appropriée : Soit $\{\hat{r}^D, D = 1, \ldots, n\}$ les procédures d'estimation des MC dans la base trigo de dimension D.
  - $E^{cv}(\hat{r}^D)$ estimateur sans biais de $E_{n-\frac{n}{k}}(\hat{r})$
  - Or le D optimal dépend de la taille d'échantillon ("$D^* = Cn^{1/(2\alpha+1)}$") :

$$\arg \min_{D=1,\ldots,n} E_{n-\frac{n}{k}}(\hat{r}^D) = \arg \min_{D=1,\ldots,n} MISE_{n-\frac{n}{k}}(\hat{r}^D)$$

$$\neq \arg \min_{D=1,\ldots,n} MISE_n(\hat{r}^D) = \arg \min_{D=1,\ldots,n} E_n(\hat{r})$$

Utilisation de la VC : comparaison d'estimateurs.

- La VC peut permettre sous certaines conditions de choisir entre plusieurs estimateurs

↪ **Exple** : Soit $\hat{r}^{kern}$ la procédure de Nadaraya Watson pour une fenêtre fixée et $\hat{r}^{kNN}$ la procédure des k plus proches voisins avec k fixé. On choisira comme meilleur estimateur en termes de prédiction :

$$\arg \min_{\hat{r} \in \{\hat{r}^{kern}, \hat{r}^{kNN}\}} E^{cv}(\hat{r})$$

- Exple où la VC n'est pas appropriée : Soit $\{\hat{r}^D, D = 1, \ldots, n\}$ les procédures d'estimation des MC dans la base trigo de dimension D.
  - $E^{cv}(\hat{r}^D)$ estimateur sans biais de $E_{n-\frac{n}{k}}(\hat{r}^D)$
  - Or le D optimal dépend de la taille d'échantillon ("$D^* = Cn^{1/(2\alpha+1)}$") :

$$\arg \min_{D=1,\ldots,n} E_{n-\frac{n}{k}}(\hat{r}^D) = \arg \min_{D=1,\ldots,n} MISE_{n-\frac{n}{k}}(\hat{r}^D)$$

$$\neq \arg \min_{D=1,\ldots,n} MISE_n(\hat{r}^D) = \arg \min_{D=1,\ldots,n} E_n(\hat{r})$$

Utilisation de la VC : comparaison d'estimateurs.

- La VC peut permettre sous certaines conditions de choisir entre plusieurs estimateurs

↪ **Exple** : Soit $\hat{r}^{kern}$ la procédure de Nadaraya Watson pour une fenêtre fixée et $\hat{r}^{kNN}$ la procédure des k plus proches voisins avec k fixé. On choisira comme meilleur estimateur en termes de prédiction :

$$\arg \min_{\hat{r} \in \{\hat{r}^{kern}, \hat{r}^{kNN}\}} E^{cv}(\hat{r})$$

- Exple où la VC n'est pas appropriée : Soit $\{\hat{r}^D, D = 1, \ldots, n\}$ les procédures d'estimation des MC dans la base trigo de dimension D.
  - $E^{cv}(\hat{r}^D)$ estimateur sans biais de $E_{n-\frac{n}{k}}(\hat{r})$
  - Or le D optimal dépend de la taille d'échantillon ("$D^* = Cn^{1/(2\alpha+1)}$") :

$$\arg \min_{D=1,\ldots,n} E_{n-\frac{n}{k}}(\hat{r}^D) = \arg \min_{D=1,\ldots,n} MISE_{n-\frac{n}{k}}(\hat{r}^D)$$

$$\neq \arg \min_{D=1,\ldots,n} MISE_n(\hat{r}^D) = \arg \min_{D=1,\ldots,n} E_n(\hat{r})$$

Utilisation de la VC : comparaison d'estimateurs.

- La VC peut permettre sous certaines conditions de choisir entre plusieurs estimateurs

↪ **Exple** : Soit $\hat{r}^{kern}$ la procédure de Nadaraya Watson pour une fenêtre fixée et $\hat{r}^{kNN}$ la procédure des k plus proches voisins avec k fixé. On choisira comme meilleur estimateur en termes de prédiction :

$$\arg \min_{\hat{r} \in \{\hat{r}^{kern}, \hat{r}^{kNN}\}} E^{cv}(\hat{r})$$

- Exple où la VC n'est pas appropriée : Soit $\{\hat{r}^D, D = 1, \ldots, n\}$ les procédures d'estimation des MC dans la base trigo de dimension D.
  - $E^{cv}(\hat{r}^D)$ estimateur sans biais de $E_{n-\frac{n}{k}}(\hat{r})$
  - Or le D optimal dépend de la taille d'échantillon ("$D^* = Cn^{1/(2\alpha+1)}$") :

$$\arg \min_{D=1,\ldots,n} E_{n-\frac{n}{k}}(\hat{r}^D) = \arg \min_{D=1,\ldots,n} MISE_{n-\frac{n}{k}}(\hat{r}^D)$$

$$\neq \arg \min_{D=1,\ldots,n} MISE_n(\hat{r}^D) = \arg \min_{D=1,\ldots,n} E_n(\hat{r})$$

- Ainsi, la VC permet de choisir entre des estimateurs selon leur capacités prédictives en supposant que le meilleur estimateur ne dépend pas de n
  - En pratique, la VC avec $k = n$ est parfois utilisée en supposant que pour les procédures $\hat{r}$ considérées :
$$MISE_{n-1}(\hat{r}) \approx MISE_n(\hat{r})$$
  - **Rq :** Dans des contextes spécifiques, il existe des méthodes rigoureuses où on définit h_n tel que
$$\mathbb{E}[h_n]\text{MISE}_{n-1} = \text{MISE}_n$$
- VC très générale : pas d'hypothèse sur les données (sauf i.i.d.) et permet de comparer des estimateurs de nature différente.
  - Exple : comparaison entre un estimateur NP et un estimateur paramétrique pour un modèle candidat.

- Ainsi, la VC permet de choisir entre des estimateurs selon leur capacités prédictives en supposant que le meilleur estimateur ne dépend pas de n
  - En pratique, la VC avec $k = n$ est parfois utilisée en supposant que pour les procédures $\hat{r}$ considérées :

$$MISE_{n-1}(\hat{r}) \approx MISE_n(\hat{r})$$

  - **Rq :** Dans des contextes spécifiques, il existe des méthodes rigoureuses où on définit h_n tel que

$$\mathbb{E}[h_n]\text{MISE}_{n-1} = \text{MISE}_n$$

- VC très générale : pas d'hypothèse sur les données (sauf i.i.d.) et permet de comparer des estimateurs de nature différente.
  - Exple : comparaison entre un estimateur NP et un estimateur paramétrique pour un modèle candidat.

- Ainsi, la VC permet de choisir entre des estimateurs selon leur capacités prédictives en supposant que le meilleur estimateur ne dépend pas de n
  - En pratique, la VC avec $k = n$ est parfois utilisée en supposant que pour les procédures $\hat{r}$ considérées :

$$MISE_{n-1}(\hat{r}) \approx MISE_n(\hat{r})$$

  - **Rq :** Dans des contextes spécifiques, il existe des méthodes rigoureuses où on définit h_n tel que

$$\mathbb{E}[h_n]\text{MISE}_{n-1} = \text{MISE}_n$$

- VC très générale : pas d'hypothèse sur les données (sauf i.i.d.) et permet de comparer des estimateurs de nature différente.
  - Exple : comparaison entre un estimateur NP et un estimateur paramétrique pour un modèle candidat.

- Ainsi, la VC permet de choisir entre des estimateurs selon leur capacités prédictives en supposant que le meilleur estimateur ne dépend pas de n
  - En pratique, la VC avec $k = n$ est parfois utilisée en supposant que pour les procédures $\hat{r}$ considérées :
$$MISE_{n-1}(\hat{r}) \approx MISE_n(\hat{r})$$
  - **Rq :** Dans des contextes spécifiques, il existe des méthodes rigoureuses où on définit h_n tel que
$$\mathbb{E}[h_n]\text{MISE}_{n-1} = \text{MISE}_n$$
- VC très générale : pas d'hypothèse sur les données (sauf i.i.d.) et permet de comparer des estimateurs de nature différente.
  - Exple : comparaison entre un estimateur NP et un estimateur paramétrique pour un modèle candidat.

