

Nonparametric regression

Hamel Elhadj

Département de mathématiques
Université Hassiba Benbouali-Chlef
2021/2022

Support de cours destiné aux étudiants Master2 mathématiques
Option :Mathématique Appliquées et statistique

Plan de la présentation

1 Modèle de régression

- Régressogramme
- Estimateur de Nadaraya-Watson (noyaux pour la régression)
- Local linear regression
- Estimateur par polynômes locaux
- Choix du paramètre de lissage
- Confidence Regions
- Nearest-Neighbor Estimator
- Estimate by Orthogonal Series

2 Multivariate Kernel Regression

Modèle de régression

Motivation :

Dans ce chapitre, on cherche à expliquer les valeurs que peut prendre une variable Y à partir des valeurs que peut prendre une variable X .

Exemples :

- Y est le taux d'insuline dans le sang, qu'on explique (ou prédit) à l'aide de $X =$ (IMC, pression du sang, concentration de molécules).
- Y est le niveau de diplôme obtenu, qu'on explique à l'aide de $X =$ (âge, sexe, revenu des parents, métier des parents).

On suppose que la variable Y est intégrable $\mathbb{E}|Y| < \infty$ et on note r la fonction de régression de Y sur X : $r(x) = \mathbb{E}(Y|X = x)$

L'objectif est d'estimer la fonction r pour expliquer et prédire Y à partir de X .

Modèle de régression

L'objectif est d'estimer la fonction r pour expliquer et prédire Y à partir de X .

Pour cela on dispose des réalisations de n couples de variables

$(X_1, Y_1), \dots, (X_n, Y_n)$. On va supposer que les (X_i, Y_i) sont indépendants.

- les Y_i sont les variables à expliquer ou les variables réponses ou variables de sortie.

↳ les X_i constituent le design, les variables explicatives, les covariables, ou variables d'entrée.

Les variables explicatives pourront être aléatoire ou déterministe. Dans ce dernier cas, on notera plutôt x_i à la place de X_i .

$$Y_i = r(X_i) + \varepsilon_i, i = 1, \dots, n,$$

Modèle de régression

On dispose de n observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ du couple (X, Y) . On suppose que

$$y_i = r(x_i) + \varepsilon_i \quad \text{pour tout } i = 1, \dots, n$$

- les x_i sont des valeurs connues non aléatoires
- r est une fonction inconnue
- ε_i sont des réalisations inconnues d'une variable aléatoire.

Pour chaque individu i , la variable aléatoire ε_i représente l'erreur commise. Généralement pour étudier le modèle "le statisticien" formule des hypothèses sur la loi des erreurs ε_i .

Basic Concepts in Regression

The Regression Problem:

Observe a **training sample** $(X_1, Y_1), \dots, (X_n, Y_n)$, use this to **estimate**

$$r(x) = \mathbb{E}(Y|X = x).$$

Equivalently: Estimate $f(x)$ when

$$Y_i = r(X_i) + \epsilon_i, \quad i = 1, 2, \dots, n$$

where $\mathbb{E}(\epsilon_i) = 0$.

Simple estimators when X is real-valued:

$$\begin{aligned}\hat{r}(x) &= \hat{\beta}_0 + \hat{\beta}_1 x && \text{parametric} \\ \hat{r}(x) &= \text{mean}\{Y_i : |X_i - x| \leq h\} && \text{nonparametric}\end{aligned}$$

la régression linéaire simple

On observe des observations bruitées $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $i = 1, \dots, n$ avec β_0 et β_1 inconnus.

- Le premier terme correspond à l'équation d'une droite.
- Le deuxième terme correspond à l'erreur et varie de façon aléatoire d'un individu à l'autre.

la régression linéaire multiple

Supposons qu'on dispose de p -variables explicatives X_1, X_2, \dots, X_p . Soit X la matrice augmentée (n lignes et $p + 1$ colonnes). Soit $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ le vecteur de coefficients inconnus.

- Modèle Théorique (sous forme vectorielle)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

- Modèle Théorique (sous forme matricielle) $Y = \mathbb{X}\beta + \varepsilon$
- Coefficients estimés (par la méthode de MC) : $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$

Non Parametric Regression: Introduction

- Parametric approach: $m(\cdot)$ is known and smooth. It is fully described by a finite set of parameters, to be estimated. Easy interpretation. For example, a linear model:

$$y_i = x_i' \beta + \varepsilon_i, \quad i = 1, \dots, N$$

- Nonparametric approach: $m(\cdot)$ is smooth, flexible, but unknown. Let the data determine the shape of $m(\cdot)$. Difficult interpretation.

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \dots, N$$

- Semi-parametric approach: $m(\cdot)$ have some parameters -to be estimated-, but some parts are determined by the data.

$$y_i = x_i' \beta + m_z(z_i) + \varepsilon_i, \quad i = 1, \dots, N$$

Contexte de la régression

- **Contexte général de la régression** : Soit $(X_1, Y_1), \dots, (X_n, Y_n)$ 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 $\{\varepsilon_i\}_{i=1,\dots,n}$ 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

$$Y_i = r(X_i) + \varepsilon_i \quad \text{avec} \quad \varepsilon_i \perp X_i, \quad \mathbb{E}[\varepsilon_i] = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2 <$$

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 :

$$\hat{Y}_{i'} = \hat{r}(X_{i'})$$

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 polynômes locaux.

Modèle de régression

Le modèle de la régression est l'un des modèle les plus fréquemment rencontrés en statistique paramétrique et non paramétrique.

Soient $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ des couples de variables aléatoires indépendantes et de même loi que (X, Y) .

Dans le modèle de régression non paramétrique on suppose l'existence d'une fonction "r" qui exprime la valeur moyenne de la variable à expliquer Y en fonction de la variable explicative X , c'est à dire :

$$Y = r(X) + \varepsilon$$

où ε est une variable centrée et indépendante de X .

