Introduction aux statistiques non-paramétriques

F. Roueff

9 février 2021

Table des matières

1 Modèles non-paramétriques **3**
- 1.1 Modèles paramétriques et non-paramétriques 3
- 1.2 Fonctions d'erreur 6
- 1.3 Régularité fonctionnelle 8

2 Estimateurs à noyau **12**
- 2.1 Estimation d'une densité 12
- 2.2 Densités à plusieurs variables 13
- 2.3 Le compromis biais - variance. 14
- 2.4 Principe de l'estimateur sans biais du risque par validation croisée 17
- 2.5 Le modèle du bruit additif 21
- 2.6 Estimateur de Nadaraya-Watson 21
- 2.7 Estimateurs de régression par polynômes locaux 23

3 Estimation par projection **27**
- 3.1 Cadre hilbertien 27
- 3.2 Définition 28
- 3.3 Estimateurs de projection pour la densité 29
- 3.4 Borne supérieure du risque intégré 30
- 3.5 Estimation sans biais du risque 32
- 3.6 Estimateur de projection pour la régression 34
- 3.7 Un exemple de critère pénalisé pour la régression 35

4 Exercices **39**
- 4.1 Des rappels autour de la régression linéaire 39
- 4.2 Estimation d'une densité par l'histogramme 41

Préambule

Ce document est une introduction à l'estimation non-paramétrique. Mentionnons les ouvrages suivants que le lecteur pourra avantageusement consulter pour approfondir ses connaissances :

— Pour une approche mathématique plus complète : Tsybakov [2004],

— Pour la pratique courante et une vision plus globale de l'apprentissage statistique : Hastie et al. [2001].

Chapitre 1

Modèles non-paramétriques

1.1 Modèles paramétriques et non-paramétriques

Un modèle paramétrique, c'est-à-dire prenant la forme $\mathcal{P} = \{\mathbb{P}_\theta \, ; \, \theta \in \Theta\}$ avec $\Theta \subseteq \mathbb{R}^d$, $d \geq 1$ peut s'avérer en pratique inadapté à la "réalité" étudiée, car trop éloigné de celle-ci. Les méthodes statistiques basées sur ce modèle peuvent alors devenir inopérantes. On est conduit de ce fait à introduire des modèles d'observation statistique $\mathcal{P} = \{\mathbb{P}_\theta \ : \ \theta \in \Theta\}$ dans lesquels l'ensemble Θ revêt un aspect *plus général* que ceux autorisés en modélisation paramétrique.

1.1 Exemple (modèle non-paramétrique):
On dispose de deux échantillons indépendants $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_p$ de lois $\mathbb{P}_X$ et $\mathbb{P}_Y$ inconnues. On cherche une méthode statistique permettant de répondre à la question : "ces deux échantillons ont-ils la même distribution ?" Si notre connaissance de la situation nous permet d'introduire un **a priori** de la forme "les X_i et les Y_i possèdent une densité", il en découlera une représentation des lois possibles (on dira : *lois admissibles*) pour $\mathbb{P}_X$ et $\mathbb{P}_Y$ respectivement sous la forme $\prod_i f(x_i)\, dx_i$ et $\prod_i g(y_i)\, dy_i$, où les éléments inconnus sont les fonctions f et g. Cette situation étant très différente du cas (paramétrique) où notre connaissance du modèle nous permettrait d'introduire p.ex. un *a priori* "les X_i et les Y_i sont gaussiennes" et où seuls les paramètres (μ_1, σ_1^2) et (μ_2, σ_2^2) des densités gaussiennes f et g resteraient inconnus.

L'ensemble Θ correspondant à notre modèle sera donc un ensemble de fonctions dont "l'étendue" dépendra de la connaissance qu'il sera possible de formuler sur les fonctions f et g. Cet ensemble Θ n'est donc plus ici un sous ensemble d'un espace vectoriel de dimension finie, comme c'est le cas pour un modèle paramétrique. $\diamondsuit$

1.2 Exemple (modèle semi-paramétrique):
On dispose d'un échantillon $X_1, \ldots, X_n$ i.i.d. de loi $f(\Delta - x)\, dx$. On cherche une méthode statistique permettant de répondre à la question : Δ est-il nul ?

On construira un modèle **semi-paramétrique** basé par exemple sur un *a priori* du type "$\Delta \in \mathbb{R}$, f densité symétrique continue sur $\mathbb{R}$" (si la connaissance du problème permet de réduire ainsi l'ensemble de fonctions f qui doivent être prises en considération). $\diamondsuit$

La plupart des notions introduites pour les modèles paramétriques s'appliquent ou se généralisent aux modèles non-paramétriques : identifiabilité, statistiques exhaustives, statistiques complètes, risque, biais, variance, vraisemblance etc. La principale difficulté réside dans la description de l'ensemble des paramètres Θ : celui-ci sera au mieux une sous-classe d'un espace métrique ou hilbertien de dimension infinie. Cette complexité de l'ensemble Θ complique toute la méthodologie de l'estimation à mettre en oeuvre.

Voici quelques exemples classiques de problèmes non-paramétriques d'estimation :

(a) **Estimation de densité**(voir fig. 1.1) $(X_i)_{i=1,\dots,n}$ sont des variables aléatoires indépendantes et identiquement distribués (i.i.d.) et on note $(X_1, \dots, X_n) \sim \mathbb{P}_{\theta,n}$, de sorte que $\mathbb{P}_{\theta,n} = \mathbb{P}_\theta^n, \quad \theta \in \Theta$. Si nous savons que les X_i admettent une densité et si le problème consiste à estimer celle-ci, on écrira $\mathbb{P}_\theta(dt) = \theta(t)\,dt$ et, en absence de toute autre information *a priori* sur $\theta(t)$, on prendra pour Θ l'ensemble de toutes les fonctions qui sont des densités de probabilité : $\Theta = \{f \in L^1(\mathbb{R}) \ : \ f \geq 0, \int f = 1\}$.

(b) **Régression** (fig. 1.2) On dispose des observations $X_i = (Y_i, Z_i)$, $i = 1, \dots, n$ avec un *a priori* $Z_i = f(Y_i, \epsilon_i)$, où $(\epsilon_i)_{i=1,\dots,n}$ sont des variables aléatoires i.i.d. et $(Y_i)_{i=1,\dots,n}$ sont des variables déterministes ou des v.a. i.i.d. indépendantes de $\{\epsilon_i\}$.
Nous avons ici $\theta = (f, \mathbb{P}_\epsilon, \mathbb{P}_Y)$, où f est le **paramètre d'intérêt**, $\mathbb{P}_\epsilon$ est la loi du **bruit** $(\epsilon_i)_{i=1,\dots,n}$ (cette loi dépendra souvent d'un **paramètre de nuisance**) et $\mathbb{P}_Y$ est la loi de $Y_1, \dots, Y_n$. L'exemple le plus simple est celui du **bruit additif** :

$$Z_i = f(Y_i) + \epsilon_i$$

avec, par exemple, $Y_i = i/n$ et $f : [0,1] \to \mathbb{R}$ (dispositif expérimental régulier).

(c) **Modèle de bruit additif** Il s'agit d'un modèle idéalisé à variable continue (par exemple le temps). Il s'écrit

$$\mathrm{d}X(t) = f(t)\,\mathrm{d}t + \mathrm{d}Z(t)\ ,$$

où f est une fonction de carré intégrable inconnue et Z désigne un processus centré à accroissements stationnaires indépendants, que l'on prendra toujours gaussien de mesure d'intensité $\sigma \mathrm{d}t$. Concrètement cela signifie que l'on observe toutes les variables

$$\int \phi\,\mathrm{d}X = \int \phi\, f + \int \phi\,\mathrm{d}Z\ ,$$

où ϕ parcourt l'ensemble L^2 des fonctions de carré intégrable et $\{\int \phi \, \mathrm{d}Z\}_{\phi \in L^2}$ est un processus gaussien centré dont la covariance est donnée par

$$\mathrm{Cov}\left(\int \phi \, \mathrm{d}Z, \int \psi \, \mathrm{d}Z\right) = \sigma^2 \int \phi\psi \; .$$

Dans ce modèle le nombre d'observations est potentiellement infini mais f ne peut être estimé de façon consistante que si σ tend vers 0.

(d) **Densité spectrale** (fig. 1.3) $(X_i)_{i=1,\dots,n}$ est un n-échantillon d'un processus stationnaire du second ordre de densité spectrale $f \in L^1(-\pi, \pi)$ qu'il s'agit d'estimer.

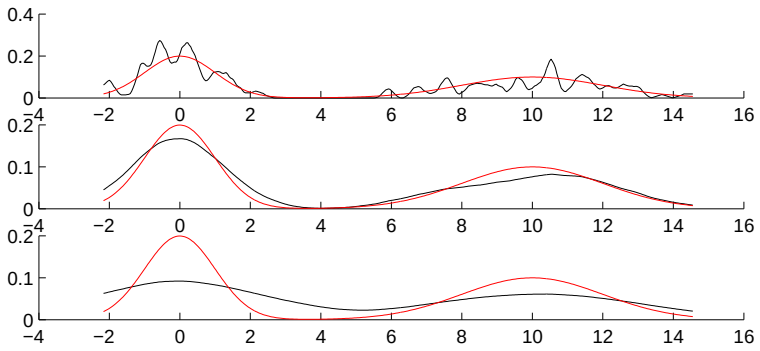

FIGURE 1.1 – Estimation de la densité d'un mélange de deux gaussiennes à partir d'un échantillon i.i.d. de taille 200. En rouge : densité réelle ; en noir : densités estimées par la méthode des noyaux pour trois différentes valeurs de h.