- Ainsi, la VC permet de choisir entre des estimateurs selon leur capacités prédictives en supposant que le meilleur estimateur ne dépend pas de n
  - En pratique, la VC avec $k = n$ est parfois utilisée en supposant que pour les procédures $\hat{r}$ considérées :
$$MISE_{n-1}(\hat{r}) \approx MISE_n(\hat{r})$$
  - **Rq :** Dans des contextes spécifiques, il existe des méthodes rigoureuses où on définit h_n tel que
$$\mathbb{E}[h_n]\text{MISE}_{n-1} = \text{MISE}_n$$
- VC très générale : pas d'hypothèse sur les données (sauf i.i.d.) et permet de comparer des estimateurs de nature différente.
  - Exple : comparaison entre un estimateur NP et un estimateur paramétrique pour un modèle candidat.

Utilisation de la VC 2 : évaluation des performances d'un estimateur

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\dots,n}$ un échantillon i.i.d, on veut évaluer les capacités de X à prédire Y

↪ *Ex : $Y =$ tp de survie d'un patient après opération, $X =$paramètres cliniques avant opération.*

- On dispose d'une procédure d'estimation "fiable" (peu d'hypothèse sous-jacentes, modèle connu, etc)
- Soit $\{1, \dots, n\} = I_1 \sqcup \cdots \sqcup I_k$.
- Pour tout $i = 1, \dots, n$ soit ℓ tq $i \in I_\ell$ et

$$\hat{Y}_i = \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i) \quad \text{la prédiction de } Y_i$$

- On regarde $(Y_i, \hat{Y}_i)_{i=1,\dots,n}$

Utilisation de la VC 2 : évaluation des performances d'un estimateur

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\dots,n}$ un échantillon i.i.d, on veut évaluer les capacités de X à prédire Y

↪ *Ex : $Y =$ tp de survie d'un patient après opération, $X =$paramètres cliniques avant opération.*

- On dispose d'une procédure d'estimation "fiable" (peu d'hypothèse sous-jacentes, modèle connu, etc)
- Soit $\{1, \dots, n\} = I_1 \sqcup \cdots \sqcup I_k$.
- Pour tout $i = 1, \dots, n$ soit ℓ tq $i \in I_\ell$ et

$$\hat{Y}_i = \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i) \quad \text{la prédiction de } Y_i$$

- On regarde $(Y_i, \hat{Y}_i)_{i=1,\dots,n}$

Utilisation de la VC 2 : évaluation des performances d'un estimateur

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\dots,n}$ un échantillon i.i.d, on veut évaluer les capacités de X à prédire Y

↪ *Ex : $Y =$ tp de survie d'un patient après opération, $X =$paramètres cliniques avant opération.*

- On dispose d'une procédure d'estimation "fiable" (peu d'hypothèse sous-jacentes, modèle connu, etc)
- Soit $\{1, \dots, n\} = I_1 \sqcup \cdots \sqcup I_k$.
- Pour tout $i = 1, \dots, n$ soit ℓ tq $i \in I_\ell$ et

$$\hat{Y}_i = \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i) \quad \text{la prédiction de } Y_i$$

- On regarde $(Y_i, \hat{Y}_i)_{i=1,\dots,n}$

Utilisation de la VC 2 : évaluation des performances d'un estimateur

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\dots,n}$ un échantillon i.i.d, on veut évaluer les capacités de X à prédire Y

↪ *Ex : $Y =$ tp de survie d'un patient après opération, $X =$paramètres cliniques avant opération.*

- On dispose d'une procédure d'estimation "fiable" (peu d'hypothèse sous-jacentes, modèle connu, etc)
- Soit $\{1, \dots, n\} = I_1 \sqcup \dots \sqcup I_k$.
- Pour tout $i = 1, \dots, n$ soit ℓ tq $i \in I_\ell$ et

$$\hat{Y}_i = \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i) \quad \text{la prédiction de } Y_i$$

- On regarde $(Y_i, \hat{Y}_i)_{i=1,\dots,n}$

Utilisation de la VC 2 : évaluation des performances d'un estimateur

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\dots,n}$ un échantillon i.i.d, on veut évaluer les capacités de X à prédire Y

↪ *Ex : $Y =$ tp de survie d'un patient après opération, $X =$paramètres cliniques avant opération.*

- On dispose d'une procédure d'estimation "fiable" (peu d'hypothèse sous-jacentes, modèle connu, etc)
- Soit $\{1, \dots, n\} = I_1 \sqcup \cdots \sqcup I_k$.
- Pour tout $i = 1, \dots, n$ soit ℓ tq $i \in I_\ell$ et

$$\hat{Y}_i = \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i) \quad \text{la prédiction de } Y_i$$

- On regarde $(Y_i, \hat{Y}_i)_{i=1,\dots,n}$

Utilisation de la VC 2 : évaluation des performances d'un estimateur

- Soit $\mathcal{Z} = \{(X_i, Y_i\}_{i=1,\dots,n}$ un échantillon i.i.d, on veut évaluer les capacités de X à prédire Y

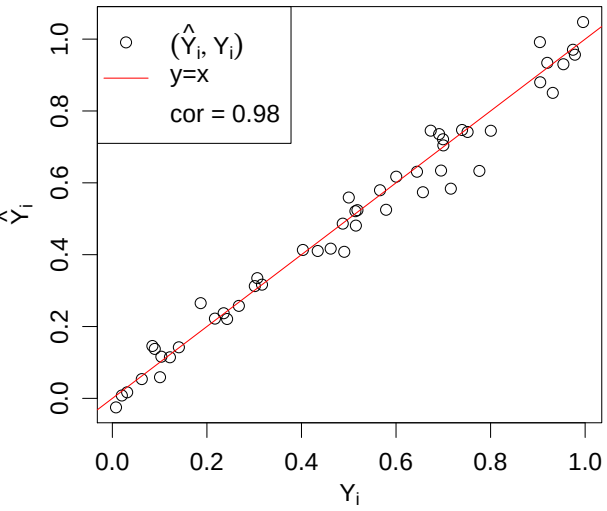
↪ *Ex : $Y =$ tp de survie d'un patient après opération, $X =$paramètres cliniques avant opération.*

- On dispose d'une procédure d'estimation "fiable" (peu d'hypothèse sous-jacentes, modèle connu, etc)
- Soit $\{1, \dots, n\} = I_1 \sqcup \cdots \sqcup I_k$.
- Pour tout $i = 1, \dots, n$ soit ℓ tq $i \in I_\ell$ et

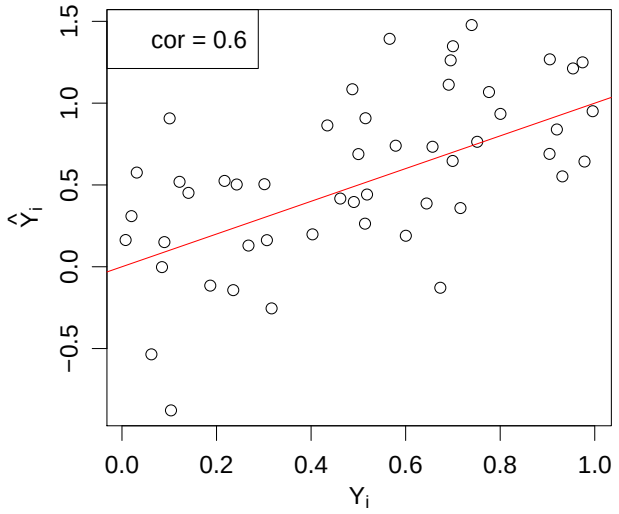
$$\hat{Y}_i = \hat{r}^{<\ell>}_{n-\frac{n}{k}}(X_i) \quad \text{la prédiction de } Y_i$$