Modèle de régression

On suppose que (X, Y) est à valeurs dans \mathbb{R}^2 , il admet une densité jointe $f(x, y)$ sur \mathbb{R}^2 et une densité marginale $f(x) > 0$ (par rapport à la mesure de Lebesgue sur \mathbb{R}).

La variable Y est supposée intégrable, c'est à dire $\mathbb{E}(|Y|) < \infty$, on peut alors définir la fonction régression $r(x)$ par :

$$r(x) = \mathbb{E}(Y/X = x) = \int yf(Y/X = x)dy = \frac{\int yf(x, y)dy}{\int f(x)dy} = r(x)$$

$r(x)$ est la fonction qui réalise la meilleure approximation de Y sachant $X = x$ au sens des moindres carrés. Dans ce problème, l'estimation de $r(x)$ est de type non paramétrique.

Modèle de régression- Exemple

Consider the joint pdf

$$f(x, y) = x + y \quad \text{for} \quad 0 \leq x \leq 1 \quad \text{and} \quad 0 \leq y \leq 1$$

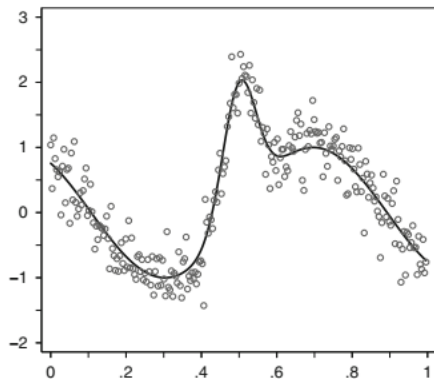
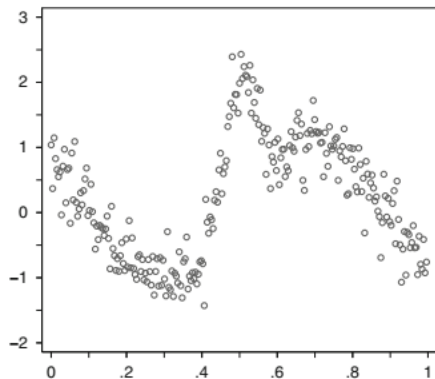
with $f(x, y) = 0$ elsewhere, and marginal pdf

$$f_X(x) = \int_0^1 f(x, y) dy = x + \frac{1}{2} \quad \text{for} \quad 0 \leq x \leq 1$$

with $f_X(x) = 0$ elsewhere. Then

$$\begin{aligned} \mathbb{E}(Y|X = x) &= \int y \frac{f(x, y)}{f_X(x)} dy \\ &= \int_0^1 y \frac{x + y}{x + \frac{1}{2}} dy \\ &= \frac{\frac{1}{2}x + \frac{1}{3}}{x + \frac{1}{2}} = m(x) \end{aligned}$$

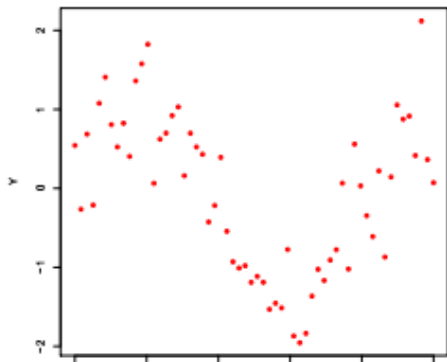
Example



Scatter plot of a simulated data set with nonlinear effect of the covariate: The *right panel* additionally shows the true covariate effect. The data have been simulated according to the model $y = f(x) + \varepsilon$ where $f(x) = \sin(2(4x - 2)) + 2 \exp(-(16^2)(x - 0.5)^2)$ and $\varepsilon \sim N(0, 0.3^2)$

An Example

- The data are $n = 60$ pairs of observations from a certain regression model.
- How to construct \hat{r}_n , an estimator of r ?



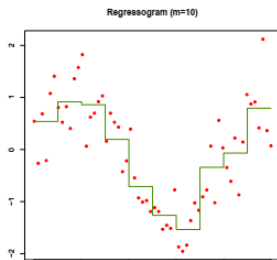
Estimator : Regressogram

On peut estimer la fonction "r" par une fonction constante par morceaux sur une partition de $[0, 1]$. (Ces estimateurs sont les analogues en régression des estimateurs par histogramme en densité, on les appelle **régressogrammes**).

On découpe $[0, 1]$ en m intervalles de même taille B_1, B_2, \dots, B_m :

$$\hat{g}_n(x) = \frac{1}{n_j} \sum_{i, x_i \in B_j} Y_i \quad \text{for } x \in B_j$$

where n_j is the number of points in B_j . Here we use the convention $\frac{0}{0} = 0$.



Principe

- ▶ Il s'agit de l'équivalent de l'histogramme pour le problème de régression. On suppose que la fonction de régression r est définie sur un **intervalle borné** $[a, b] \subset \mathbb{R}$ et $r \in \mathbb{L}_2([a, b])$.

Définition

- ▶ Soit $I = (I_k)_{1 \leq k \leq D}$ une **partition** de $[a, b]$ (i.e. intervalles disjoints dont l'union est $[a, b]$),
- ▶ On note $n_k = \text{Card}\{i; X_i \in I_k\}$ le nombre de variables X dans I_k .
- ▶ L'**estimateur par regressogramme** de r est défini par

$$\hat{r}_{I,n}(x) = \sum_{k=1}^D \left[\sum_{i=1}^n \frac{Y_i}{n_k} 1_{X_i \in I_k} \right] 1_{I_k}(x).$$

- ▶ Il affecte à chaque intervalle I_k une valeur égale à la **moyenne des observations Y** dans cet intervalle, **renormalisée** par le nombre de variables X de cet intervalle.