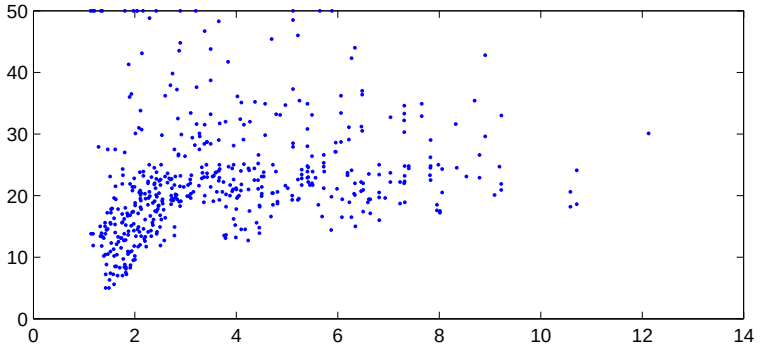

FIGURE 1.2 – Graphe nuage : prix maison (unité 1000$) médian d'une ville proche de Boston VS distance d'éloignement des 5 grands centres de travail

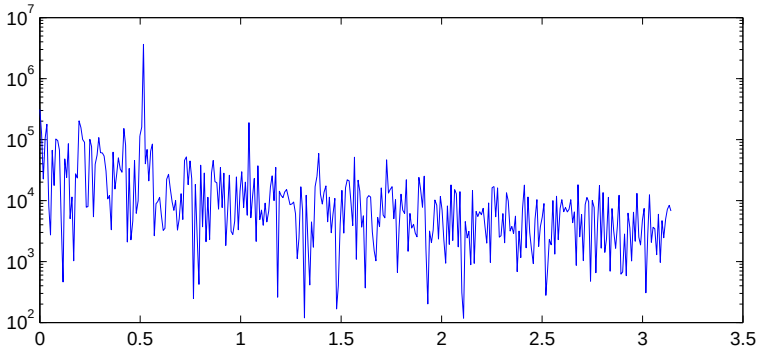

FIGURE 1.3 – Log-périodogramme d'une série temporelle issue de relevés mensuels des niveaux d'une rivière californienne

1.2 Fonctions d'erreur

Dans les trois exemples décrits ci-dessus, la loi du vecteur d'observation $X = (X_1, \cdots, X_n)$ dépend de la taille n de l'échantillon d'observation et d'une fonction inconnue $f : I \to \mathbb{R}$ (la même pour tout n). Cette fonction inconnue constitue le **paramètre d'intérêt** dans les situations traitées par ces exemples. Une estimation $\widehat{f}$ de la valeur $f(t)$ de f au point t est une fonction de t et de l'observation x que l'on a effectuée. Autrement dit, un *estimateur* de $f(t)$ est une fonction $\widehat{f}(t, X)$ de t et du vecteur d'observation X, où chaque $\widehat{f}(t, x)$ est la valeur par laquelle on estime $f(t)$ lorsque l'on a observé x. Il y a donc lieu de se rappeler que l'écriture $\widehat{f}$ ou $\widehat{f}(t)$ souvent utilisée constitue un raccourci.

On doit se doter maintenant d'un critère d'évaluation de l'estimateur $\widehat{f}$ qui permettrait de dire quelle est l'erreur que l'on commet en substituant à l'inconnue f son estimée $\widehat{f}$. c'est donc "l'écart" entre les deux fonctions f et $\widehat{f}$ qui doit être évalué. La méthode consiste à dire qu'on estime f en tant qu'élément d'un espace fonctionnel $(H, \|\cdot\|_H)$ et que si l'on a également $\widehat{f} \in H$, alors l'erreur commise en utilisant $\widehat{f}$ quand f est la vraie valeur de l'inconnue vaut $l(\widehat{f}, f) = \|\widehat{f} - f\|_H^2$. Par exemple, quand $H = L^2(I)$, on a :

$$l(\widehat{f}, f) = \|\widehat{f} - f\|_2^2 = \int_I (\widehat{f}(t) - f(t))^2 \, dt. \tag{1.1}$$

ce qui constitue également un raccourci, car l'erreur écrite dans 1.1 est dépendante de l'observation x et l'on devrait en réalité écrire :

$$l(\widehat{f}, f)(x) = \|\widehat{f}(., x) - f\|_2^2 = \int_I (\widehat{f}(t, x) - f(t))^2 \, dt \tag{1.2}$$

On introduit ensuite deux types de risque : le **Risque Quadratique** (MSE en anglais : mean square error), ainsi que le **Risque Quadratique Intégré** (MISE en anglais : mean integrated square error). Le risque MISE

(cf. ci-dessous) apparaît comme directement associé à cette distance (erreur) quadratique définie par (1.2). Plus précisément :

1. le **Risque Quadratique** (MSE). Si on regarde l'estimation de $f(t)$ à t fixé, on retrouve le cadre d'un problème *paramétrique* d'estimation du paramètre inconnu $(f(t))_f$ et on peut introduire le risque quadratique associé à ce problème :

$$\mathrm{MSE}_f(t) = \mathbb{E}_f[(\widehat{f}(t,X) - f(t))^2] \tag{1.3}$$

C'est la valeur moyenne de l'écart $(f(t)-\widehat{f}(t,X))^2$ sous l'hypothèse que f est la vraie fonction inconnue du problème (p.ex. la vraie densité dans le cas 1. cité ci-dessus). Pour un estimateur $\widehat{f}_n(t,X_1,\cdots,X_n)$ basée sur l'observation d'une suite i.i.d. de densité inconnue f à estimer, cette expression s'écrit :

$$\mathrm{MSE}_f(t) = \int_{\mathbb{R}^n} (\widehat{f}_n(t,x_1,\cdots,x_n) - f(t))^2 f(x_1)\cdots f(x_n)\, dx_1\cdots dx_n. \tag{1.4}$$

Lorsqu'il n'y aura pas de confusion possible, on écrira $\mathrm{MSE}(t)$ au lieu de $\mathrm{MSE}_f(t)$.

2. le **Risque Quadratique Intégré** (MISE). C'est la valeur moyenne de écarts $l(\widehat{f},f)(x)$ de (1.1), sous l'hypothèse que la fonction inconnue coïncide avec f :

$$\begin{aligned}
\mathrm{MISE}(\widehat{f},f) &= \mathbb{E}_f\left[\,\|\widehat{f}(.,X) - f\|_2^2\,\right] \\
&= \mathbb{E}_f\left[\int_{\mathbb{R}} (\widehat{f}(t,X) - f(t))^2\, dt\right] \\
&= \int_{\mathbb{R}} \mathbb{E}_f[(\widehat{f}(t,X) - f(t))^2]\, dt \\
&= \int_{\mathbb{R}} \mathrm{MSE}_f(t) dt. \qquad (1.5)
\end{aligned}$$

l'inversion des espérances dans (1.5) découlant du théorème de Fubini. Ici encore, l'écriture $\mathrm{MISE}(\widehat{f},f)$ pourra être simplifiée en $\mathrm{MISE}(f)$ ou simplement MISE quand aucune confusion ne sera à craindre. Pour un estimateur $\widehat{f}_n(t,X_1,\cdots,X_n)$ basée sur l'observation d'une suite i.i.d. de densité inconnue f à estimer, la formule 1.5 devient :

$$\mathrm{MISE}(f) = \int_{t\in\mathbb{R}} \int_{x_1^n\in\mathbb{R}^n} (\widehat{f}_n(t,x_1^n) - f(t))^2\, f(x_1)\cdots f(x_n)\, dx_1\cdots dx_n dt.$$

1.3 Régularité fonctionnelle

Transposer au cadre non-paramétrique les techniques d'estimation paramétrique nécessite certaines précautions. Par exemple, si l'on veut estimer la densité inconnue f d'un échantillon $(X_1, \ldots, X_n)$ par une méthode du type du maximum de vraisemblance, on ne cherchera pas $\widehat{f}$ correspondant à la condition :

$$\widehat{f} = \underset{f \in L^1(\mathbb{R}): f \geq 0, \int f = 1}{\operatorname{argmax}} \left\{ \sum_{k=1}^{n} \log(f(X_i)) \right\} \tag{1.6}$$

puisque la classe L^1 est beaucoup trop vaste. En effet, il est clair que pour chaque point $x \in \mathbb{R}$, on peut trouver une suite de densités de probabilité f_n telle que $f_n(x) \to +\infty$, ce qui nous montre que le problème énoncé par (1.6) n'a simplement pas de solution dans la classe L^1. Dans certains cas, on remplacera L^1 par un ensemble plus restreint avec l'idée d'une *projection* sur un sous-ensemble de L^1. D'autres méthodes (méthodes à noyaux, pénalisations) permettront d'aboutir au même type de résultat : *maximiser une vraisemblance ou une fonction d'erreur empirique en imposant une opération de lissage.*

Il est donc important d'introduire des notions de **classes de régularité**. Celles-ci sont définies assez simplement à l'aide d'opérations usuelle de l'analyse (dérivées, incréments, normes L^p etc) mais elles sont toujours reliées à des propriétés du type suivant :

Pour $f \in \mathcal{C}$, quelle est l'erreur d'approximation de f par f_h, où h est un paramètre de lissage.

Pour illustrer ce principe, nous introduisons deux définitions simples de classes de régularité. On note

$$\langle f, g \rangle = \int_I f(t)\overline{g(t)}\, dt$$

le produit hermitien usuel sur $L^2(0,1)$ et $\{e_k(t) = e^{2\,i\,\pi\,k\,t},\ k \in \mathbb{Z}\}$ la base de Fourier associée.

Définition 1.3.1 (Espaces de Sobolev sur $L^2(0,1)$)
Pour $s > 0$, on appelle espace de Sobolev $\mathcal{W}^s(0,1)$ l'espace des fonctions $f \in L^2(0,1)$ qui vérifient la condition :

$$\|f\|_{\mathcal{W}^s} = \left(\sum_{k \in \mathbb{Z}} |\langle f, e_k \rangle|^2 \, (1 + |k|)^{2s} \right)^{\frac{1}{2}} < \infty \tag{1.7}$$

Une définition intéressante de classes de régularité est alors

$$\mathcal{W}(s, L) := \{ f \in \mathcal{W}^s(0,1) \ : \ \|f\|_{\mathcal{W}^s} \leq L \}.$$

Cette classe traduit de façon évidente la qualité de l'approximation de f par sa projection

$$f_n = \sum_{k=-n}^{n} \langle f, e_k \rangle \, e_k$$

En effet comme (e_k) est une base orthonormée de $L^2(0,1)$, l'erreur d'approximation L^2, s'écrit

$$\begin{aligned} \|f - f_n\|_2 &= \left(\sum_{|k|>n} |\langle f, e_k \rangle|^2 \right)^{\frac{1}{2}} \\ &\leq \left(\sum_{|k|>n} |\langle f, e_k \rangle|^2 \frac{(1+|k|)^{2s}}{(2+n)^{2s}} \right)^{\frac{1}{2}} \\ &\leq L \frac{1}{(2+n)^s}, \end{aligned} \tag{1.8}$$

où la dernière inégalité est vérifiée pour tout $f \in \mathcal{W}(s, L)$.