- On regarde $(Y_i, \hat{Y}_i)_{i=1,\dots,n}$

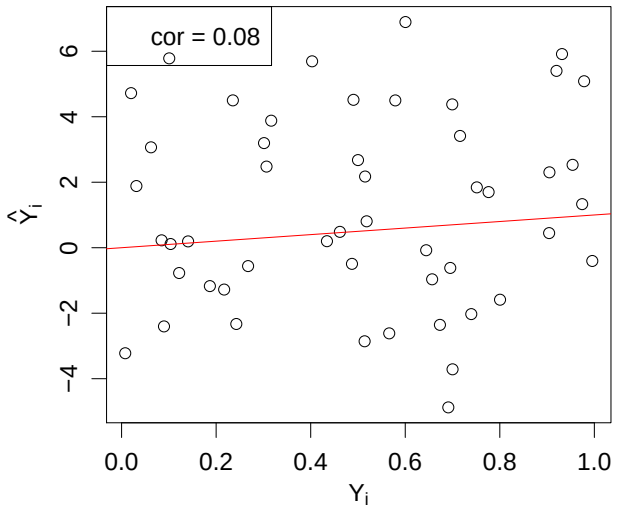
- Exemple



Prédiction ++
Corrélation ++



Prédiction -
Corrélation +

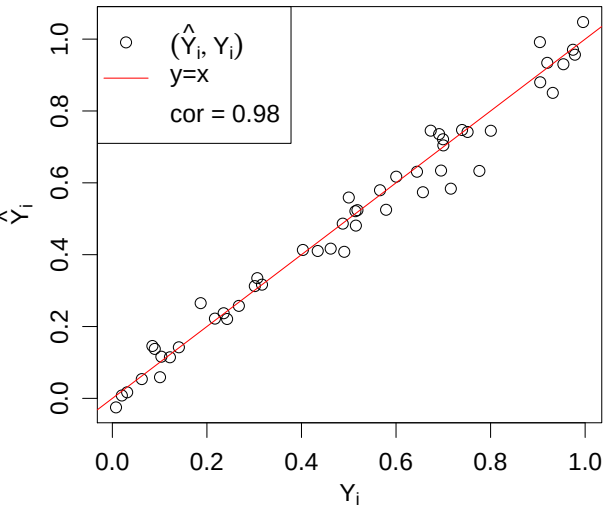


Prédiction –
Corrélation -

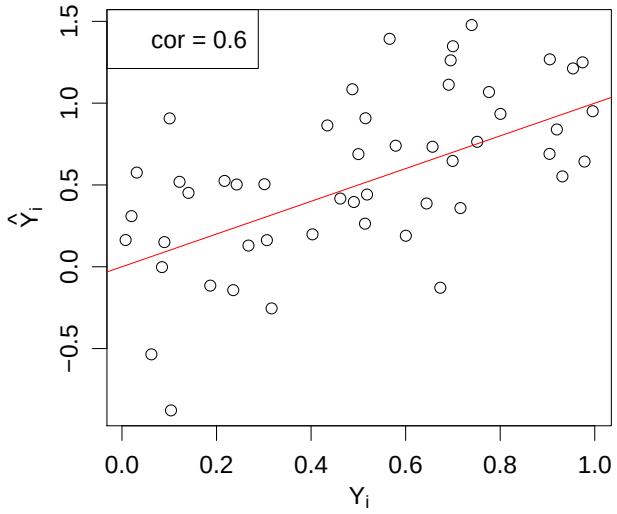
- La correlation de Pearson est descriptive, on ne peut pas directement effectuer un test car les échantillons (Y_i) et $(\hat{Y}_i)$ ne sont pas indépendants.
- Pour évaluer les performance d'un estimateur tous les paramètres de l'estimateur $\hat{r}^{<\ell>}_{n-\frac{n}{k}}$ doivent être calculés sur l'ensemble d'apprentissage $\mathcal{Z}^{train,\ell}$.

↪ Si on veut évaluer des paramètres par VC, il faut refaire une VC à l'intérieur de chaque $\mathcal{Z}^{train,\ell}$.

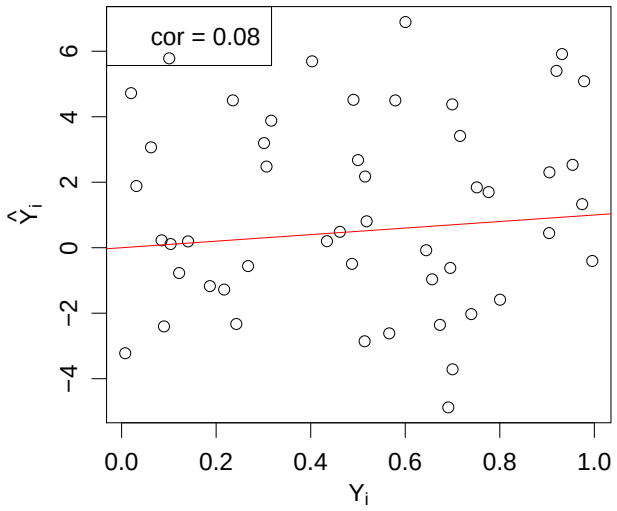
- Exemple



Prédiction ++
Corrélation ++



Prédiction -
Corrélation +

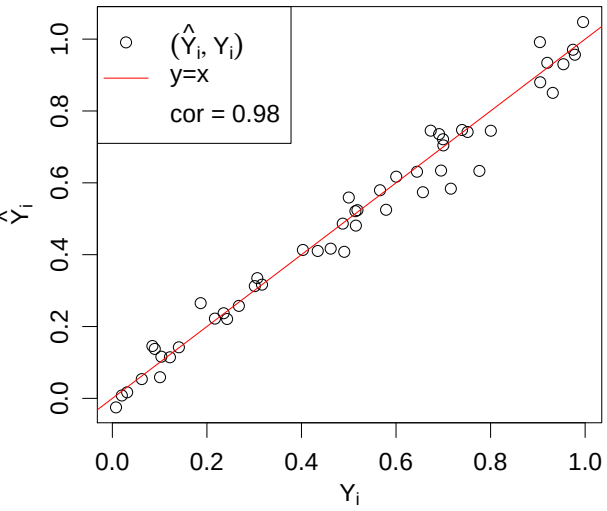


Prédiction –
Corrélation -

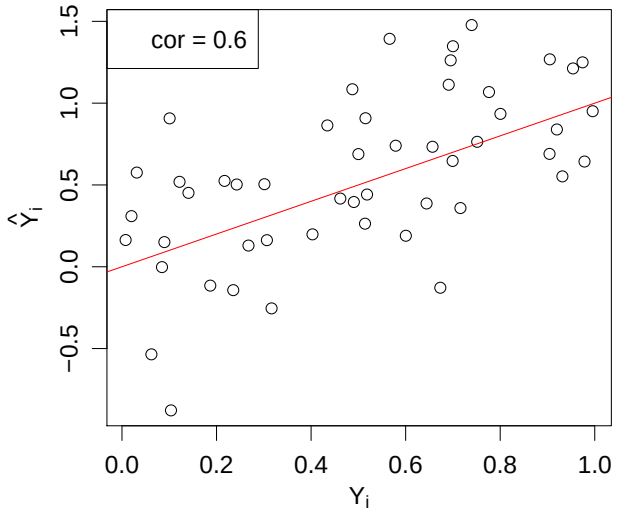
- La correlation de Pearson est descriptive, on ne peut pas directement effectuer un test car les échantillons (Y_i) et $(\hat{Y}_i)$ ne sont pas indépendants.
- Pour évaluer les performance d'un estimateur tous les paramètres de l'estimateur $\hat{r}^{<\ell>}_{n-\frac{n}{k}}$ doivent être calculés sur l'ensemble d'apprentissage $\mathcal{Z}^{train,\ell}$.

↪ Si on veut évaluer des paramètres par VC, il faut refaire une VC à l'intérieur de chaque $\mathcal{Z}^{train,\ell}$.