Estimator : Regressogram

Illustration

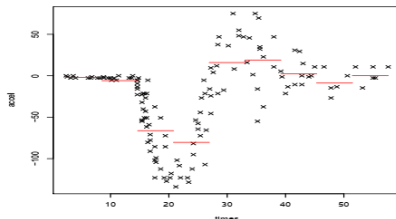


Figure : Régressogramme sur les données *mcycle* de la librairie *MASS*.

Remarques :

Comme pour les histogrammes,

- Fixer les intervalles à l'avance n'est pas la meilleure chose à faire ;
- On peut voir la moyenne sur un intervalle comme le résultat d'un lissage par un noyau rectangulaire, et donc préférer des noyaux plus réguliers.

On introduit donc les noyaux pour la régression.

Estimateur de Nadaraya-Watson

Supposons que (X, Y) a une densité $f : (x, y) \longrightarrow f(x, y)$ sur \mathbb{R}^2 et que $f_X : x \longrightarrow f_X(x) = \int f(x, y)dy > 0$ (densité de X). Alors,

$$\forall x \in \mathbb{R}; r(x) = E(Y/X = x) = \frac{\int yf(x, y)dy}{f_X(x)}$$

Les densités $f(x, y)$ et $f_X(x)$ sont inconnues mais on peut les estimer via

$$\forall (x, y) \in \mathbb{R}^2; \hat{f}_n(x, y) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right)$$

$$\forall (x) \in \mathbb{R}; \hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

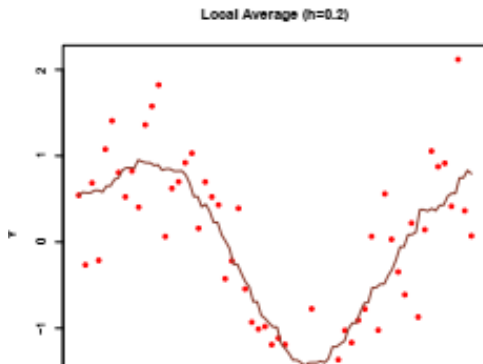
puis on considère l'estimateur de la régression

$$\forall (x, y) \in \mathbb{R}^2; \hat{r}_n(x) = \frac{\int y\hat{f}(x, y)dy}{\hat{f}_X(x)}.$$

Nadaraya-Watson Estimator- *naive kernel estimator*

$$\text{Fixe } h > 0, \hat{r}_n(x) = \frac{\sum_{i=1}^n I_{(x-h < x_i \leq x+h)} Y_i}{\sum_{i=1}^n I_{(x-h < x_i \leq x+h)}} = \frac{\sum_{i=1}^n \frac{1}{2} I_{(-1 < x_i \leq 1)} \left(\frac{x - X_i}{h} \right) Y_i}{\sum_{i=1}^n \frac{1}{2} I_{(-1 < x_i \leq 1)} \left(\frac{x - X_i}{h} \right)}$$

This Estimator called naive kernel estimator.

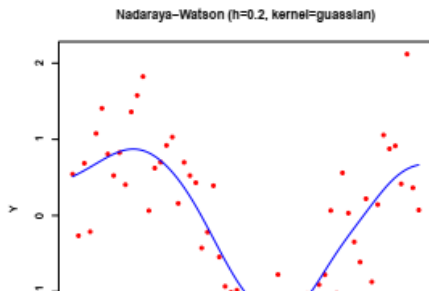


Nadaraya-Watson Estimator

Nadaraya-Watson Estimator

Replacing the box kernel by a general kernel in the local average estimator, we obtain the Nadaraya-Watson estimator of r :

$$\hat{r}_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}.$$



L'estimateur de Nadaraya-Watson

Nous nous présentons maintenant le célèbre estimateur à noyau de la fonction de régression introduit par Nadaraya et Watson 1964.

Il est construit à partir d'une fonction noyau $K(\cdot)$ et d'une fenêtre (h_n) , de manière analogue à l'estimateur à noyau de la fonction de densité introduit par Rosenblatt [?] et Parzen [?].

Soit

$$m(x) = \mathbb{E}[Y|X = x] = \int_{\mathbb{R}^d} yf(y/x)dy.$$

Si on remplace $f(x, y)$ par $\hat{f}_n(x, y)$ et $f(x)$ par $\hat{f}_n(x)$, on obtient l'estimateur de NW.

Définition

L'estimateur de Nadaraya-Watson défini par :

$$\hat{r}_n^{NW}(x) = \frac{\int_{\mathbb{R}^d} y\hat{f}_n(x, y)dy}{\hat{f}_n(x)},$$

L'estimateur de Nadaraya-Watson

Proposition

Si K est un noyau d'ordre 1 satisfait

$$\int_{\mathbb{R}^d} uK(u)du = 0 \text{ et } \int_{\mathbb{R}^d} K(u)du = 1 ,$$

alors l'estimateur de Nadaraya et Watson est défini par

$$\hat{m}_n^{NW}(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)} & \text{si } \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \neq 0 \\ \frac{1}{n} \sum_{i=1}^n Y_i & \text{sinon} \end{cases} \quad (1)$$

où h_n est une paramètre de lissage, satisfait

Remarque *On peut écrire*

$$\hat{r}_n(x) = \sum_{i=1}^n \omega_{n,i}(x) Y_i$$

$$\text{où } \omega_{n,i}(x) = \begin{cases} \frac{K(\frac{X_i-x}{h})}{\sum_{i=1}^n K(\frac{X_i-x}{h})} & \text{si } \sum_{i=1}^n K(\frac{X_i-x}{h}) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

Remarquons aussi que, si $\sum_{i=1}^n K(\frac{X_i-x}{h}) = 0$, i.e. si x se trouve dans une zone où il n'y a pas de X_i , alors $\hat{r}(x) = 0$. Et sinon, comme $\sum_{i=1}^n \omega_{n,i}(x) = 1$, alors Y_i est une moyenne pondérée des Y_i qui correspondent aux points X_i proches de x .