Un autre point important des espaces de Sobolev concerne le cas où s est un entier.

Proposition 1.3.2
Soient $n \geq 1$ et $f \in L^2(0,1)$. Alors les deux propriétés suivantes sont équivalentes :

(i) $f^{(n)} \in L^2(0,1)$,

(ii) *Il existe un polynôme p de degré n tel que $f + p \in \mathcal{W}^n(0,1)$.*

Démonstration La preuve se base sur le fait que, pour tout f tel que $f' \in L^2(0,1)$, une intégration par partie donne

$$(2i\pi k)\langle f, e_k \rangle = (f(1) - f(0)) - \langle f', e_k \rangle.$$

De plus une fonction f est dans $L^2(0,1)$ si et seulement si la série $\sum_{k\in\mathbb{Z}} |\langle f, e_k \rangle|^2$ est finie. ∎

Notons que la définition des espaces de Sobolev existe aussi sur $\mathbb{R}$: $f \in L^2(\mathbb{R})$ est dans $W^s(\mathbb{R})$ si et seulement si

$$\int |\widehat{f}(\xi)|^2 (1+|\xi|)^{2s} \, d\xi < \infty,$$

où $\widehat{f}$ est la transformée de Fourier sur $L^2(\mathbb{R})$. Pour s entier, cette condition revient à supposer $f^{(s)} \in L^2(\mathbb{R})$.

Dans la suite on note $\Lambda(\alpha, L)$ la classe de Hölder définie par :

Définition 1.3.3 (Classes de Hölder)
Soit $k \in \mathbb{N}$, $\beta \in (0,1]$, et $L > 0$. On définit la classe de Hölder $\Lambda(k+\beta, L)$ comme étant l'ensemble des fonctions $f : \mathbb{R} \to \mathbb{R}$ telles que f soit k fois continûment dérivable ($f \in C^k$) et $f^{(k)}$ satisfait la condition :

$$|f^{(k)}(t) - f^{(k)}(s)| \leq L\,|t-s|^{\beta} \quad \forall s,t \in \mathbb{R}$$

Ici on a directement défini une classe $\Lambda(\alpha, L)$ en introduisant la constante L dans la définition qui a le même rôle que la semi-norme $\|\cdot\|_{\mathcal{W}^s}$ pour les classes de Sobolev. On a encore un résultat d'approximation, un peu moins élémentaire que pour les classes de Sobolev, qui relie calsses de Hölder et approximations par un noyau.

Définition 1.3.4 (Noyaux)
On appelle **noyau** *une fonction $K : \mathbb{R} \to \mathbb{R}$, intégrable, telle que $\int_{\mathbb{R}} K = 1$. On dit que le noyaux K est d'ordre j ($j \in \mathbb{N}$), si K vérifie les conditions :*

$$\int_{\mathbb{R}} K(t)\, t^i\, dt = 0 \qquad (i = 1, \ldots, j). \tag{1.9}$$

On remarque qu'un noyau paire vérifie (1.9) pour tout i impair dés que l'intégrale est définie.

Une approximation de $f(t_0)$ est alors obtenu en calculant

$$\frac{1}{h} \int f(t)\, K\left(\frac{t-t_0}{h}\right) dt$$

Le résultat donne la qualité de cette approximation.

Proposition 1.3.5
Soit $f \in \Lambda(\alpha = l + \beta, L)$ et supposons que K est un noyau d'ordre l tel que

$$C_\alpha = \int |t|^\alpha |K(t)|\, dt < \infty.$$

Alors, pour tout $h > 0$,

$$\sup_{t_0} \left| \frac{1}{h} \int f(t)\, K\left(\frac{t-t_0}{h}\right) dt - f(t_0) \right| \leq \frac{L\, C_\alpha}{l!}\, h^\alpha. \tag{1.10}$$

Démonstration L'erreur d'approximation ci-dessus s'écrit aisément (en utilisant que $\int K = 1$)

$$b_f(t_0) := \int_{\mathbb{R}} K(u)[f(hu + t_0) - f(t_0)]du. \tag{1.11}$$

On développe f à l'ordre l au voisinage de t_0 :

$$f(t_0+hu) = f(t_0) + f'(t_0)hu + f''(t_0)\frac{(hu)^2}{2} + \ldots + f^{(l)}(t_0+hu\tau(u))\frac{(hu)^l}{l!}$$

avec $0 \leq \tau(u) \leq 1$. On a alors

$$b_f(t_0) = \sum_1^{l-1} \int_{\mathbb{R}} f^{(k)}(t_0)K(u)\frac{(hu)^k}{k!}du + \frac{h^l}{l!}\int_{\mathbb{R}} f^{(l)}(t_0+hu\tau(u))K(u)u^l du$$

Le noyau K étant d'ordre l, tous les termes à droite ci-dessus sont nuls, à l'exception du dernier pour lequel on intoduit le terme suivant :

$$b_f(t_0) = \frac{h^l}{l!}\int_{\mathbb{R}} \left[f^{(l)}(t_0+hu\tau(u)) - f^{(l)}(t_0)\right] K(u)u^l du + \frac{h^l}{l!}\int_{\mathbb{R}} f^{(l)}(t_0)K(u)u^l du$$

où, pour la même raison encore, seul la première intégrale à droite est non-nulle. Maintenant, nous pouvons majorer grace à la propriété hölderienne de f :

$$|b_f(t_0)| \;\leq\; \frac{Lh^l}{l!}\int |hu\tau(u)|^{\alpha-l}|K(u)||u|^l du \;\leq\; \frac{Lh^\alpha}{l!}\int |u|^\alpha |K(u)|du \quad (1.12)$$

ce qui donne bien le résultat. ∎

Il convient aussi de mentionner à ce point que dans les récentes décennies, l'apport des **ondelettes** s'est révélé un outil très puissant pour analyser la régularité d'une fonction définie sur $\mathbb{R}$. Ce type de résultat sort néanmoins du cadre de ce cours.

Chapitre 2

Estimateurs à noyau

2.1 Estimation d'une densité

Dans cette section, le modèle d'observation étudié sera toujours associé à un vecteur d'observations $(X_1, \cdots, X_n)$, où les $(X_i)_{i\in\mathbb{N}}$ sont des variables aléatoires i.i.d. admettant une densité $f(t)$ inconnue. On s'intéresse à la construction des estimateurs de la densité $f(t)$. Le modèle d'observation correspond donc à la famille de lois : $\mathbb{P}_{f,n} = \prod_1^n f(x_i)dx_i$. On étudiera ici les **estimateurs** de $f(t)$ **associés à un noyau** K.

Définition 2.1.1 (Estimateur de densité associé à un noyau)
Le modèle d'observation étant de la forme $\mathbb{P}_{f,n} = \prod_1^n f(x_i)dx_i$, et un noyau K étant donné, on définit, pour $n \in \mathbb{N}$ et $h > 0$ fixés, l'estimateur :

$$\widehat{f}_{n,h}(t, X_1, \cdots, X_n) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - t}{h}\right). \tag{2.1}$$

On dit que $\widehat{f}_{n,h}(t, X_1, \cdots, X_n)$ est un estimateur de $f(t)$ **associé au noyau** *K et basé sur n observations. La constante h est appelée* **fenêtre** *de l'estimateur $\widehat{f}_{n,h}$.*

Motivation. Si les (X_i) sont i.i.d., les variables aléatoires $Y_i = \mathbb{1}_{]t-h,t+h]}(X_i)$ le sont également, avec comme loi commune la loi de Bernouilli dont le paramètre est égal à la constante $P[t - h < X_i \leq t + h]$ et, d'après la loi forte des grands nombres, la suite de variables aléatoires :

$$P_{n,h}(t, X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]t-h,t+h]}(X_i) \tag{2.2}$$

convergera presque sûrement vers $P[t - h < X_i \leq t + h]$:

$$P_{n,h}(t, X_1, \cdots, X_n) \longrightarrow P[t - h < X_i \leq t + h] \quad (p.s.) \tag{2.3}$$

(ceci pour tout t). Par ailleurs, si $f(t)$ est la densité commune des X_i, alors :

$$f(t) \approx \frac{1}{2h}\int_{t-h}^{t+h} f(u)du = \frac{P[t-h<X_i\leq t+h]}{2h} \tag{2.4}$$

pour h petit *et le terme de droite convergera réellement vers le terme de gauche pour* $h \to 0$ *si la fonction* f *est assez régulière* (il suffit p.ex. qu'elle soit continue). Il est donc naturel de s'interroger sur les qualités de l'estimateur de $f(t)$ obtenu en remplaçant $P[t-h<X_i\leq t+h]$ par $P_{n,h}(t,X_1,\cdots,X_n)$ dans (2.4) :

$$\widehat{f}_{n,h}(t,X_1,\cdots,X_n) = \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{1}_{]t-h,t+h]}(X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i-t}{h}\right). \tag{2.5}$$

avec $K(u) = \frac{1}{2}\mathbb{1}_{]-1,1]}(u)$.

L'étude portera donc ici sur les condition sous lesquelles un estimateur du type (2.1) est un bon estimateur de $f(t)$ et, en particulier, sur l'évolution (l'amélioration) de ses performances lorsque $n \to \infty$ et $h \to 0$. Le choix de $K(u)$ sera un autre élément permettant d'obtenir de bons estimateurs à noyau. Ce choix dépendra en général de notre connaissance *a priori* du modèle d'expérimentation statistique.