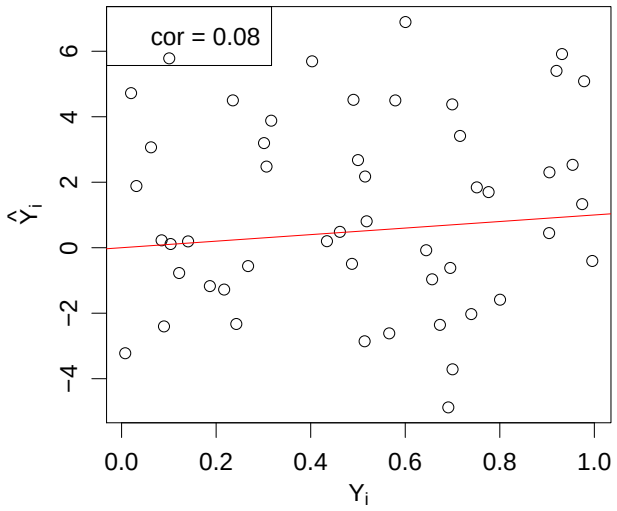
- Exemple



Prédiction ++
Corrélation ++



Prédiction -
Corrélation +

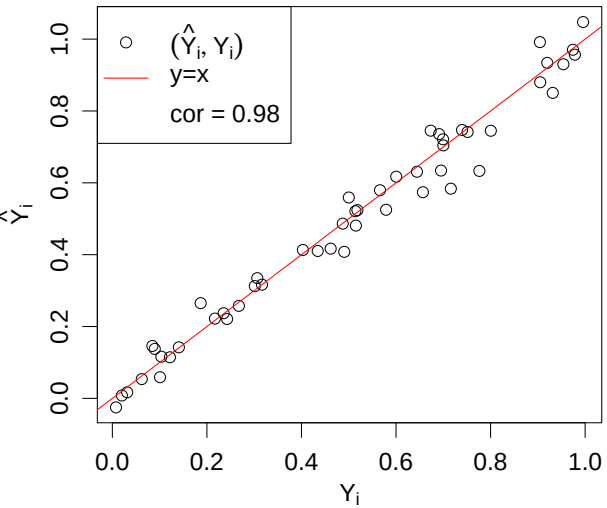


Prédiction –
Corrélation -

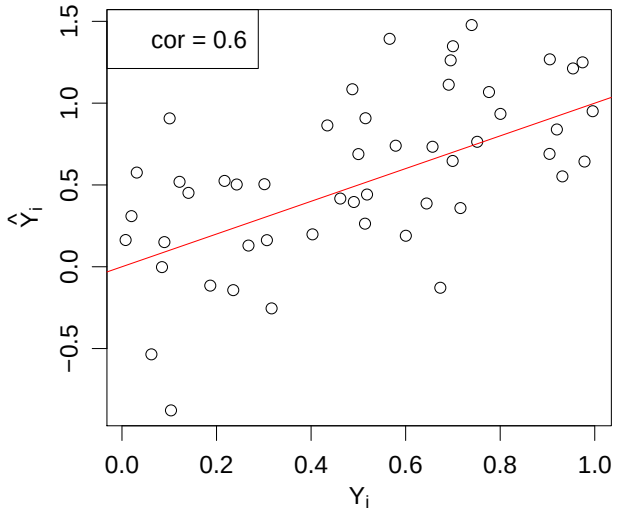
- La correlation de Pearson est descriptive, on ne peut pas directement effectuer un test car les échantillons (Y_i) et $(\hat{Y}_i)$ ne sont pas indépendants.
- Pour évaluer les performance d'un estimateur tous les paramètres de l'estimateur $\hat{r}^{<\ell>}_{n-\frac{n}{k}}$ doivent être calculés sur l'ensemble d'apprentissage $\mathcal{Z}^{train,\ell}$.

↪ Si on veut évaluer des paramètres par VC, il faut refaire une VC à l'intérieur de chaque $\mathcal{Z}^{train,\ell}$.

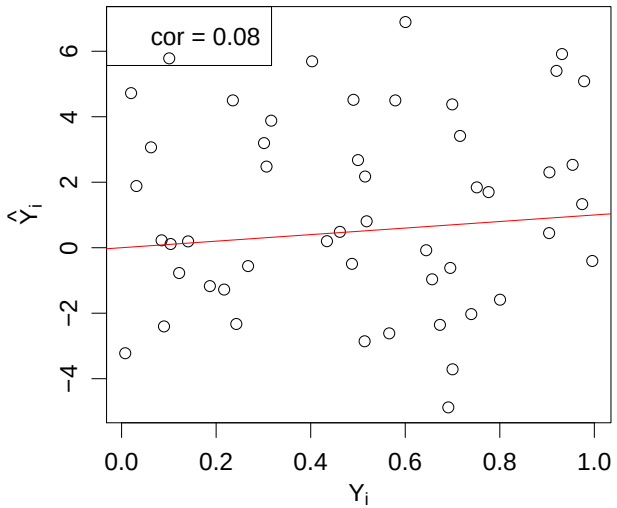
- Exemple



Prédiction ++
Corrélation ++



Prédiction -
Corrélation +



Prédiction –
Corrélation -

- La correlation de Pearson est descriptive, on ne peut pas directement effectuer un test car les échantillons (Y_i) et $(\hat{Y}_i)$ ne sont pas indépendants.
- Pour évaluer les performance d'un estimateur tous les paramètres de l'estimateur $\hat{r}^{<\ell>}_{n-\frac{n}{k}}$ doivent être calculés sur l'ensemble d'apprentissage $\mathcal{Z}^{train,\ell}$.

↪ Si on veut évaluer des paramètres par VC, il faut refaire une VC à l'intérieur de chaque $\mathcal{Z}^{train,\ell}$.

Choix du nombre de "paquets" k

- Si on divise les observations en k validation sets :
  - Chaque $\hat{r}^{\ell}$ est calculé à partir de $n(1-1/k)$ observations.
  - Nombre total d'observations utilisées pour la validation : n
- D'un point de vue statistique, on a intérêt à choisir k le plus grand possible, mais d'un point de vue numérique l'estimateur doit être calculé k fois. La VC admet deux dénominations :
  - CV "leave-one-out" avec $k = n$

    $\hookrightarrow$ Petits échantillons et/ou estimateurs rapides à calculer.
  - CV "k-fold" : $k < n$ (souvent $k = 10$)

    $\hookrightarrow$ Grands échantillons et/ou estimateurs lourds en temps de calcul.

Choix du nombre de "paquets" k

- Si on divise les observations en k validation sets :
  - Chaque $\hat{r}^{\ell}$ est calculé à partir de $n(1-1/k)$ observations.
  - Nombre total d'observations utilisées pour la validation : n
- D'un point de vue statistique, on a intérêt à choisir k le plus grand possible, mais d'un point de vue numérique l'estimateur doit être calculé k fois. La VC admet deux dénominations :
  - CV "leave-one-out" avec $k = n$
    $\hookrightarrow$ Petits échantillons et/ou estimateurs rapides à calculer.
  - CV "k-fold" : $k < n$ (souvent $k = 10$)
    $\hookrightarrow$ Grands échantillons et/ou estimateurs lourds en temps de calcul.

Choix du nombre de "paquets" k

- Si on divise les observations en k validation sets :
  - Chaque $\hat{r}^{\ell}$ est calculé à partir de $n(1-1/k)$ observations.
  - Nombre total d'observations utilisées pour la validation : n
- D'un point de vue statistique, on a intérêt à choisir k le plus grand possible, mais d'un point de vue numérique l'estimateur doit être calculé k fois. La VC admet deux dénominations :
  - CV "leave-one-out" avec $k = n$

    $\hookrightarrow$ Petits échantillons et/ou estimateurs rapides à calculer.
  - CV "k-fold" : $k < n$ (souvent $k = 10$)

    $\hookrightarrow$ Grands échantillons et/ou estimateurs lourds en temps de calcul.

Choix du nombre de "paquets" k

- Si on divise les observations en k validation sets :
  - Chaque $\hat{r}^{\ell}$ est calculé à partir de $n(1-1/k)$ observations.
  - Nombre total d'observations utilisées pour la validation : n
- D'un point de vue statistique, on a intérêt à choisir k le plus grand possible, mais d'un point de vue numérique l'estimateur doit être calculé k fois. La VC admet deux dénominations :
  - CV "leave-one-out" avec $k = n$

    $\hookrightarrow$ Petits échantillons et/ou estimateurs rapides à calculer.
  - CV "k-fold" : $k < n$ (souvent $k = 10$)

    $\hookrightarrow$ Grands échantillons et/ou estimateurs lourds en temps de calcul.