Consistency of Nadaraya-Watson Estimator

- Here we consider the random design. There are n pairs of IID observations $(X_1, Y_1), \dots, (X_n, Y_n)$ and

$$Y_i = r(X_i) + \varepsilon_i; i = 1, \dots, n;$$

where ε_i 's and X_i 's are independent, and $\mathbb{E}(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$.

- Recall that for chosen smoothing parameter h_n and kernel K , the Nadaraya-Watson estimator of r is given by

$$\hat{r}_n(x) = \hat{r}_n^{NW}(x) = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)}.$$

Consistency of Nadaraya-Watson Estimator

Théorème

Let $h \rightarrow 0$, $nh \rightarrow +\infty$ as $n \rightarrow +\infty$. Let f denote the density of X and Let $\mathbb{E}(Y^2) < +\infty$. Then for any x for which $r(x)$ and $f(x)$ are continuous and $f(x) > 0$, the Nadaraya-Watson estimator $\hat{r}_n(x)$ is a consistent estimator of $r(x)$, that is,

$$\hat{r}_n(x) \xrightarrow{P} r(x) ; \text{ as } n \rightarrow +\infty.$$

Consistency of Nadaraya-Watson Estimator

Proof of the theorem

To prove this theorem, we need to use the following result.

Lemme

(Theorem 1A in Parzen (1962))

Suppose that $w(y)$ is bounded and integrable function satisfying $\lim_{y \rightarrow +\infty} |yw(y)| = 0$. Let g be an integrable function. Then for h such that $h \rightarrow 0$ as $n \rightarrow +\infty$,

$$\lim_{n \rightarrow +\infty} \frac{1}{h} \int w\left(\frac{u-x}{h}\right) g(u) du = g(x) \int w(u) du.$$

for every continuity point x of g .

Proof of the theorem

- In the proof, we drop the subscript of n in h_n . Denote

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$
$$\hat{\psi}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) Y_i.$$

Then $\hat{r}_n(x) = \frac{\hat{\psi}_n(x)}{\hat{f}_n(x)}$. Note that $\hat{f}_n(x)$ is the kernel estimator of $f(x)$.

- It suffices to prove that $\hat{f}_n(x) \xrightarrow{P} f(x)$ and $\hat{\psi}_n(x) \xrightarrow{P} r(x)f(x)$.
- We will prove the latter using the lemma. The proof of the former is similar and simpler.

Consistency of Nadaraya-Watson Estimator

Proof of the theorem: $\hat{\psi}_n(x) \xrightarrow{P} r(x)f(x)$

First we have,

$$\begin{aligned} E(\hat{\psi}_n(x_0)) &\stackrel{IID}{=} \frac{1}{h} E\left(K\left(\frac{x - X_1}{h}\right) Y_1\right) \\ &= \frac{1}{h} E\left(K\left(\frac{x - X_1}{h}\right) (r(X_1) + \epsilon_1)\right) \\ &\stackrel{E(\epsilon)=0}{=} \frac{1}{h} \int K\left(\frac{x - x}{h}\right) r(x) f(x) dx \rightarrow r(x)f(x) \end{aligned}$$

Note that the kernel K satisfies the conditions on w of the lemma. The last convergence follows from the lemma and the symmetry of K .

Similarly we can show that

$$nh \text{Var}(\hat{\psi}_n(x)) \rightarrow (r^2(x) + \sigma^2) f(x) \int K^2(u) du.$$

MISE of the Nadaraya-Watson estimator

Theorem 5.44 in Wasserman (2005)

The mean integrated square error of the Nadaraya-Watson estimator is

$$\begin{aligned} MISE(\hat{r}_n) = & \frac{h_n^4}{4} \left(\int x^2 K(x) dx \right) \int \left(r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right)^2 dx \\ & + \frac{\sigma^2}{nh_n} \int K^2(x) dx \int \frac{1}{f(x)} dx + o(nh_n^{-1} + h_n^4) \end{aligned}$$

- The first term is the squared bias. The term $r'(x) \frac{f'(x)}{f(x)}$ is called the design bias as it depends on the design, that is, the distribution of X_i 's.
- It is known that the kernel estimator has high bias near the boundaries of the data. This is known as boundary bias.

Propriétés asymptotiques

La construction de l'estimateur à noyau de Nadaraya-Watson dépend de deux paramètres, le paramètre de lissage h dont le choix est crucial pour obtenir de bonnes propriétés asymptotiques (citées ci-dessous) et le noyau K dont on ne peut pas négliger le rôle pour la réduction du biais.

Local linear regression

Local linear regression

Suppose that we want to estimate $r(x)$ and X_i is an observation close to x . By Taylor expansion,

$$r(X_i) \approx r(x) + r'(x)(X_i - x) =: a + b(X_i - x).$$

Thus the problem of estimating $r(x)$ is equivalent to estimating a ! Now, we replace $r(X_i)$ with Y_i as we only observe Y_i but not $r(X_i)$. We want to find an a such that $(Y_i - (a + b(X_i - x)))^2$ is small. Take into account all the observations and let \hat{a} and \hat{b} be given by

$$(\hat{a}, \hat{b}) = \operatorname{argmin}_{a, b} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) (Y_i - (a + b(X_i - x)))^2.$$

The local linear estimator is defined as: $\tilde{r}_n(x) := \hat{a}$.