2.2 Densités à plusieurs variables

Ce qui vient d'être dit peut être généralisé à des observations vectorielles. Limitons-nous au cas où les observations constitueraient une suite i.i.d. de vecteurs aléatoires de dimension 2, $(X_1,Y_1),\cdots,(X_n,Y_n),\cdots$ de densité commune $f(u,v)$. Poursuivons l'approche heuristique développée ci-dessus, en l'adaptant à la dimension 2. A la place des v.a. Y_i, on introduira les v.a. $Z_i = \mathbb{1}_{]x-h,x+h]\times]y-h,y+h]}(X_i,Y_i)$ et on aura :

$$\begin{aligned} P_{n,h}(x,y,(X_1,Y_1),\cdots,(X_n,Y_n)) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]x-h,x+h]\times]y-h,y+h]}(X_i,Y_i) \\ &\longrightarrow P[x-h<X_i\leq x+h\ ,\ y-h<Y_i\leq y+h] \end{aligned}$$

la convergence étant presque sûre, quel que soit (x,y). Et par ailleurs :

$$f(x,y)) \approx \frac{1}{4h^2}\int_{x-h}^{x+h}\int_{y-h}^{y+h} f(u,v)dudv = \frac{P[x-h<X_i\leq x+h\ ,\ y-h<Y_i\leq y+h]}{4h^2}$$

et ici également, il y a convergence pour $h \to 0$ dès que f est suffisamment régulière. En remplaçant donc la probabilité à droite dans cette équation par son estimée$P_{n,k}$, on définit l'estimateur :

$$\begin{aligned}\widehat{f}_{n,h}(x,y,(X_1,Y_1),\cdots,(X_n,Y_n)) &= \frac{1}{4nh^2}\sum_{i=1}^{n} \mathbb{1}_{]x-h,x+h]}(X_i)\mathbb{1}_{]y-h,y+h]}(Y_i) \\ &= \frac{1}{nh^2}\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{1}_{[-1,1]}\left(\frac{X_i-x}{h}\right)\right)\left(\frac{1}{2}\mathbb{1}_{[-1,1]}\left(\frac{Y_i-y}{h}\right)\right).\end{aligned}$$

qui est la forme particulière, pour $K(u) = \frac{1}{2}\mathbb{1}_{[-1,1]}(u)$, de l'estimateur :

$$\widehat{f}_{n,h}(x,y,(X_1,Y_1),\cdots,(X_n,Y_n)) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)K\left(\frac{Y_i-y}{h}\right). \tag{2.6}$$

avec $K(u)$ un noyau arbitraire à une variable. On retiendra (2.6) comme une généralisation possible de la définition des estimateurs à noyau pour l'estimation des densités à deux variables.

2.3 Le compromis biais - variance.

Sous le modèle $\mathbb{P}_{f,n} = \prod_1^n f(x_i)dx_i$, on considère un estimateur de la densité inconnue f associé à un noyau K :

$$\widehat{f}_{n,h}(t) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{X_k-t}{h}\right). \tag{2.7}$$

(pour simplifier les formules, le vecteur d'observation ne figure plus explicitement dans le membre de gauche). Le risque quadratique $\mathrm{MSE}_f(t)$ (cf. 1.3) associé à $\widehat{f}_{n,h}$ se décompose, comme dans la théorie paramétrique en somme de deux termes : la variance et le carré du biais. A la différence que tous les termes évoqués sont ici dépendants de la variable t :

$$\begin{aligned}\mathrm{MSE}_f(t_0) &= \mathbb{E}_f[(\widehat{f}_{n,h}(t_0) - f(t_0))^2] \\ &= \mathrm{Var}_f(\widehat{f}_{n,h}(t_0)) \;+\; \Big[\mathbb{E}_f(\widehat{f}_{n,h}(t_0)) - f(t_0)\Big]^2 = \sigma_f^2(t_0) \;+\; b_f^2(t_0)\end{aligned} \tag{2.8}$$

Les v.a. X_i étant équi-réparties, on trouve immédiatement que sous f (i.e. à supposer que la densité commune des X_i soit f), le biais s'écrit :

$$b_f(t_0) = \mathbb{E}_f[\widehat{f}_{n,h}(t_0)] - f(t_0) = \frac{1}{h}\int_{\mathbb{R}} f(t)\,K\left(\frac{t-t_0}{h}\right)dt - f(t_0) \tag{2.9}$$

Par ailleurs, les X_i étant indépendantes, on a pour la variance de $\widehat{f}_{n,h}(t_0)$:

$$\mathrm{Var}_f[\widehat{f}_{n,h}(t_0)] = \frac{1}{nh^2}\mathrm{Var}_f[K((X_1-t_0)/h)] \leq \frac{1}{nh^2}\mathbb{E}_f\left[K^2\left(\frac{X_i-t_0}{h}\right)\right] \leq \frac{1}{nh}\,(\sup_t f(t))\int (K(t))^2\,dt. \quad (2.10)$$

Ceci appelle déjà une première remarque. Supposons que le noyau K soit de carré sommable et que notre connaissance *a priori* du modèle d'observation nous permette de réduire la classe des densités admissibles de telle sorte qu'elles soient toutes uniformement bornées par une même constante finie. Il suffira alors de construire la suite des $\widehat{f}_{n,h}$ de manière à avoir $nh \to \infty$, pour que $\mathrm{Var}_f[\widehat{f}_{n,h}(t_0)] \to 0$ pour tout t_0. D'un autre coté, le biais a tendance à se conduire de façon "concurrente" à la variance : d'après la proposition 1.3.5,

$$|b_f(t_0)| \leq \frac{LC_\alpha}{l!}h^\alpha, \quad f \in \Lambda(\alpha, L).$$

pour peu que K soit un noyau d'ordre suffisemment élevé. Ceci conduit à construire la suite des $\widehat{f}_{n,h}$ avec $h \to 0$ pour s'assurer que $b_f(x) \to 0$ lorsque n croit. Dans de tels cas, la construction de bons estimateurs à noyau nécessite de réaliser un *compromis biais - variance* : faire tendre h vers zéro quand $n \to \infty$ (pour le biais), mais ne pas le faire décroitre trop rapidement, afin de préserver la condition $nh \to \infty$ (pour la variance).

Nous pouvons maintenant préciser les données du *compromis biais-variance* dans les hypothèses de la Proposition 1.3.5. La majoration du biais, sous ces conditions, obéit à l'inégalité (1.10). Cela nous donne *une majoration du biais uniforme en t_0 et en f*, puisque le membre de droite dans (1.10) ne dépend plus ni de t_0, ni de f. Pour la variance, nous avons la majoration (2.10) dans laquelle le membre de droite ne dépend plus de t_0, mais il dépend de f par l'intermédiaire de la constante $\|f\|_\infty = \sup_t f(t)$. Récapitulons la majoration du MSE obtenu.

Proposition 2.3.1
Soit $\alpha, L > 0$ et K un noyau d'ordre k, où k est l'entier tel que $\alpha - k \in (0, 1]$). Supposons que

$$C_\alpha = \int |t|^\alpha |K(t)|\,dt < \infty.$$

Alors

$$\sup_{f\in\mathcal{C}(\alpha,L_1,L_2)}\sup_{t_0\in\mathbb{R}}\mathbb{E}_f[(\widehat{f}_{n,h}(t_0)-f(t_0))^2] \leq \frac{L_1\|K\|_2^2}{nh} + \left(\frac{L_2C_\alpha}{k!}\right)^2 h^{2\alpha} \quad (2.11)$$

où $\mathcal{C}(\alpha, L_1, L_2) = \{f; 0 \leq f \leq L_1, \int f = 1\} \cap \Lambda(\alpha, L_2)$.

La majoration de (2.11) étant uniforme sur $\mathcal{C}(\alpha, L_1, L_2)$, on peut construire un estimateur $\widehat{f}_{n,h}$ dont l'erreur $\sup_{t_0 \in \mathbb{R}} \mathbb{E}_f[(\widehat{f}_{n,h}(t_0) - f(t_0))^2]$ est uniformément majoré sur cette classe tout simplement en minimisant, pour tout n, la majoration en h. On obtient alors une solution de la forme

$$h_n = c\, n^{-\frac{1}{2\alpha+1}}$$

où c ne dépend que des constantes L_1, L_2 et du noyau K.

On peut de plus montrer que *toutes les densités de la classe $\Lambda(\alpha, L)$ sont uniformement bornées par une même constante finie*, i.e.

Lemme 2.3.2
Pour tout $\alpha, L > 0$, il existe $L'(\alpha, L)$ tel que, pour tout $f \in \Lambda(\alpha, L)$ satisfaisant $f \geq 0$ et $\int f = 1$,

$$\sup_t f(t) \leq L'(\alpha, L).$$

Il s'ensuit le résultat fondmental suivant

Théorème 2.3.3
Soit $\alpha, L > 0$, $h_n = cn^{-\frac{1}{1+2\alpha}}$ et supposons sur K les mêmes hypothèses que pour la proposition 2.3.1. Alors

$$\sup_{f \in \mathcal{C}(\alpha,L)} \sup_{t_0 \in \mathbb{R}} \mathbb{E}_f[(\widehat{f}_{n,h_n}(t_0) - f(t_0))^2] \leq M\, n^{-\frac{2\alpha}{1+2\alpha}}. \tag{2.12}$$

où, cette fois, $\mathcal{C}(\alpha, L) = \Lambda(\alpha, L) \cap \{f; f \geq 0, \int f = 1\}$, et où $M = M(\alpha, L, K, c)$ est une constante qui ne dépend que de α, L, K et c.

On trouve ainsi que, uniformement en t_0 et f, le risque quadratique associé à la suite d'estimateurs $\widehat{f}_{n,h_n}$ tend vers zéro à la vitesse de $n^{-\frac{2\alpha}{1+2\alpha}}$. On peut alors se poser la question de savoir si cette vitesse peut être améliorée. Or nous avons le résultat qui suit.