Choix du nombre de "paquets" k

- Si on divise les observations en k validation sets :
  - Chaque $\hat{r}^{\ell}$ est calculé à partir de $n(1-1/k)$ observations.
  - Nombre total d'observations utilisées pour la validation : n
- D'un point de vue statistique, on a intérêt à choisir k le plus grand possible, mais d'un point de vue numérique l'estimateur doit être calculé k fois. La VC admet deux dénominations :
  - CV "leave-one-out" avec $k = n$

    $\hookrightarrow$ Petits échantillons et/ou estimateurs rapides à calculer.
  - CV "k-fold" : $k < n$ (souvent $k = 10$)

    $\hookrightarrow$ Grands échantillons et/ou estimateurs lourds en temps de calcul.

Validation croisée : conclusion

- L'erreur de VC est un estimateur non biaisée de l'erreur de prédiction.

  $\hookrightarrow$ Dans le cadre de la régression avec bruit additif, il est égal à cte près au MISE.
- La VC s'applique dans un cadre général pour
  - Evaluer les capacités prédictives d'une procédure estimation
  - Comparer les performances de plusieurs procédures

  quand ces performances ne dépendent pas de la taille d'échantillon
- Théoriquement, la VC simple ne peut pas être utilisée pour choisir la dimension de modèle/la fenêtre. En pratique, la VC "leave-one-out" peut donner de meilleurs résultats que les procédures de sélection basées sur des considérations semi-asymptotiques.
- Pour évaluer les performances d'un estimateur qui dépend d'un paramètre, tous les paramètres doivent être estimés sur l'ensemble d'apprentissage

Validation croisée : conclusion

- L'erreur de VC est un estimateur non biaisée de l'erreur de prédiction.

  $\hookrightarrow$ Dans le cadre de la régression avec bruit additif, il est égal à cte près au MISE.
- La VC s'applique dans un cadre général pour
  - Evaluer les capacités prédictives d'une procédure estimation
  - Comparer les performances de plusieurs procédures

  quand ces performances ne dépendent pas de la taille d'échantillon
- Théoriquement, la VC simple ne peut pas être utilisée pour choisir la dimension de modèle/la fenêtre. En pratique, la VC "leave-one-out" peut donner de meilleurs résultats que les procédures de sélection basées sur des considérations semi-asymptotiques.
- Pour évaluer les performances d'un estimateur qui dépend d'un paramètre, tous les paramètres doivent être estimés sur l'ensemble d'apprentissage

Validation croisée : conclusion

- L'erreur de VC est un estimateur non biaisée de l'erreur de prédiction.

  $\hookrightarrow$ Dans le cadre de la régression avec bruit additif, il est égal à cte près au MISE.
- La VC s'applique dans un cadre général pour
  - Evaluer les capacités prédictives d'une procédure estimation
  - Comparer les performances de plusieurs procédures

  quand ces performances ne dépendent pas de la taille d'échantillon
- Théoriquement, la VC simple ne peut pas être utilisée pour choisir la dimension de modèle/la fenêtre. En pratique, la VC "leave-one-out" peut donner de meilleurs résultats que les procédures de sélection basées sur des considérations semi-asymptotiques.
- Pour évaluer les performances d'un estimateur qui dépend d'un paramètre, tous les paramètres doivent être estimés sur l'ensemble d'apprentissage

Validation croisée : conclusion

- L'erreur de VC est un estimateur non biaisée de l'erreur de prédiction.

  $\hookrightarrow$ Dans le cadre de la régression avec bruit additif, il est égal à cte près au MISE.
- La VC s'applique dans un cadre général pour
  - Evaluer les capacités prédictives d'une procédure estimation
  - Comparer les performances de plusieurs procédures

  quand ces performances ne dépendent pas de la taille d'échantillon
- Théoriquement, la VC simple ne peut pas être utilisée pour choisir la dimension de modèle/la fenêtre. En pratique, la VC "leave-one-out" peut donner de meilleurs résultats que les procédures de sélection basées sur des considérations semi-asymptotiques.
- Pour évaluer les performances d'un estimateur qui dépend d'un paramètre, tous les paramètres doivent être estimés sur l'ensemble d'apprentissage

Conclusion sur la régression NP

- La fonction de régression est estimée aux points du design $\{X_i\}$, et interpolée sur l'intervalle d'estimation.
  - Elle ne peut pas être correctement estimée sur une zone avec très peu d'observations X_i.
  - Le contrôle du MISE $\mathbb{E}[\|\hat{r} - r\|_{f_X}^2]$ "autorise" $(\hat{r} - r)^2(x)$ à être grand si f_X est petit.
- MC : fonction qui explique le mieux les $\{Y_i\}$ en fonction des $\{X_i\}$ dans un espace de dimension fini donné, i.e. qui minimise l'erreur d'approximation :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2$$

- La moyenne des carrés des résidus (définie comme l'erreur d'approximation pour $t = \hat{r}$) diminue systématiquement quand le modèle augmente : ce n'est pas un indicateur de la qualité de l'estimateur.
- Le MISE est estimé à constante près par l'erreur de VC.

Conclusion sur la régression NP

- La fonction de régression est estimée aux points du design $\{X_i\}$, et interpolée sur l'intervalle d'estimation.
  - Elle ne peut pas être correctement estimée sur une zone avec très peu d'observations X_i.
  - Le contrôle du MISE $\mathbb{E}[\|\hat{r} - r\|_{f_X}^2]$ "autorise" $(\hat{r} - r)^2(x)$ à être grand si f_X est petit.
- MC : fonction qui explique le mieux les $\{Y_i\}$ en fonction des $\{X_i\}$ dans un espace de dimension fini donné, i.e. qui minimise l'erreur d'approximation :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2$$

- La moyenne des carrés des résidus (définie comme l'erreur d'approximation pour $t = \hat{r}$) diminue systématiquement quand le modèle augmente : ce n'est pas un indicateur de la qualité de l'estimateur.
- Le MISE est estimé à constante près par l'erreur de VC.

Conclusion sur la régression NP

- La fonction de régression est estimée aux points du design $\{X_i\}$, et interpolée sur l'intervalle d'estimation.
  - Elle ne peut pas être correctement estimée sur une zone avec très peu d'observations X_i.
  - Le contrôle du MISE $\mathbb{E}[\|\hat{r} - r\|_{f_X}^2]$ "autorise" $(\hat{r} - r)^2(x)$ à être grand si f_X est petit.
- MC : fonction qui explique le mieux les $\{Y_i\}$ en fonction des $\{X_i\}$ dans un espace de dimension fini donné, i.e. qui minimise l'erreur d'approximation :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2$$

- La moyenne des carrés des résidus (définie comme l'erreur d'approximation pour $t = \hat{r}$) diminue systématiquement quand le modèle augmente : ce n'est pas un indicateur de la qualité de l'estimateur.
- Le MISE est estimé à constante près par l'erreur de VC.

Conclusion sur la régression NP

- La fonction de régression est estimée aux points du design $\{X_i\}$, et interpolée sur l'intervalle d'estimation.
  - Elle ne peut pas être correctement estimée sur une zone avec très peu d'observations X_i.
  - Le contrôle du MISE $\mathbb{E}[\|\hat{r} - r\|_{f_X}^2]$ "autorise" $(\hat{r} - r)^2(x)$ à être grand si f_X est petit.
- MC : fonction qui explique le mieux les $\{Y_i\}$ en fonction des $\{X_i\}$ dans un espace de dimension fini donné, i.e. qui minimise l'erreur d'approximation :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2$$

- La moyenne des carrés des résidus (définie comme l'erreur d'approximation pour $t = \hat{r}$) diminue systématiquement quand le modèle augmente : ce n'est pas un indicateur de la qualité de l'estimateur.
- Le MISE est estimé à constante près par l'erreur de VC.

Conclusion sur la régression NP

- La fonction de régression est estimée aux points du design $\{X_i\}$, et interpolée sur l'intervalle d'estimation.
  - Elle ne peut pas être correctement estimée sur une zone avec très peu d'observations X_i.
  - Le contrôle du MISE $\mathbb{E}[\|\hat{r} - r\|_{f_X}^2]$ "autorise" $(\hat{r} - r)^2(x)$ à être grand si f_X est petit.
- MC : fonction qui explique le mieux les $\{Y_i\}$ en fonction des $\{X_i\}$ dans un espace de dimension fini donné, i.e. qui minimise l'erreur d'approximation :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2$$

- La moyenne des carrés des résidus (définie comme l'erreur d'approximation pour $t = \hat{r}$) diminue systématiquement quand le modèle augmente : ce n'est pas un indicateur de la qualité de l'estimateur.
- Le MISE est estimé à constante près par l'erreur de VC.