Compare it with the Nadaraya-Watson estimator

$$\hat{r}_n(x) = \operatorname{argmin}_{c \in \mathbf{R}} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) (Y_i - c)^2.$$

Local linear regression

Write $L(a, b) = \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) (Y_i - (a + b(X_i - x)))^2$. Solving the following equation, with $k_i = K\left(\frac{x-x_i}{h}\right)$ and $z_i = X_i - x$,

$$\frac{\partial L(a, b)}{\partial a} = a \sum_{i=1}^n k_i + b \sum_{i=1}^n k_i z_i - \sum_{i=1}^n k_i Y_i = 0$$

$$\frac{\partial L(a, b)}{\partial b} = a \sum_{i=1}^n k_i z_i + b \sum_{i=1}^n k_i z_i^2 - \sum_{i=1}^n k_i Y_i z_i = 0,$$

yields $\hat{a} = \sum_{i=1}^n w_i(x) Y_i / \sum_{i=1}^n w_i(x)$, and thus

$$\tilde{r}_n(x) = \sum_{i=1}^n w_i(x) Y_i / \sum_{i=1}^n w_i(x)$$

Local linear regression

- A linear smoother is defined by the following weighted average :
 $\sum_{i=1} \ell_i(x) Y_i$. Clearly the local linear estimator is a linear smoother, so are the regressogram and the kernel estimator.
- Like Nadaraya-Watson estimator, $\tilde{r}_n(x)$ depends on h . We also need to choose h when using the linear estimator. The cross validation can be done in the same manner as that for N-W estimator.

Comparison

Nadaraya-Watson estimator V.S. Local linear estimator

Theorem 5.65 in Wasserman (2005); see also Fan (1992)

Let $h_n \rightarrow 0$ and $nh_n \rightarrow \infty$, as $n \rightarrow \infty$. Under some smoothing conditions on $f(x)$ and $r(x)$, both $\hat{r}_n(x)$ and $\tilde{r}_n(x)$ have variance

$$\frac{\sigma^2}{nh_n f(x)} \int K^2(u) du + o\left(\frac{1}{nh}\right).$$

The bias of $\hat{r}_n(x)$ is

$$h_n^2 \left(\frac{1}{2} r''(x) + r'(x) \frac{f'(x)}{f(x)} \right) \left(\int u^2 K(u) du \right) + o(h_n^2)$$

whereas $\tilde{r}_n(x)$ has bias $\frac{1}{2} h_n^2 r''(x) \left(\int u^2 K(u) du \right) + o(h_n^2)$.

At the boundary points, the NW estimator typically bears high bias due to the large absolute value of $\frac{f'(x)}{f(x)}$. In this sense, local linear estimation eliminates boundary bias and is free from design bias.

Estimateur par polynomes locaux

Proposition

Si \hat{r}_n est l'estimateur de Nadaraya-Watson associé à un noyau $K \geq 0$ alors \hat{r}_n est solution de

$$\hat{r}_n(x) = \arg \min_{\theta \in \mathbb{R}} \sum K\left(\frac{x - X_i}{h}\right) (Y_i - \theta)^2$$

$\hat{r}_n(x)$ est donc un estimateur des moindres carrés pondéré si $\sum K\left(\frac{x - X_i}{h}\right) \neq 0$

$$\Leftrightarrow \theta = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

C'est un minimum car $\tau'' \equiv 2 \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \geq 0$.

Choix du paramètre de lissage

L'erreur quadratique moyenne MSE (mean square error) est une mesure permettant d'évaluer la similarité de r_n par rapport à la fonction de régression inconnue r , au point x donné de \mathbb{R} .

Nous constatons d'une part que les expressions du biais de $\hat{r}_n(x, h_n)$ et de la variance de $\hat{r}_n(x, h_n)$ permettent de conclure qu'une **grande valeur de h_n donne une augmentation du biais et une diminution de la variance** (estimation fortement biaisée) et qu'un faible paramètre h_n , donne une diminution du biais et une augmentation de la variance (phénomène de sous lissage).

Smoothing Parameter Selection

Afin de déterminer le paramètre de lissage optimal, on s'intéresse à l'optimisation de l'erreur moyenne quadratique $MSE(r_n)$

$$MSE(r_n) = \mathbb{E}(r_n(x) - r(x))^2 \quad \text{et} \quad MISE(r_n) = \int \mathbb{E}(r_n(x) - r(x))^2 f(x) dx$$

Le paramètre de lissage optimal au sens du MSE et du MISE respectivement

$$h_{op1} = \left(\frac{\frac{\sigma^2(x)}{f(x)} [K^2]}{\left((r''(x) + 2r'(x) \frac{f'(x)}{f(x)}) [u^2 K] \right)^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$

$$h_{op2} = \left(\frac{\int \frac{\sigma^2(x)}{f(x)} [K^2] dx}{\int \left((r''(x) + 2r'(x) \frac{f'(x)}{f(x)}) [u^2 K] \right)^2 dx} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

Comment choisir sa fenêtre pour la régression ?

- ▶ Même logique, calculs différents
- ▶ Fenêtre optimale :

$$h_{opt} = \left[\frac{\int \sigma^2(x) f^{-1}(x) dx \int K^2(u) du}{\int \{2 r'(x) f'(x) f^{-1}(x) + r''(x)\}^2 dx \kappa_2^2} \right]^{1/5} n^{-1/5}$$

- ▶ Plug-in : Ben “YAKA” estimer tout ces trucs et remplacer...
- ▶ Règle du pouce : $h_{ROT} \propto \sigma(x) \cdot n^{-\frac{1}{5}}$
- ▶ Validation croisée :

$$h_{CV} = \arg \min_h \frac{1}{n} \sum_{i=1}^N \left(Y_i - \widehat{r}^{-i}(X_i) \right)^2$$

Kernel Regression

Random design

Observations $(X_i, Y_i), i = 1, \dots, n$ from bivariate distribution $f(x, y)$.
The distribution of $f_X(x)$ is unknown.