Notons $R_n^\star(\mathcal{C}(\alpha, L))$ le *risque minimax* associé à la classe $\mathcal{C}(\alpha, L)$, défini ainsi :

$$R_n^\star(\mathcal{C}(\alpha, L)) = \inf_{t_0} \inf_{\tilde{f} \in \mathcal{S}_n} \sup_{f \in \mathcal{C}(\alpha,L)} \mathbb{E}_f[(\tilde{f} - f(t_0))^2] \tag{2.13}$$

où $\mathcal{S}_n$ est l'ensemble de tous les estimateurs basés sur n observations, c'est-à-dire de la forme $\tilde{f}(X_1, \cdots X_n)$. Alors, on peut montrer qu'il existe une constante $m = m(\alpha, L)$ telle que, pour tout $t_0 \in \mathbb{R}$:

$$R_n^\star(\mathcal{C}(\alpha, L)) \geq m\, n^{-\frac{2\alpha}{1+2\alpha}}, \tag{2.14}$$

On dira que $\widehat{f}_{n,h_n}$ converge à la *vitesse du minimax* dans la classe $\mathcal{C}(\alpha, L)$.

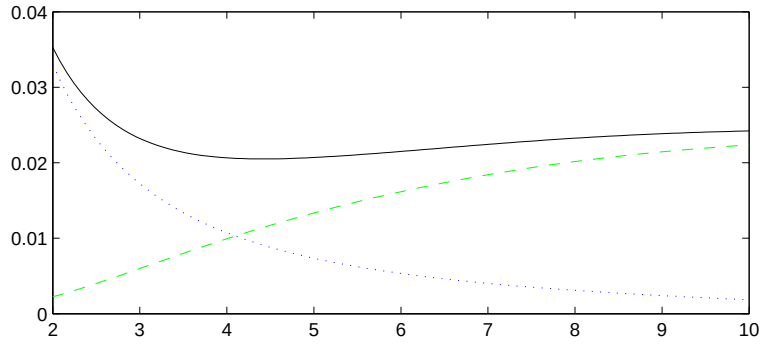

FIGURE 2.1 – Compromis biais variance : en abscisse est représenté le paramètre d'approximation h. Le carré du biais est en vert interrompu, la variance en trait bleu pointillé, l'erreur moyenne quadratique en trait noir plein. Le compromis consiste à choisir h au minimum de l'erreur quadratique moyenne.

2.4 Principe de l'estimateur sans biais du risque par validation croisée

L'inégalité (2.11) permet de déterminer une fenêtre h_n qui confère à un estimateur à noyau donné une certaine propriété d'optimalité. Toutefois, cela nécessite, d'une part, que les conditions de la Proposition 2.3.1 soient satisfaites (avec α connu par le statisticien) et, d'autre part, cela conduit à une fenêtre dont l'optimalité n'est qu'approchée : on a minimisé non-pas le risque quadratique, mais un majorant de celui-ci. Pour une suite d'estimateurs $\{\widehat{f}_{h,n}\}$ associés à un noyau donné et pour un risque quadratique R_Q fixé, l'estimateur optimal $\widehat{f}_{h^\star,n}$ de f correspondrait au choix de la fenêtre :

$$h^\star = \operatorname*{argmin}_h R_Q(f, \widehat{f}_{h,n}) \tag{2.15}$$

Examinons cette formule dans le cas des deux risques quadratiques que nous connaissons :

$$R_1(f,n,h,t_0) = \mathrm{MSE}_f(t_0) = \mathbb{E}_f[(\widehat{f}_{n,h}(t_0,X_1,\cdots,X_n) - f(t_0))^2] \tag{2.16}$$

$$R_2(f,n,h) = \mathrm{MISE}(f,\widehat{f}_{n,h}) = \int_{\mathbb{R}} \mathrm{MSE}_f(t)dt \tag{2.17}$$

Dans le cas (2.16), la solution de (2.15) dépendra de f et de t_0. Dans le cas de (2.17), elle dépendra de f. C'est cette dépendance de f qui rend impraticable l'estimation (de f) au moyen de $\widehat{f}_{h^\star,n}$ avec $h^\star$ défini par (2.15). On appelle **oracle** un estimateur (ou plutôt : pseudo-estimateur) qui, comme

$\widehat{f}_{h^\star,n}$, bénéficie d'une propriéte d'optimalité mais dont la dépendance relativement à l'inconnue f le prive de toute utilité pratique pour l'estimation de ce paramètre.

Il découle de ce qui précède que la construction de $\widehat{f}_{h^\star,n}$ à partir de (2.15) doit être modifiée d'une manière qui l'affranchisse de la dépence relativement au paramètre inconnu. Le **principe de l'estimateur sans biais du risque** répond à ce besoin. Il consiste à chercher un estimateur sans biais $\widehat{R_Q}(n, h, t_0, X_1, \cdots, X_n)$ de la fonction $R_Q(f, \widehat{f}_{n,h})$ du paramètre inconnu f et à appliquer la méthode qui précède à $\widehat{R_Q}$ au lieu de l'appliquer à R_Q. Cela revient donc à calculer la suite d'estimateurs $\widehat{f}_{\widehat{h}_n,n}$, avec :

$$\widehat{h}_n = \underset{h}{\operatorname{argmin}}\, \widehat{R_Q}(n, h, t_0, X_1, \cdots, X_n) \tag{2.18}$$

La fenêtre $\widehat{h}_n$ pourra dépendre de t_0 (c'est le cas quand $R_Q = \text{MSE}$) : dans ce cas, elle devra être ajustée à chaque valeur de la variable t en laquelle on cherchera à estimer $f(t)$. Dans tous les cas, cette fenêtre dépend des observations : sa forme générale est celle d'une *fenêtre aléatoire* $\widehat{h}_n(t_0, X_1, \cdots, X_n)$. La méthode peut être justifiée dans certains cas, notamment lorsque l'estimateur $\widehat{R_Q}$ possède de bonnes propriétés asymptotiques, de sorte que l'on puisse en déduire une relation de la forme :

$$R_Q(f, \widehat{f}_{\widehat{h}_n,n}) = R_Q(f, \widehat{f}_{h^\star,n})\,(1 + O_p(1/n)) \tag{2.19}$$

On peut montrer que ceci est bien le cas lorsqu'on applique la démarche décrite ci-dessus pour $R_Q = \text{MISE}$. Dans ce cas, pour construire un estimateur sans biais du risque, on appliquera le **principe de validation croisée** dont voici une description heuristique.

Soit $\widehat{f}_{n,h}$ un estimateur à noyau d'une densité inconnue f telle que $\int (f(t))^2\, dt < \infty$. En notant dans ce qui suit X le vecteur d'observation ($X = (X_1, \cdots, X_n)$), nous avons :

$$\begin{aligned}
\text{MISE}(f, \widehat{f}_{n,h}) &= \mathbb{E}_f\left[\int (\widehat{f}_{n,h}(t, X) - f(t))^2\, dt\right] \\
&= \mathbb{E}_f\left[\int (\widehat{f}_{n,h}(t, X))^2\, dt - 2\int \widehat{f}_{n,h}(t, X) f(t)\, dt\right] \\
&\qquad + \int (f(t))^2\, dt \qquad (2.20) \\
&= J(f, \widehat{f}_{n,h}) + \int (f(t))^2\, dt
\end{aligned}$$

La minimisation de cette expression en h revient à celle de $J(f, \widehat{f}_{n,h})$, l'intégrale de $f^2(t)$ étant indépendante de h. Le principe de l'estimateur

sans biais du risque nous conduit donc à rechercher un estimateur sans biais de $J(f, \widehat{f}_{n,h})$ qui sera à minimiser ensuite en h.

De façon évidente, la fonction de f définie par $\mathbb{E}_f\left[\int(\widehat{f}_{n,h}(t, X_1, \cdots, X_n))^2\, dt\right]$ admet un estimateur sans biais de la forme

$$U(X_1, \cdots, X_n) = \int [\widehat{f}_{n,h}(t, X_1, \cdots, X_n)]^2\, dt.$$

Reste à trouver un estimateur sans biais de la fonction de f définie par $\mathbb{E}_f\left[\int \widehat{f}_{n,h}(t, X_1, \cdots, X_n) f(t)\, dt\right]$. Nous avons :

$$\begin{aligned} \mathbb{E}_f\left[\int \widehat{f}_{n,h}(t, X_1, \cdots, X_n) f(t)\, dt\right] &= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{E}_f\left[\int_{\mathbb{R}} K\left(\frac{X_i - t}{h}\right) f(t)\, dt\right] = \\ &= \frac{1}{h}\int_{\mathbb{R}} f(y)\left[\int_{\mathbb{R}} K\left(\frac{y-t}{h}\right) f(t)\, dt\right] dy \end{aligned} \tag{2.21}$$

ce qui coincide avec l'espérance sous f de l'estimateur :

$$\widehat{f}_{n,-i}(X) = \frac{1}{(n-1)h}\sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right) \tag{2.22}$$

ainsi qu'on le vérifie immédiatement, puisque pour $j \neq i$ on a bien :

$$\mathbb{E}_f\left[K\left(\frac{X_j - X_i}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}^2} K\left(\frac{u-v}{h}\right) f(u) f(v) du dv \tag{2.23}$$

Il s'ensuit que $\widehat{f}_{n,-i}(X)$ est un estimateur sans biais de

$$\mathbb{E}_f\left[\int \widehat{f}_{n,h}(t, X_1, \cdots, X_n) f(t)\, dt\right]$$

et il en sera de même de l'estimateur $\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{n,-i}(X)$. Le *principe de validation croisée* que l'on a appliqué pour construire cet estimateur peut s'énoncer de la façon suivante :

Principe de validation croisée : *on a construit l'estimateur $\widehat{f}_{n,-i}$ à partir de l'estimateur de départ $\widehat{f}_{n,h}(t, X_1, ..., X_n)$ en éliminant dans ce dernier la variable X_i du vecteur d'observation $(X_1, ..., X_n)$ et en l'introduisant à la place de la variable t. Puis on construit la moyenne des n estimateurs obtenus en appliquant la méthode pour chaque i.*

Proposition 2.4.1 (Validation croisée pour estimateurs à noyau) *Supposons que K soit un noyau borné. On note $\widehat{f}_{n,h}(t, X)$ l'estimateur à noyau associé à K pour une fenêtre h et pour n observations et $\widehat{f}_{n,-i}(X)$ l'estimateur défini par (2.22). On pose :*

$$\mathrm{CV}(h, X) = \int [\widehat{f}_{n,h}(t, X)]^2 \, dt - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{n,-i}(X) \tag{2.24}$$