Conclusion sur l'estimation NP

Conclusion sur l'estimation NP

- Quand utiliser des estimateurs NP ?
  - Quand on n'a pas de modèles paramétriques envisageables a priori.
  - Quand les modèles paramétriques envisageables ne collent pas aux données
  - Quand on a suffisamment d'observations
- Différences entre estimateurs paramétriques et NP
  - En paramétrique : théoriquement, on suppose que le modèle est vrai.
  - En NP : on considère un modèle d'approximation (on sait qu'il n'est pas exact) dépendant d'un paramètre de régularisation (dimension du modèle, fenêtre,etc), et on choisit ce paramètre en fonction des données.
    $\hookrightarrow$ Le modèle dépend notamment de la taille d'échantillon.
- Choix du paramètre de régularisation :
  - Méthodes théoriquement rigoureuses, mais basées sur des résultats semi-asymptotiques ou asymptotiques.
  - Méthodes ad-hoc, moins rigoureuses mais parfois plus efficaces.

Conclusion sur l'estimation NP

- Quand utiliser des estimateurs NP ?
  - Quand on n'a pas de modèles paramétriques envisageables a priori.
  - Quand les modèles paramétriques envisageables ne collent pas aux données
  - Quand on a suffisamment d'observations
- Différences entre estimateurs paramétriques et NP
  - En paramétrique : théoriquement, on suppose que le modèle est vrai.
  - En NP : on considère un modèle d'approximation (on sait qu'il n'est pas exact) dépendant d'un paramètre de régularisation (dimension du modèle, fenêtre,etc), et on choisit ce paramètre en fonction des données.
    $\hookrightarrow$ Le modèle dépend notamment de la taille d'échantillon.
- Choix du paramètre de régularisation :
  - Méthodes théoriquement rigoureuses, mais basées sur des résultats semi-asymptotiques ou asymptotiques.
  - Méthodes ad-hoc, moins rigoureuses mais parfois plus efficaces.

Conclusion sur l'estimation NP

- Quand utiliser des estimateurs NP ?
  - Quand on n'a pas de modèles paramétriques envisageables a priori.
  - Quand les modèles paramétriques envisageables ne collent pas aux données
  - Quand on a suffisamment d'observations
- Différences entre estimateurs paramétriques et NP
  - En paramétrique : théoriquement, on suppose que le modèle est vrai.
  - En NP : on considère un modèle d'approximation (on sait qu'il n'est pas exact) dépendant d'un paramètre de régularisation (dimension du modèle, fenêtre,etc), et on choisit ce paramètre en fonction des données.
    $\hookrightarrow$ Le modèle dépend notamment de la taille d'échantillon.
- Choix du paramètre de régularisation :
  - Méthodes théoriquement rigoureuses, mais basées sur des résultats semi-asymptotiques ou asymptotiques.
  - Méthodes ad-hoc, moins rigoureuses mais parfois plus efficaces.

Quelques mots sur le cas multivarié

Ex : Estimation de densité par histogramme. Soit $(X_i)_{i=1,\dots,n}$ i.i.d

Uni-varié : $X_i \in \mathbb{R}$

- $n = 50$ observations
- Hist à $D = 5$ morceaux.
- En moyenne : $n/D = 10$ observations par intervalle

Bi-varié : $X_i = (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$.

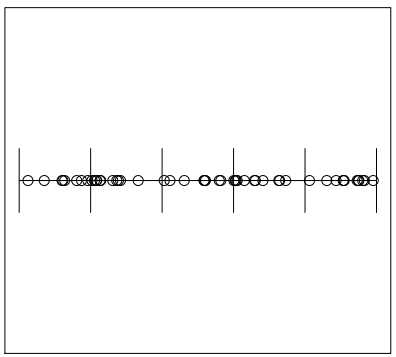
- $n = 50$ observations
- $D = 5$ morceaux dans chaque dimension
- En moyenne : $n/D^2 = 2$ obs par case
- Pas d'observations dans certaines cases

Quelques mots sur le cas multivarié

Ex : Estimation de densité par histogramme. Soit $(X_i)_{i=1,\dots,n}$ i.i.d

Uni-varié : $X_i \in \mathbb{R}$

- $n = 50$ observations
- Hist à $D = 5$ morceaux.



- En moyenne : $n/D = 10$ observations par intervalle

Bi-varié : $X_i = (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$.

- $n = 50$ observations
- $D = 5$ morceaux dans chaque dimension

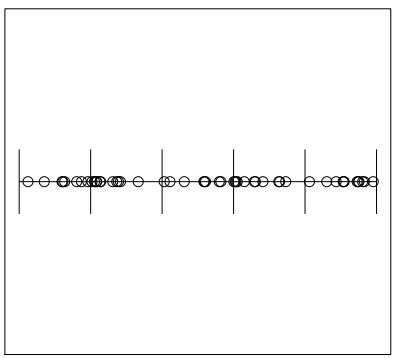
- En moyenne : $n/D^2 = 2$ obs par case
- Pas d'observations dans certaines cases

Quelques mots sur le cas multivarié

Ex : Estimation de densité par histogramme. Soit $(X_i)_{i=1,\dots,n}$ i.i.d

Uni-varié : $X_i \in \mathbb{R}$

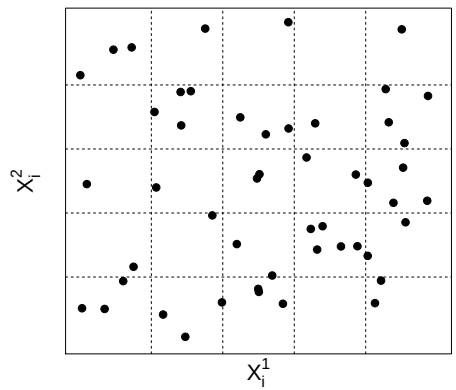
- $n = 50$ observations
- Hist à $D = 5$ morceaux.



- En moyenne : $n/D = 10$ observations par intervalle

Bi-varié : $X_i = (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$.

- $n = 50$ observations
- $D = 5$ morceaux dans chaque dimension



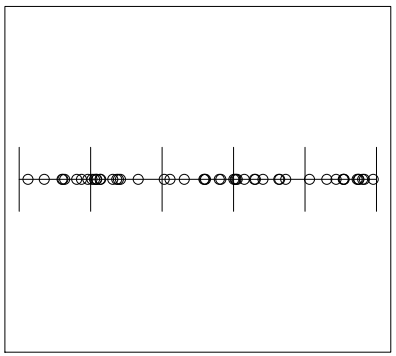
- En moyenne : $n/D^2 = 2$ obs par case
- Pas d'observations dans certaines cases

Quelques mots sur le cas multivarié

Ex : Estimation de densité par histogramme. Soit $(X_i)_{i=1,\dots,n}$ i.i.d

Uni-varié : $X_i \in \mathbb{R}$

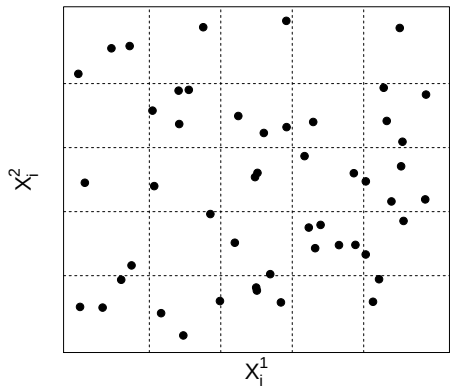
- $n = 50$ observations
- Hist à $D = 5$ morceaux.



- En moyenne : $n/D = 10$ observations par intervalle

Bi-varié : $X_i = (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$.