Fix design

Control the predictor variable, X , then Y is the only random variable. (Eg. the beer-example).
The distribution of $f_X(x)$ is known.

Kernel Regression

Random design

The Nadaraya-Watson estimator

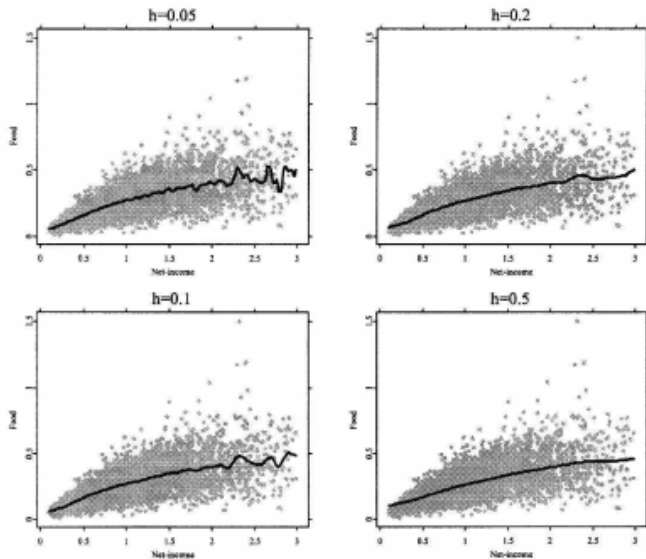
$$\hat{m}(x) = \frac{n^{-1} \sum_{i=1}^n K_h(x - X_i) Y_i}{n^{-1} \sum_{j=1}^n K_h(x - X_j)}$$

Rewrite the Nadaraya-Watson estimator

$$\begin{aligned}\hat{m}(x) &= \frac{1}{n} \sum_{i=1}^n \left(\frac{K_h(x - X_i)}{n^{-1} \sum_{j=1}^n K_h(x - X_j)} \right) Y_i \\ &= \frac{1}{n} \sum_{i=1}^n W_{hi}(x) Y_i\end{aligned}$$

- ▶ Weighted (local) average of Y_i (note: $\frac{1}{n} \sum_{i=1}^n W_{hi}(x) = 1$)
- ▶ h determines the degree of smoothness.
- ▶ What happens if the denominator of $W_{hi}(x)$ is equal to 0? Then the numerator is also equal to 0, and the estimator is not defined. This

Kernel Regression



Kernel Regression

Fix design

$f_X(x)$ is known. Weights of the form

$$W_{hi}^{FD}(x) = \frac{K_h(x - x_i)}{f_X(x)}$$

Simpler structure and therefore easier to analyse.

One particular fixed design kernel regression estimator:

Gasser-Müller

For the case of ordered design points $x_{(i)}, i = 1, \dots, n$ from $[a, b]$

$$W_{hi}^{GM}(x) = n \int_{s_{i-1}}^{s_i} K_h(x - u) du,$$

where $s_i = \frac{x_{(i)} + x_{(i+1)}}{2}$, $s_0 = a$, $s_{n+1} = b$.

Note that $W_{hi}^{GM}(x)$ sums to 1.

Confidence Regions

How close is the smoothed curve to the true curve?

Theorem 4.5: Asymp. distribution of the Nadaraya-Watson est.

Suppose m and f_X are twice differentiable.

$\int |K(u)|^{2+\kappa} du < \infty$ for some $\kappa > 0$,

x is a continuity point of $\sigma^2(x)$, $\mathbb{E}(|Y|^{2+\kappa}|X=x)$ and $f_X(x) > 0$.

Take $h = cn^{-1/5}$. Then

$$n^{2/5} \{ \hat{m}_h(x) - m(x) \} \xrightarrow{L} N(b_x, v_x^2)$$

where

$$b_x = c^2 \mu_2(K) \left\{ \frac{m''(x)}{2} + \frac{m'(x)f'_X(x)}{f_X(x)} \right\}$$

$$v_x^2 = \frac{\sigma^2(x) \|K\|_2^2}{cf_X(x)}$$

Confidence Regions

Suppose bias is of negligible magnitude compare to the variance, e.g. h sufficiently small.

Approx. confidence intervals:

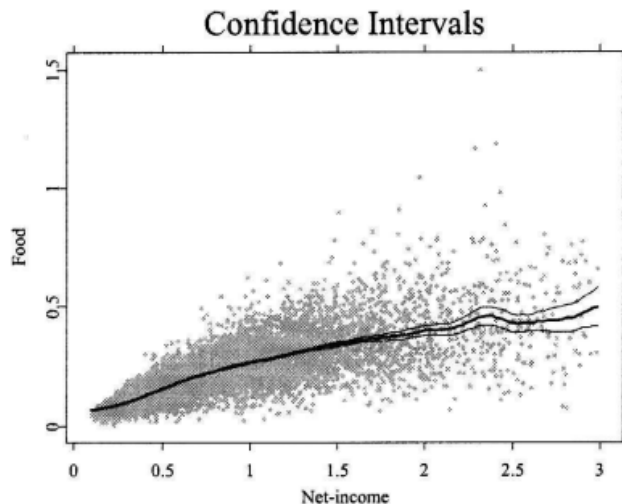
$$\left[\hat{m}_h(x) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\|K\|_2 \hat{\sigma}^2(x)}{n h \hat{f}_h(x)}}, \hat{m}_h(x) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\|K\|_2 \hat{\sigma}^2(x)}{n h \hat{f}_h(x)}} \right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ -quantile of the normal distribution and the estimate of $\sigma^2(x)$ is

$$\hat{\sigma}^2(x) = \frac{1}{n} \sum_{i=1}^n W_{hi}(x) \{Y_i - \hat{m}_h(x)\}^2$$

where W_{hi} is the weights from the Nadaraya-Watson est.