Alors, pour tout $h > 0$ *:*

$$\mathbb{E}_f[\mathrm{CV}(h, X)] = \mathbb{E}_f \left[\int (\widehat{f}_{n,h}(t, X) - f(t))^2 \, dt \right] - \int [f(t)]^2 \, dt < \infty.$$

A partir de cette donnée, on ne peut construire qu'un pseudo-estimateur (car dépendant de f) sans biais de $\mathrm{MISE}(f, \widehat{f}_{n,h})$, à savoir $\widehat{\mathrm{MISE}}(h, X) = \mathrm{CV}(h, X) + \int [f(t)]^2 \, dt$. L'important étant toutefois que sa minimisation en h, qui revient à la minimisation en h de l'estimateur $\mathrm{CV}(h, X)$, donne bien un résultat indépendant de f. On définira donc un estimateur **adaptatif** $\widehat{f}_{\widehat{h}_n, n}$ à partir de la suite $\widehat{f}_{n,h}$ répondant à la condition :

$$\widehat{h}_n(X) = \operatorname*{argmin}_{h>0} \mathrm{CV}(h, X) = \operatorname*{argmin}_{h>0} \widehat{\mathrm{MISE}}(h, X) \tag{2.25}$$

Ainsi qu'on l'a déjà signalé, on peut montrer que le risque quadratique intégré (MISE) de l'estimateur $\widehat{f}_{\widehat{h}_n, n}$ est asymptotiquement équivalent à celui du pseudo-estimateur (oracle) optimal $\widehat{f}_{h^\star, n}$ défini par (2.15).

2.1 Exemple (Estimation de densité par noyaux et validation croisée): La démarche de validation croisée pour l'estimation sans biais du risque est illustrée par la figure 2.2. ♢

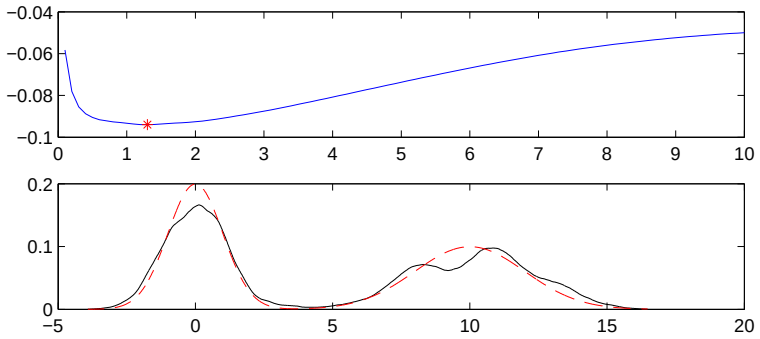

FIGURE 2.2 – On reprend les observations de la figure 1.1 (200 observations i.i.d. d'un mélange de deux gaussiennes). Figure du haut : $\mathrm{CV}(h)$ en fonction de h. Figure du bas : l'estimateur à noyau du h minimisant $\mathrm{CV}(h)$ est représenté en noir plein, la vraie densité étant en rouge pointillé.

2.5 Le modèle du bruit additif

Pour introduire le modèle général de la régression, rappelons que lorsque (Y, Z) est un vecteur aléatoire à valeurs dans $\mathbb{R}^2$, le meilleur prédicteur de Z au moyen de Y au sens de la distance de l'espace L^2, est donné par $\psi(Y)$, où $\psi(y) = \mathbb{E}[Z|Y = y]$ est définie de façon unique par la condition :

$$\mathbb{E}[\psi(Y)g(Y)] = \mathbb{E}[Zg(Y)] \tag{2.26}$$

relation qui doit valoir pour toute $g : \mathbb{R} \longrightarrow \mathbb{R}$ borélienne bornée. Si, pour une suite d'observations i.i.d. $(Y_1, Z_1), \cdots, (Y_n, Z_n)$, on pose un modèle de régression de la forme :

$$Z_i = f(Y_i) + \epsilon_i \quad i = 1, \ldots, n \tag{2.27}$$

avec, pour tout i, ϵ_i indépendante de Y_i et $\mathbb{E}[\epsilon_i] = 0$ (modèle de bruit additif), alors on dit que f est *la fonction de régression des Z_i en Y_i* et on doit avoir $f(y) = \mathbb{E}[Z_i|Y_i = y]$. En effet, la fonction f dans (2.27) satisfait (2.26), puisque pour tout g bornée, on a :

$$\mathbb{E}[f(Y_i)g(Y_i)] = \mathbb{E}[(Z_i - \epsilon_i)g(Y_i)] = \mathbb{E}[Z_i g(Y_i)] \tag{2.28}$$

étant donné que ϵ_i est indépendante de Y_i et $\mathbb{E}[\epsilon_i] = 0$. Le problème statistique non-paramétrique étudié ici sera celui de l'estimation de la fonction de régression f considérée comme inconnue. C'est donc le problème de l'estimation de la meilleure approximation de Z_i par une fonction de Y_i.

2.6 Estimateur de Nadaraya-Watson

Les estimateurs les plus connus, pour le problème d'estimation de f dans (2.27), sont les estimateurs de Nadaraya-Watson. Leur définition repose sur le même type d'approximation que les estimateurs à noyau pour la densité.

Définition 2.6.1 (estimateur de Nadaraya-Watson)
On considère un modèle de régression non-linéaire (2.27) et on donne un noyau $K(u)$. Notons $X = ((Y_1, Z_1), \cdots, (Y_n, Z_n))$. On appelle estimateur de Nadaraya-Watson de f *associé au noyau K (pour n observations et pour une fenêtre h fixée), l'estimateur :*

$$\widehat{f}_{(h,n)}(y, X) = \begin{cases} \dfrac{\sum_{i=1}^n Z_i K\left(\frac{Y_i - y}{h}\right)}{\sum_{j=1}^n K\left(\frac{Y_i - y}{h}\right)} & si \ \sum_{j=1}^n K\left(\frac{Y_i - y}{h}\right) \neq 0 \\ 0 & sinon \end{cases} \tag{2.29}$$

Motivation. Si les (Y_i, Z_i) possèdent une densité $p(y, z)$, alors nous savons que :

$$f(y) = \mathbb{E}[Z|Y = y] = \int_{\mathbb{R}} z\, p_{Z|Y}(z|y)\, dz$$

où $p_{Z|Y}(z|y)$ est la densité conditionnelle $\frac{p(y,z)}{p(y)}$, $p(y)$ étant la densité des Y_i (avec la convention habituelle $0/0 = 0$). Si l'on remplace dans le terme de droite de cette formule les densités $p(y, z)$ et $p(y)$ par leurs estimées au moyen des estimateurs à noyaux respectif $\widehat{p}_{n,h}(y, z, (Y_1, Z_1), \cdots, (Y_n, Z_n))$ et $\widehat{p}_{n,h}(y, Y_1, \cdots, Y_n)$ (cf. (2.6) et (2.1)), on obtient :

$$\begin{aligned} & \frac{1}{h\sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right)} \int_{\mathbb{R}} z \sum_{i=1}^n K\left(\frac{Z_i - z}{h}\right) K\left(\frac{Y_i - y}{h}\right) dz \; = \\ = \quad & \frac{1}{h\sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right)} \sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right) \int_{\mathbb{R}} zK\left(\frac{Z_i - z}{h}\right) dz \end{aligned}$$

et on a :

$$\int_{\mathbb{R}} zK\left(\frac{Z_i - z}{h}\right) dz = h \int_{\mathbb{R}} (Z_i - hu)K(u)du = Z_i$$

dès l'instant que $K(u)$ est un noyau d'ordre 1, c'est-à-dire $\int uK(u)du = 0$ (et, par définition d'un noyau, on a toujours $\int K(u)du = 1$). Par conséquent :

Théorème 2.6.2
Soit K un noyau d'ordre 1 et $\widehat{p}_{n,h}(y, z)$, $\widehat{p}_{n,h}(y)$ les estimateurs à noyau respectif de la densité $p(y, z)$ des (Y_i, Z_i) et $p(y)$ des Y_i associés à K. Alors on a :

$$\widehat{f}_{(h,n)}(y, (Y_1, Z_1), \cdots, (Y_n, Z_n)) = \int_{\mathbb{R}} z \frac{\widehat{p}_{n,h}(y, z, (Y_1, Z_1), \cdots, (Y_n, Z_n))}{\widehat{p}_{n,h}(y, Y_1, \cdots, Y_n)} dz$$

Cas où la loi $\mathbb{P}_Y$ est connue. Si la densité $p(y)$ de Y est connue, on modifiera l'expression de l'estimateur de Nadaraya - Watson de la manière suivante :

$$\widehat{f}_{h,n}(y) = \frac{1}{nhp(y)} \sum_{i=1}^n Z_i K((Y_i - y)/h) = \int z \frac{\widehat{p}_{h,n}(y, z)}{p(y)}\, dz \tag{2.30}$$

la deuxième identité ayant lieu dans les conditions du Théorème 2.6.2.