- $n = 50$ observations
- $D = 5$ morceaux dans chaque dimension



- En moyenne : $n/D^2 = 2$ obs par case
- Pas d'observations dans certaines cases

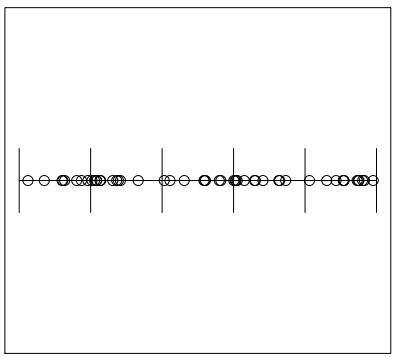
D doit être très petit pour ne pas que la variance explose.

Quelques mots sur le cas multivarié

Ex : Estimation de densité par histogramme. Soit $(X_i)_{i=1,\dots,n}$ i.i.d

Uni-varié : $X_i \in \mathbb{R}$

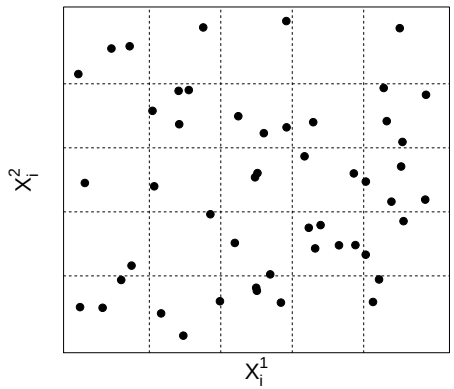
- $n = 50$ observations
- Hist à $D = 5$ morceaux.



- En moyenne : $n/D = 10$ observations par intervalle

Bi-varié : $X_i = (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$.

- $n = 50$ observations
- $D = 5$ morceaux dans chaque dimension



- En moyenne : $n/D^2 = 2$ obs par case
- Pas d'observations dans certaines cases

D doit être très petit pour ne pas que la variance explose.

Point crucial : rapport entre le nombre d'observations n et le nombre de paramètres

Soit $(X_i, Y_i)_{i=1,\ldots,n}$ avec $X_i \in [0,1]^k$ et $Y_i \in \mathbb{R}$.

- Estimateur des MC : D^k paramètres à estimer. On est restreint à des modèles D tq $D^k << n$ pour limiter la variance. Il faut donc
  - Un échantillon n très grand
  - Un D très petit c'est à dire un modèle peu flexible
- Même problème avec les noyaux : on doit choisir une grande fenêtre $\rightarrow$ la fonction est lissée.
- Plus généralement, si n n'est pas très grand, il y a des "zones sans observations" sur $[0,1]^k$. Sans hypothèse supplémentaires, l'estimation NP est très peu performante.

Soit $(X_i, Y_i)_{i=1,\ldots,n}$ avec $X_i \in [0,1]^k$ et $Y_i \in \mathbb{R}$.

- Estimateur des MC : D^k paramètres à estimer. On est restreint à des modèles D tq $D^k << n$ pour limiter la variance. Il faut donc
  - Un échantillon n très grand
  - Un D très petit c'est à dire un modèle peu flexible
- Même problème avec les noyaux : on doit choisir une grande fenêtre $\rightarrow$ la fonction est lissée.
- Plus généralement, si n n'est pas très grand, il y a des "zones sans observations" sur $[0,1]^k$. Sans hypothèse supplémentaires, l'estimation NP est très peu performante.

Soit $(X_i, Y_i)_{i=1,\ldots,n}$ avec $X_i \in [0,1]^k$ et $Y_i \in \mathbb{R}$.

- Estimateur des MC : D^k paramètres à estimer. On est restreint à des modèles D tq $D^k << n$ pour limiter la variance. Il faut donc
  - Un échantillon n très grand
  - Un D très petit c'est à dire un modèle peu flexible
- Même problème avec les noyaux : on doit choisir une grande fenêtre $\rightarrow$ la fonction est lissée.
- Plus généralement, si n n'est pas très grand, il y a des "zones sans observations" sur $[0,1]^k$. Sans hypothèse supplémentaires, l'estimation NP est très peu performante.

Que faire en régression multivariée ?

On doit poser plus d'hypothèses sur le modèle.

- Exple de modèle semi-paramétrique : modèle linéaire généralisé. On suppose que l'effet des variables se résume en une combinaison linéaire :

$$f(x) = g(\sum_{j=1}^{k} \alpha_j x_j) \quad D + k \text{ paramètres}$$

  avec g estimée non-paramétriquement.

- Modèle additif : on suppose que les effets des variables $\{x_j, j = 1, \ldots, k\}$ sont additifs :

$$f(x) = \sum_{j=1}^{k} g_j(x_j) \quad D \times k \text{ paramètres}$$

  avec g_j estimée non-paramétriquement.

- Si le nombre de variables k est proche de n, on tombe dans le contexte de la grande dimension.

Que faire en régression multivariée ?

On doit poser plus d'hypothèses sur le modèle.

- Exple de modèle semi-paramétrique : modèle linéaire généralisé. On suppose que l'effet des variables se résume en une combinaison linéaire :

$$f(x) = g(\sum_{j=1}^{k} \alpha_j x_j) \quad D + k \text{ paramètres}$$

avec g estimée non-paramétriquement.

- Modèle additif : on suppose que les effets des variables $\{x_j, j = 1, \ldots, k\}$ sont additifs :

$$f(x) = \sum_{j=1}^{k} g_j(x_j) \quad D \times k \text{ paramètres}$$

avec g_j estimée non-paramétriquement.

- Si le nombre de variables k est proche de n, on tombe dans le contexte de la grande dimension.

Que faire en régression multivariée ?

On doit poser plus d'hypothèses sur le modèle.

- Exple de modèle semi-paramétrique : modèle linéaire généralisé. On suppose que l'effet des variables se résume en une combinaison linéaire :

$$f(x) = g(\sum_{j=1}^{k} \alpha_j x_j) \quad D + k \text{ paramètres}$$

avec g estimée non-paramétriquement.

- Modèle additif : on suppose que les effets des variables $\{x_j, j = 1, \ldots, k\}$ sont additifs :

$$f(x) = \sum_{j=1}^{k} g_j(x_j) \quad D \times k \text{ paramètres}$$

avec g_j estimée non-paramétriquement.

- Si le nombre de variables k est proche de n, on tombe dans le contexte de la grande dimension.

Que faire en régression multivariée ?

On doit poser plus d'hypothèses sur le modèle.

- Exple de modèle semi-paramétrique : modèle linéaire généralisé. On suppose que l'effet des variables se résume en une combinaison linéaire :

$$f(x) = g(\sum_{j=1}^{k} \alpha_j x_j) \quad D + k \text{ paramètres}$$

avec g estimée non-paramétriquement.

- Modèle additif : on suppose que les effets des variables $\{x_j, j = 1, \ldots, k\}$ sont additifs :

$$f(x) = \sum_{j=1}^{k} g_j(x_j) \quad D \times k \text{ paramètres}$$

avec g_j estimée non-paramétriquement.

- Si le nombre de variables k est proche de n, on tombe dans le contexte de la grande dimension.

La grande dimension : des problématiques similaires à l'estimation NP

- Définition : On est dans un contexte de grande dimension quand on travaille sur un échantillon $(Z_i)_{i=1,\dots,n}$ avec $k = \dim(Z_i) \geq n$ (ou au moins de l'ordre de n).
- Dans ce contexte, les estimateurs NP ou le modèle linéaire ne sont pas identifiables.
- *Ex : régression linéaire :* $Y_i = \sum_{j=1}^{k} \theta_j X_{i,j} + \varepsilon_i, \ \ i = 1, \dots, n$

  L'estimateur des MC satisfait :

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^k} \sum_{i=1} (Y_i - \sum_{j=1}^{k} \theta_j X_{i,j})^2 \quad \Leftrightarrow \quad \mathbb{X}^t \mathbb{X} \hat{\theta} = \mathbb{X}^t \mathbf{Y}$$

  Or

$$\left. \begin{array}{l} rang(\mathbb{X}^t\mathbb{X}) = rang(t\mathbb{X}) \leq k \\ dim(\mathbb{X}^t\mathbb{X}) = n * n \end{array} \right\} \Rightarrow \mathbb{X}^t\mathbb{X} \text{ non inversible}$$