Confidence Regions



The Nadaraya-Watson kernel regression and 95% confidence intervals, $h = 0.2$.

Nearest-Neighbor Estimator

Other nonparametric smoothing techniques that differ from the kernel method.

Nearest-Neighbor Estimator

Kernel regression: Weighted averages Y_i in a *fixed* neighborhood (governed by h) around x .

k -NN: Weighted averages of Y_i in a neighborhood around x . The width is *variable*. Using the k nearest observations to x .

$$\hat{m}_k(x) = \frac{1}{n} \sum_{i=1}^n W_{ki}(x) Y_i$$

where

$$W_{ki}(x) = \begin{cases} n/k & \text{if } i \in J_x \\ 0 & \text{otherwise} \end{cases}$$

with the set of indices

$$J_x = \{i : X_i \text{ is one of the } k \text{ nearest observations to } x\}$$

Nearest-Neighbor Estimator

Remark:

- ▶ When we estimate $m(\cdot)$ at a point x with sparse data, then the k nearest neighbors are rather far away from x .
- ▶ k is the smoothing parameter.
Increasing k makes the estimate smoother.

The k -NN estimator is the **kernel estimator** with **uniform** kernel $K(u) = \frac{1}{2}1_{(|u| \leq 1)}$ and **variable bandwidth** $R = R(k)$ with $R(k)$ being the distance between x and its furthest k -nearest neighbor:

$$\hat{m}_k(x) = \frac{\sum_{i=1}^n K_R(x - X_i) Y_i}{\sum_{i=1}^n K_R(x - X_i)}$$

The k -NN estimator can be generalised by considering other kernel functions.

Nearest-Neighbor Estimator

Bias and variance:

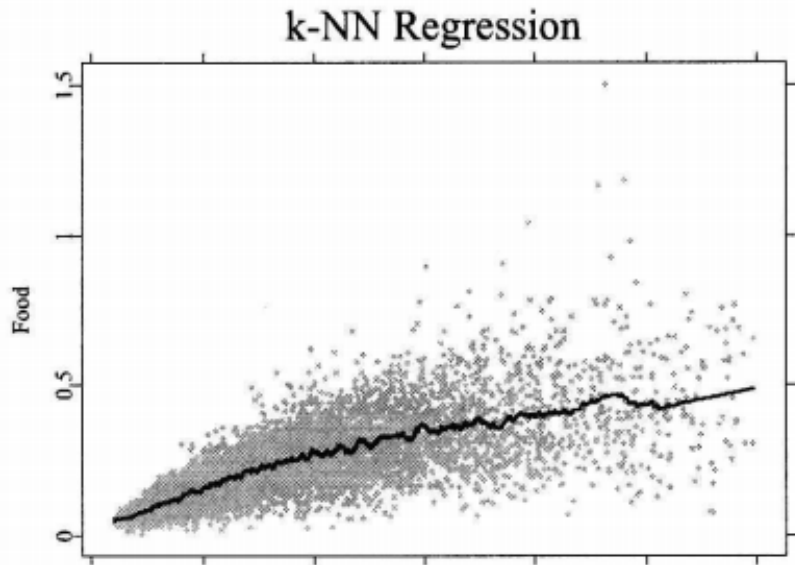
Let $k \rightarrow \infty$, $k/n \rightarrow 0$ and $n \rightarrow \infty$. Then

$$\begin{aligned}\mathbb{E}\{\hat{m}_k(x)\} - m(x) &\approx \frac{\mu_2(K)}{8f_X(x)^2} \left\{ m''(x) + 2 \frac{m'(x)f'_X(x)}{f_X(x)} \right\} \left(\frac{k}{n} \right)^2 \\ \mathbb{V}\{\hat{m}_k(x)\} &\approx 2\|K\|_2^2 \frac{\sigma^2(x)}{k}\end{aligned}$$

Remark:

- ▶ The variance does not depend of $f_X(x)$ (unlike LC) because k -NN always averages over k observations.
- ▶ k -NN approx. identical to LC with bandwidth h wht. MSE: Choose $k = 2nhf_X(x)$.

Nearest-Neighbor Estimator



Orthogonal Series

Suppose that $m(\cdot)$ can be represented by a **Fourier series**

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x)$$

where $\{\varphi_j(x)\}_{j=0}^{\infty}$ is a known basis of functions and $\{\beta_j\}_{j=0}^{\infty}$ are the unknown Fourier coefficients.

Goal: Estimate the unknown Fourier coefficients.

Cannot estimate a infinite number of coefficients from a finite number of observations. Therefore choose N that will be included in the Fourier series representaion.

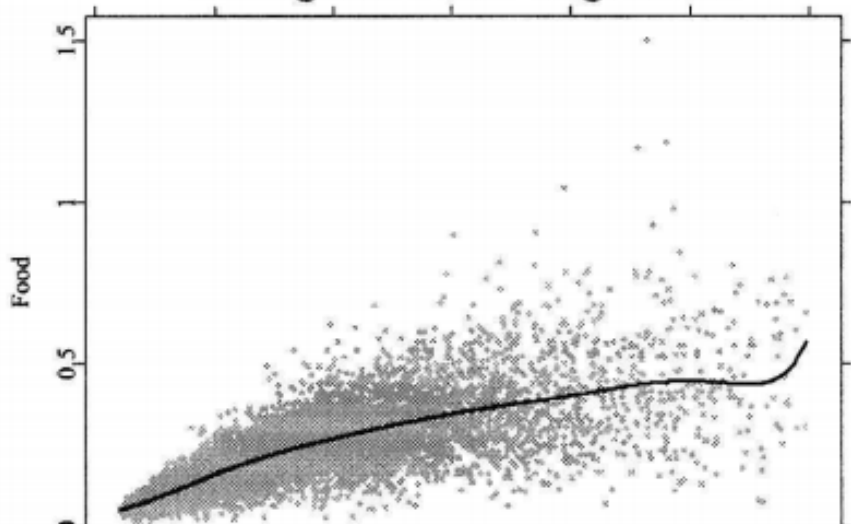
Estimation proceeds in three steps:

- ▶ Select a basis of functions,
- ▶ Select $N < n$ (integer),
- ▶ Estimate N unknown coefficients

N : Smoothing parameter.