2.2 Exemple (Variable explicative uniformes):
Si Y est de loi uniforme sur$[0,1]$, sa densité est $p(y) = \mathbb{1}_{[0,1]}(y)$. Soit σ^2 la variance du bruit blanc $\{\epsilon_i\}$. On trouve alors :

$$\begin{aligned}
\text{biais}_\theta(\widehat{f}_{h,n}(y)) &= \frac{1}{h}\int_0^1 f(u)K((u-y)/h)\,du - f(y),\\
\text{Var}_\theta(\widehat{f}_{h,n}(y)) &= \frac{1}{nh^2}\left[\text{Var}(f(Y_1)K((Y_1-y)/h)) + \sigma^2\int_0^1 (K((u-y)/h))^2\,du\right]\\
&\leq \frac{1}{nh}\left[\sup_t(f(t))^2+\sigma^2\right]\left(\int (K(u))^2\,du\right).
\end{aligned}$$

où σ^2 est la variance de $\{\epsilon_i\}$. ♢

2.7 Estimateurs de régression par polynômes locaux

Pour un noyau K positif, l'estimateur de Nadaraya-Watson satisfait la condition :

$$\widehat{f}_{(h,n)}(y,X) = \underset{z\in\mathbb{R}}{\text{argmin}} \sum_{k=1}^n (Z_k - z)^2\; K\left(\frac{Y_k - y}{h}\right)$$

(vérification immédiate). Cela suggére une généralisation de la méthode en définissant un estimateur à valeurs vectorielles :

$$\widehat{\mathbf{f}}_l(y) = \underset{\mathbf{z}\in\mathbb{R}^{l+1}}{\text{argmin}} \sum_{k=1}^n \left(Z_k - \mathbf{z}^T\mathbf{u}(Y_k-y)\right)^2\; K\left(\frac{Y_k-y}{h}\right) \tag{2.31}$$

où $\mathbf{u}(t) = \left[1, t, \ldots, t^l/l!\right]^T$. En effet, $f(Y_k)$ est la meilleure approximation de Z_k en moyenne quadratique et on peut écrire par ailleurs :

$$\begin{aligned}
f\,(Y_k) &= f(y) + (Y_k - y)\; f'(y) + \ldots + \frac{(Y_k-y)^l}{l!}\, f^{(l)}(y) + R_l(Y_k,y) =\\
&= \mathbf{f}_l(y)^T\mathbf{u}(Y_k-y) + R_l(Y_k,y)
\end{aligned}$$

où $\mathbf{f}_l(y) = [f(y), f'(y), \ldots, f^{(l)}(y)]^T$ et $R_l(Y_k,y) = O(|Y_k-y|^\beta)$ dès que $f \in \Lambda(\beta, L)$ avec $l < \beta \leq l+1$. Il est donc assez naturel de se demander dans quelle mesure $\widehat{\mathbf{f}}_l(y)$ dans (2.31) peut être un bon estimateur de $\mathbf{f}_l(y)$. On remarquera que la première coordonnée de $\widehat{\mathbf{f}}_l(y)$ apparaît comme un nouvel estimateur de $f(y)$. L'estimateur $\widehat{\mathbf{f}}_l$ s'appelle *estimateur par polynômes locaux d'ordre l*. Compte-tenu de la façon dont il a été défini, on peut s'attendre à ce que, sous des conditions appropriées, les performances de l'estimateur par polynômes locaux s'améliorent lorsque l augmente. Nous nous limiterons ici aux questions concernant la détermination de $\widehat{\mathbf{f}}_l$.

Détermination de $\widehat{\mathbf{f}}_l$. C'est un problème de moindres carrés pondérés, avec les $K\left(\frac{Y_k-y}{h}\right)$ comme coefficients de pondération. Posons :

$$\begin{aligned}
&\mathbf{Z}_n = [Z_1, \ldots, Z_n]^T \\
&\Delta_{h,n}(y) = \text{Diag}\left([K((Y_1-y)/h), \ldots, K((Y_n-y)/h)]\right) \\
&\Delta_{h,n}^{1/2}(y) = \text{Diag}\left([K((Y_1-y)/h)^{1/2}, \ldots, K((Y_n-y)/h)^{1/2}]\right) \\
&\mathbf{U}_n(y) = [\mathbf{u}(Y_1-y), \ldots, \mathbf{u}(Y_n-y)]
\end{aligned}$$

La matrice $\Delta_{h,n}(y)$ est diagonale $n \times n$, avec la diagonale formée des $K((Y_i - y)/h)$, la matrice $\mathbf{U}_n(y)$ est $(l+1) \times n$ et sa k-ème colonne coincide avec $\mathbf{u}(Y_k - y)$. Enfin, pour pouvoir parler de la matrice $\Delta_{h,n}^{1/2}(y)$, on doit avoir un noyau K positif.

La solution $\widehat{\mathbf{f}}_l(y)$ de (2.31) répond à la condition :

$$\widehat{\mathbf{f}}_l(y) = \underset{\mathbf{z}\in\mathbb{R}^{l+1}}{\text{argmin}} \left\| \Delta_{h,n}^{1/2}(y).\mathbf{Z}_n - \Delta_{h,n}^{1/2}(y).\mathbf{U}_n^T(y).\mathbf{z} \right\|^2$$

autrement-dit : $\Delta_{h,n}^{1/2}(y).\mathbf{U}_n^T(y).\widehat{\mathbf{f}}_l(y)$ est la projection orthogonale (dans l'espace $\mathbb{R}^n$) de $\Delta_{h,n}^{1/2}(y).\mathbf{Z}_n$ sur l'espace $Im\left(\Delta_{h,n}^{1/2}(y).\mathbf{U}_n^T(y)\right)$. Le vecteur

$$\Delta_{h,n}^{1/2}(y)\left(\mathbf{Z}_n - \mathbf{U}_n^T(y).\widehat{\mathbf{f}}_l(y)\right)$$

doit donc être orthogonal à $Im\left(\Delta_{h,n}^{1/2}(y).\mathbf{U}_n^T(y)\right)$ ou, ce qui est équivalent, il doit être dans le noyau de la transposée de $\Delta_{h,n}^{1/2}(y).\mathbf{U}_n^T(y)$:

$$\left[\Delta_{h,n}^{1/2}(y).\mathbf{U}_n^T(y)\right]^T \Delta_{h,n}^{1/2}(y)\left(\mathbf{Z}_n - \mathbf{U}_n^T(y).\widehat{\mathbf{f}}_l(y)\right) \;=\; 0$$

c'est-à-dire :

$$\mathbf{U}_n(y)\Delta_{h,n}(y)\mathbf{Z}_n = \mathbf{U}_n(y)\Delta_{h,n}(y)\mathbf{U}_n^T(y).\widehat{\mathbf{f}}_l(y)$$

(équation normale). Si la matrice $\mathbf{U}_n(y)\Delta_{h,n}(y)\mathbf{U}_n^T(y)$ est définie positive, alors :

$$\widehat{\mathbf{f}}_l(y) = \left(\mathbf{U}_n(y)\Delta_{h,n}(y)\mathbf{U}_n(y)^T\right)^{-1}\mathbf{U}_n(y)\Delta_{h,n}(y)\mathbf{Z}_n \tag{2.32}$$

On peut en fait montrer le résultat suivant (voir Tsybakov [2004]).

Théorème 2.7.1
Soit $\alpha, L > 0$ et $\widehat{f}_{(l,h,n)}$ l'estimateur défini comme la première coordonée de (2.31) avec $l = [\alpha]$. On suppose que le modèle (2.27) vérifie

(a) $Y_1, \ldots, Y_n$ *sont déterministes dans* $[0,1]$ *et vérifie la propriété de non-concentration suivante : il existe* $a_0 > 0$ *tel que pour tout intervalle* $A \subset [0,1]$ *et tout entier* n*,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i \in A) \leq a_0 \max(|A|, A/n)$$

(b) La matrice

$$B_{n,x} = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{u}((Y_k - y)/h)\mathbf{u}^T((Y_k - y)/h)K((Y_k - y)/h)$$

a toutes ses valeurs propres supérieures à $\lambda_0 > 0$ *pour* n *suffisemment grand et pour tout* $x \in [0,1]$*.*

(c) $\epsilon_1, \ldots, \epsilon_n$ *sont indépendantes, centrées de variances minorées par* σ^2*.*

Et l'on choisi un noyau K *à support dans* $[-1,1]$ *et borné par* $K_{\max} < \infty$*.*

Alors pour $h_n = cn^{\frac{-1}{1+2\alpha}}$*, on a*

$$\sup_{f \in \Lambda(\alpha, L)} \sup_{t_0 \in [0,1]} \mathbb{E}_f[(\widehat{f}_{n,h_n}(t_0) - f(t_0))^2] \leq M\, n^{-\frac{2\alpha}{1+2\alpha}}.$$

où $M = M(\alpha, L, \lambda_0, \alpha_0, K_{\max}, c)$ *est une constante qui ne dépend que de* $\alpha, L, \lambda_0, \alpha_0, K_{\max}$ *et* c*.*

Ce théorème s"applique en particulier pour $Y_i = i/n$ et K positif borné à support dans $[-1,1]$ et minoré sur $[-\epsilon, \epsilon]$ par une constante strictement positive.

2.3 Exemple (Housing price de Boston):
Sur la figure 2.3, estimateurs par polynômes locaux et estimateur simple de Nadaraya-Watson sont comparés. ♢

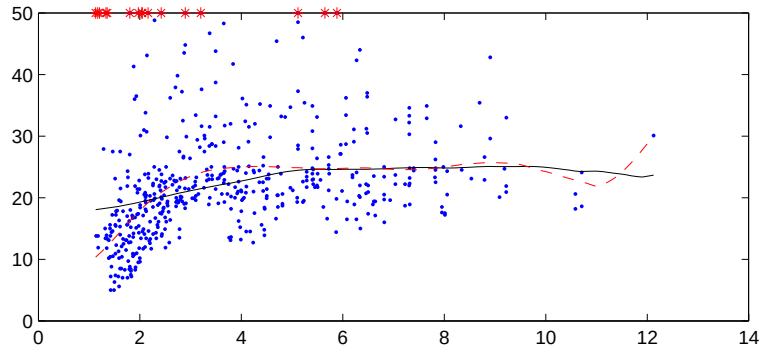

FIGURE 2.3 – Estimation de la régression pour les données Housing price de Boston. Estimateur Nadaraya-Watson (trait noir plein) et estimateur par polynômes locaux d'ordre 2 (trait rouge pointillé), fenêtre triangulaire de support $[-3, 3]$. Les points $*$rouges ne sont pas pris en compte.