- *L'estimateur des MC n'est pas unique, de plus il "colle" exactement aux données :* $\sum_{i=1} (Y_i - \sum_{j=1}^{k} \hat{\theta}_j X_{i,j})^2 = 0$

La grande dimension : des problématiques similaires à l'estimation NP

- Définition : On est dans un contexte de grande dimension quand on travaille sur un échantillon $(Z_i)_{i=1,\dots,n}$ avec $k = \dim(Z_i) \geq n$ (ou au moins de l'ordre de n).
- Dans ce contexte, les estimateurs NP ou le modèle linéaire ne sont pas identifiables.
- *Ex : régression linéaire :* $Y_i = \sum_{j=1}^{k} \theta_j X_{i,j} + \varepsilon_i, \ \ i = 1, \dots, n$

  L'estimateur des MC satisfait :

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^k} \sum_{i=1} (Y_i - \sum_{j=1}^{k} \theta_j X_{i,j})^2 \quad \Leftrightarrow \quad \mathbb{X}^t \mathbb{X} \hat{\theta} = \mathbb{X}^t \mathbf{Y}$$

  Or

$$\left. \begin{array}{l} rang(\mathbb{X}^t\mathbb{X}) = rang(t\mathbb{X}) \leq k \\ dim(\mathbb{X}^t\mathbb{X}) = n * n \end{array} \right\} \Rightarrow \mathbb{X}^t\mathbb{X} \text{ non inversible}$$

- *L'estimateur des MC n'est pas unique, de plus il "colle" exactement aux données :* $\sum_{i=1} (Y_i - \sum_{j=1}^{k} \hat{\theta}_j X_{i,j})^2 = 0$

La grande dimension : des problématiques similaires à l'estimation NP

- Définition : On est dans un contexte de grande dimension quand on travaille sur un échantillon $(Z_i)_{i=1,\dots,n}$ avec $k = \dim(Z_i) \geq n$ (ou au moins de l'ordre de n).
- Dans ce contexte, les estimateurs NP ou le modèle linéaire ne sont pas identifiables.
- *Ex : régression linéaire :* $Y_i = \sum_{j=1}^{k} \theta_j X_{i,j} + \varepsilon_i, \ \ i = 1, \dots, n$

  L'estimateur des MC satisfait :

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^k} \sum_{i=1} (Y_i - \sum_{j=1}^{k} \theta_j X_{i,j})^2 \quad \Leftrightarrow \quad \mathbb{X}^t \mathbb{X} \hat{\theta} = \mathbb{X}^t \mathbf{Y}$$

  Or

$$\left. \begin{array}{l} rang(\mathbb{X}^t\mathbb{X}) = rang(t\mathbb{X}) \leq k \\ dim(\mathbb{X}^t\mathbb{X}) = n * n \end{array} \right\} \Rightarrow \mathbb{X}^t\mathbb{X} \text{ non inversible}$$

- *L'estimateur des MC n'est pas unique, de plus il "colle" exactement aux données :* $\sum_{i=1} (Y_i - \sum_{j=1}^{k} \hat{\theta}_j X_{i,j})^2 = 0$

La grande dimension : des problématiques similaires à l'estimation NP

- Définition : On est dans un contexte de grande dimension quand on travaille sur un échantillon $(Z_i)_{i=1,\dots,n}$ avec $k = \dim(Z_i) \geq n$ (ou au moins de l'ordre de n).
- Dans ce contexte, les estimateurs NP ou le modèle linéaire ne sont pas identifiables.
- *Ex : régression linéaire :* $Y_i = \sum_{j=1}^{k} \theta_j X_{i,j} + \varepsilon_i, \ \ i = 1, \dots, n$

  L'estimateur des MC satisfait :

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^k} \sum_{i=1} (Y_i - \sum_{j=1}^{k} \theta_j X_{i,j})^2 \quad \Leftrightarrow \quad \mathbb{X}^t \mathbb{X} \hat{\theta} = \mathbb{X}^t \mathbf{Y}$$

  Or

$$\left. \begin{array}{l} \textit{rang}(\mathbb{X}^t\mathbb{X}) = \textit{rang}(t\mathbb{X}) \leq k \\ \textit{dim}(\mathbb{X}^t\mathbb{X}) = n * n \end{array} \right\} \Rightarrow \mathbb{X}^t\mathbb{X} \textit{ non inversible}$$

- *L'estimateur des MC n'est pas unique, de plus il "colle" exactement aux données :* $\sum_{i=1} (Y_i - \sum_{j=1}^{k} \hat{\theta}_j X_{i,j})^2 = 0$

Régression pénalisée

De manière similaire à la sélection de modèles en régression NP univariée, le problème d'unicité et d'estimateur qui "colle trop" aux données

Rappel : sélection de modèles en régression NP

$$\hat{D} = \arg\min_{D} \left(\min_{\theta \in \mathbb{R}^D} \frac{1}{n} \|\mathbf{Y} - \theta^D \mathbb{X}\|^2 + C \frac{D}{n} \right)$$

et estimateur $\hat{\theta}^{\hat{D}}$ avec $\hat{\theta}^D = \arg\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \theta^D \mathbb{X}\|$.

- $\min_{\theta \in \mathbb{R}^D} \|\mathbf{Y} - \theta^D \mathbb{X}\|^2 \searrow$ quand $D \nearrow$
- La pénalité CD/n qui $\nearrow$ quand $D \nearrow$ est ajoutée pour compenser

Régression linéaire en grande dimension

Pour $\|\theta\|_0$ une norme de $\mathbb{R}^k$ donnée

$$\min_{\{\theta, \|\theta\|_0 \leq R\}} \|\mathbf{Y} - \theta\mathbb{X}\|^2 \searrow \text{ quand R } \nearrow$$

De manière analogue, on compense ce phénomène en ajoutant une pénalité :

$$\arg\min_{R>0} \left(\min_{\{\theta, \|\theta\|_0 \leq R\}} \|\mathbf{Y} - \theta\mathbb{X}\| + \lambda R \right)$$

que l'on reformule en :

$$\hat{\theta} = \arg\min_{\theta} \left(\|\mathbf{Y} - \theta\mathbb{X}\|^2 + \lambda \|\theta\|_0 \right)$$

Grande dimension : conclusion

- Il existe des méthodes pour contourner les limitations de la stat classique (nb de paramètres $<$ taille d'échantillon)
- En grande dimension, comme en stats NP, un bon estimateur réalise un compromis entre
  - Expliquer suffisamment l'échantillon d'observations (faible biais)
  - Avoir des paramètres suffisamment stables par rapport à l'échantillon d'estimation particulier (faible variance)
- Lorsque des connaissances a priori **fiables** sont disponibles :
  - modèle paramétrique plutôt qu'estimateur NP,
  - petit jeu de variables pertinentes plutôt que grande dimension,

  l'estimation sera toujours meilleure en utilisant ces informations !

Grande dimension : conclusion

- Il existe des méthodes pour contourner les limitations de la stat classique (nb de paramètres $<$ taille d'échantillon)
- En grande dimension, comme en stats NP, un bon estimateur réalise un compromis entre
  - ◇ Expliquer suffisamment l'échantillon d'observations (faible biais)
  - ◇ Avoir des paramètres suffisamment stables par rapport à l'échantillon d'estimation particulier (faible variance)
- Lorsque des connaissances a priori **fiables** sont disponibles :
  - ◇ modèle paramétrique plutôt qu'estimateur NP,
  - ◇ petit jeu de variables pertinentes plutôt que grande dimension,

  l'estimation sera toujours meilleure en utilisant ces informations !

Grande dimension : conclusion

- Il existe des méthodes pour contourner les limitations de la stat classique (nb de paramètres $<$ taille d'échantillon)
- En grande dimension, comme en stats NP, un bon estimateur réalise un compromis entre
  - ◇ Expliquer suffisamment l'échantillon d'observations (faible biais)
  - ◇ Avoir des paramètres suffisamment stables par rapport à l'échantillon d'estimation particulier (faible variance)
- Lorsque des connaissances a priori **fiables** sont disponibles :
  - ◇ modèle paramétrique plutôt qu'estimateur NP,
  - ◇ petit jeu de variables pertinentes plutôt que grande dimension,

  l'estimation sera toujours meilleure en utilisant ces informations !