- ▶ Large $N \Rightarrow$ many terms in the Fourier series \Rightarrow interpolation,
- ▶ Small $N \Rightarrow$ few terms \Rightarrow smooth estimates.

Orthogonal Series Regression



Multivariate Kernel Regression

How does Y depends on a *vector* of variables, \mathbf{X} ?

We want to estimate

$$\mathbb{E}(Y|\mathbf{X}) = \mathbb{E}(Y|X_1, \dots, X_d) = m(\mathbf{X})$$

where $\mathbf{X} = (X_1, \dots, X_d)^T$.

We have

$$\begin{aligned}\mathbb{E}(Y|\mathbf{X}) &= \int y f(y|\mathbf{x}) dy \\ &= \frac{\int y f(y, \mathbf{x}) dy}{f_{\mathbf{X}}(\mathbf{x})}\end{aligned}$$

Multivariate Nadaraya- Watson estimator: (Local constant estimator)

Replace with kernel densities

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) Y_i}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})}$$

Weighted sum of the observed Y_i .

Multivariate Kernel Regression

Local linear estimator:

The minimization problem

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n \left\{ Y_i - \beta_0 - \beta_1^T (\mathbf{X}_i - \mathbf{x}) \right\}^2 K_H(\mathbf{X} - \mathbf{x})$$

Solution

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1^T)^T = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

$$\text{where } \mathbf{X} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T \\ \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ and}$$

$$\mathbf{W} = \text{diag}(K_H(\mathbf{X}_1 - \mathbf{x}), \dots, K_H(\mathbf{X}_n - \mathbf{x}))$$

$\hat{\beta}_0$ estimates the regression function. $\hat{m}_{1,H}(\mathbf{x}) = \hat{\beta}_0(\mathbf{x})$.

$\hat{\beta}_1$ estimates the partial derivatives wrt. the components \mathbf{x} .

Multivariate Kernel Regression

Statistical properties of the Nadaraya-Watson estimator:

Theorem 4.8: Asymptotic bias and variance

The conditional asymptotic bias and variance of the multivariate Nadaraya-Watson kernel regression estimators are

$$\begin{aligned}\text{Bias}(\hat{m}_{\mathbf{H}}|\mathbf{X}_1, \dots, \mathbf{X}_n) &\approx \mu_2(\mathcal{K}) \frac{\nabla_m(\mathbf{x})^T \mathbf{H} \mathbf{H}^T \nabla_f(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \\ &\quad + \frac{1}{2} \mu_2(\mathcal{K}) \text{tr}\{\mathbf{H}^T \mathcal{H}_m(\mathbf{x}) \mathbf{H}\} \\ \mathbb{V}(\hat{m}_{\mathbf{H}}|\mathbf{X}_1, \dots, \mathbf{X}_n) &\approx \frac{1}{n \det(\mathbf{H})} \|\mathcal{K}\|_2^2 \frac{\sigma(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\end{aligned}$$

in the interior of the support of $f_{\mathbf{X}}$.

Multivariate Kernel Regression

Statistical properties of the local linear estimator:

Theorem 4.9: Asymptotic bias and variance

The conditional asymptotic bias and variance of the multivariate local linear kernel regression estimator are

$$\begin{aligned}\text{Bias}(\hat{m}_{1,\mathbf{H}}|\mathbf{X}_1, \dots, \mathbf{X}_n) &\approx \frac{1}{2}\mu_2(\mathcal{K})\text{tr}\{\mathbf{H}^T\mathcal{H}_m(\mathbf{x})\mathbf{H}\} \\ \mathbb{V}(\hat{m}_{1,\mathbf{H}}|\mathbf{X}_1, \dots, \mathbf{X}_n) &\approx \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\frac{\sigma(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\end{aligned}$$

in the interior of the support of $f_{\mathbf{X}}$.

Multivariate Kernel Regression

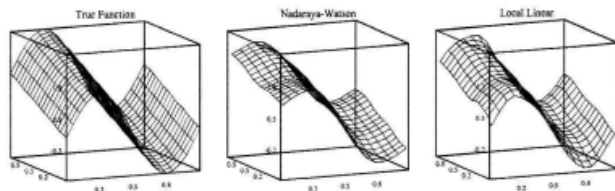
Practical aspects:

d=2:






Comparison of Nadaraya-Watson and the local linear two-dimensional estimate for simulated data. 500 design points uniformly distributed in $[0, 1] \times [0, 1]$ and

$$m(\mathbf{x}) = \sin(2\pi x_1) + x_2$$

and error term $N(0, \frac{1}{4})$. Bandwidth $h_1 = h_2 = 0.3$



Bibliographie

-  Fan, J. and Gijbels, I., (1996). Local Polynomial Modelling and its Applications. In : Monographs on Statistics and Applied Probability. vol. 66. Chapman & Hall.
-  Bosq, D., and Lecoutre, J.P. Theorie de l'estimation fonctionnelle (in french). Economica.
-  Gasser, T. and Miiller, H.-G. (1979). Kernel estimation of regression functions. In Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics, 757, 23-68. Springer-Verlag, New York.
-  Gasser, T. and Miiller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. Scand. J. of Statist., 11, 171-185.
-  Gyorfi, L., Hardle, W., Sarda, P. and Vieu, P. (1989). Nonparametric Curve Estimation from Time Series. Lecture Notes in Statistics, 60. Springer-Verlag, Berlin.