Chapitre 3

Estimation par projection

3.1 Cadre hilbertien

Etant donné un espace de Hilbert $(H, \langle \cdot, \cdot \rangle)$ dont on note $\langle \cdot, \cdot \rangle$ le produit scalaire, ainsi qu'une base orthonormée de cet espace (base hilbertienne) $\{g_k,\, k \geq 0\}$, nous savons que tout élément f de H se décompose sur cette base en sa série de la manière suivante :

$$f = \sum_{k=0}^{\infty} \langle f, g_k \rangle \, g_k.$$

On appelle alors *projection d'ordre J de f* le terme $f_J := \sum_{k=0}^{J} \langle f, g_k \rangle \, g_k$ et *l'erreur d'approximation de f à l'ordre J* est représentée évidemment par

$$\|f - f_J\|_H^2 = \sum_{k>J} |\langle f, g_k \rangle|^2.$$

Le lemme suivant résume le comportement asymptotique de l'approximation selon une base hilbertienne :

Lemme 3.1.1
Sous les hypothèses et définitions ci-dessus, on a

a *Soit $0 < \rho < 1$. Alors $|\langle f, g_k \rangle| = O(\rho^k) \Leftrightarrow ||f - f_J|| = O(\rho^J)$.*

b *Soit $\alpha > 1$. Alors $|\langle f, g_k \rangle| = O(k^{-1-\alpha}) \Rightarrow ||f - f_J|| = O(J^{-\alpha})$ et la réciproque est fausse.*

c *Soit $s > 0$. Alors $\sum_{k \geq 0} k^{2s} |\langle f, g_k \rangle|^2 < \infty \Leftrightarrow \sum_{J \geq 1} J^{2s-1} ||f - f_J||^2 < \infty$.*

Démonstration elle est élémentaire, on se contentera de l'indiquer pour le premier point, le reste étant laissé à titre d'exercice. Si $0 \leq |\langle f, g_k \rangle| \leq B\rho^k$, alors :

$$\|f - f_J\|^2 = \sum_{k \geq J} |\langle f, g_k \rangle|^2 \leq B^2 \rho^{2J} \sum_{i=0}^{\infty} \rho^i = \tilde{B} \rho^{2J}$$

ce qui donne bien $||f - f_J|| = O(\rho^J)$. Inversement, si on a cette dernière relation , alors pour $k \geq J$, on aura $|\langle f, g_k \rangle| \leq ||f - f_J|| = O(\rho^J)$. ∎

Il ressort du lemme 3.1.1 que la bonne notion de régularité pour contrôler l'erreur de projection $||f - f_J||$ est de contôler la somme pondérée $\sum_{k \geq 0} k^{2s} |\langle f, g_k \rangle|^2$. C'est pour cette raison que nous analyserons le risque sous des hypothèses de régularité de type Sobolev.

3.2 Définition

Considérons, pour un modèle $\{\mathbb{P}_{n,f},\ f \in \mathcal{F}\}$ d'observation $(X_1, \cdots, X_n)$, le problème d'estimation de $\phi(f)$, où $\phi : \mathcal{F} \to H$ est une fonction à valeurs dans un espace de Hilbert H.

Supposons que l'on ait trouvé une fonction $\mathcal{C} : H \times \mathcal{M} \to \mathbb{R}_+$ ($\mathcal{M}$: ensemble des mesures de probabilité sur $\mathbb{R}$) permettant d'écrire :

$$\phi(f) = \operatorname*{argmin}_{g \in H} \mathcal{C}(g, \mathbb{P}_f) \qquad (\forall f \in \mathcal{F}) \tag{3.1}$$

Le coté droit de (3.1) dépend de f et il ne dépend pas des observations : deux raisons qui font que, tel quel, cette relation ne peut pas servir à construire un estimateur de $\phi(f)$. Supposons toutefois que l'on sache déterminer une suite $P_n = P^{(X_1,\ldots,X_n)}$ de probabilités sur $\mathbb{R}$ (P_n est fonction des observations $(X_1, ..., X_n)$ et elle ne dépend plus de f) telle que l'on ait *sous chaque* $\mathbb{P}_f$ *:* $\mathcal{C}(g, P_n) \to \mathcal{C}(g, P_f)$ quand $n \to \infty$. Alors, d'une part, la relation :

$$\widehat{\phi}_n(X_1, ..., X_n) = \operatorname*{argmin}_{g \in H} \mathcal{C}(g, P_n) \tag{3.2}$$

définit un estimateur (dépendance des observations, indépendance par rapportà f) et, d'autre part, la convergence $\mathcal{C}(g, P_n) \to \mathcal{C}(g, P_f)$ supposée ci-dessus permettra en général d'avoir une *suite consistante d'estimateurs* de $\phi(f)$, c'est-à-dire satisfaisant la condition : $\forall f, \widehat{\phi}_n(X_1, ..., X_n) \to \phi(f)$ sous $\mathbb{P}_f$, quand $n \to \infty$.

Il est fréquent de prendre $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ (cf. par exemple le cas de l'estimation d'une densité qui suit). Par ailleurs, pour atteindre $\widehat{\phi}_n$, il serait intéressant d'avoir une méthode permettant de faire le calcul du minimum dans (3.2) sur des espaces où ce calcul est facile à faire (p.ex., des espaces de dimension finie) puis de passer à la limite. Autrement-dit, étant donnée une suite croissante $\{V_m,\ m \geq 1\}$ de sous-espaces vectoriels de H telle que :

$$\overline{\bigcup_{m\geq 1} V_m} = H \tag{3.3}$$

on définira pour tout m et n, l'estimateur de projection de $\phi(f)$ sur V_m par :

$$\widehat{\phi}_{m,n}(X_1, ..., X_n) = \underset{g\in V_m}{\operatorname{argmin}}\, \mathcal{C}(g, P_n) \tag{3.4}$$

et la propriéte de consistance correspondant à cette nouvelle construction s'écrira : $\forall f, \widehat{\phi}_{m,n}(X_1, ..., X_n) \to \phi(f)$ sous $\mathbb{P}_f$, quand $m, n \to \infty$.

En fait comme pour le paramètres h des estimateurs de type noyau, il ne faudra pas faire tendre m vers l'infini à n fixé mais bien choisir m en fonction de n de tel sorte que $\widehat{\phi}_{m_n,n}$ converge vers $\phi(f)$ sous $\mathbb{P}_f$.

3.3 Estimateurs de projection pour la densité

On considère désormais le cas où $\phi(f) = f$ pour le modèle $\{\mathbb{P}_{n,f} = \mathbb{P}_f^n\}$ avec $\mathbb{P}_f(dx) = f(x)\,dx$, où f est la densité commune des X_i à estimer. On suppose que $f \in H = L^2(I), I = [0, 1]$. On écrit, pour tout $f \in H$:

$$f = \underset{g\in H}{\operatorname{argmin}} \|g - f\|_H^2 = \underset{g\in H}{\operatorname{argmin}} \left(\|g\|_H^2 - 2 \left(\int_I g(x)\, \mathbb{P}_f(dx) \right) \right)$$
$$=: \underset{g\in H}{\operatorname{argmin}}\, \mathcal{C}(g, \mathbb{P}_f)$$

Soit $\{g_k,\, k \in \mathbb{N}\}$ une base orthonormée de H et notons $V_m = \text{vect}(g_k,\, 0 \leq k \leq m)$ le sous-espace vectoriel de H engendré par les g_k dont les indices sont au plus m. Nous utiliserons ci-dessous des bases de Fourier mais dans bien des cas les bases d'ondelettes sont plus adaptées (voir Hastie et al. [2001]). Posons également $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ (mesure empirique associée aux observations $X_1, \ldots, X_n$). En poursuivant la méthode décrite ci-dessus, on calcule maintenant l'estimateur :

$$\begin{aligned} \widehat{f}_{m,n} &= \underset{g\in V_m}{\operatorname{argmin}}\, \mathcal{C}(g, P_n) = \underset{g\in V_m}{\operatorname{argmin}} \left(\|g\|_H^2 - 2 \left(\int_I g(x)\, P_n(dx) \right) \right) \\ &= \underset{g\in V_m}{\operatorname{argmin}} \left(\|g\|_H^2 - 2 \left(\frac{1}{n} \sum_{n=1}^n g(X_i) \right) \right) \end{aligned} \tag{3.5}$$

et on voit qu'on a bien $\mathcal{C}(g, P_n) \to \mathcal{C}(g, P_f)$ sous chaque $\mathbb{P}_f$. En effet, si f est la densité commune des X_i, alors en vertu de la loi des grands nombres, on a, presque sûrement,

$$\int_I g(x)\, P_n(dx) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \longrightarrow \mathbb{E}_f[g(X_i)] = \int_I g(x)\, \mathbb{P}_f(dx) \qquad (3.6)$$

Compte-tenu du choix de nos espaces V_m, le calcul des estimateurs $\widehat{f}_{m,n}(X_1, ..., X_n)$ est presque immédiat. Notons $\{a_k(g),\, k \in \mathbb{N}\}$ les coefficients d'une fonction $g \in H$, calculés selon la base (g_k). Dire que $g \in V_m$ équivaut à $g = \sum_{k=0}^{m} a_k(g) g_k$. Par conséquent, on a pour $g \in V_m$ (cf. aussi (3.5)) :

$$\begin{aligned}
\mathcal{C}(g, P_n) &= \sum_{k=0}^{m} |a_k(g)|^2 - 2\sum_{k=0}^{m} a_k(g) \left(\frac{1}{n}\sum_{i=1}^{n} g_k(X_i)\right) \\
&= \sum_{k=0}^{m} \left(|a_k(g)|^2 - 2\left[a_k(g).\frac{1}{n}\sum_{i=1}^{n} g_k(X_i)\right]\right) \\
&= \sum_{k=0}^{m} \left|a_k(g) - \frac{1}{n}\sum_{i=1}^{n} g_k(X_i)\right|^2 - \frac{1}{n}\sum_{k=0}^{m}\left|\sum_{i=1}^{n} g_k(X_i)\right|^2 .
\end{aligned}$$

Seul le premier terme de la différence ci-dessus dépend de g. La fonction $g \in V_m$ qui minimise ce terme correspond aux $a_k(g) = \frac{1}{n}\sum_{i=1}^{n} \overline{g_k}(X_i)$, donc :

$$\widehat{f}_{m,n}(t, X_1, ..., X_n) = \frac{1}{n}\sum_{k=0}^{m}\left(\sum_{j=1}^{n} g_k(X_j)\right) g_k(t) \qquad (3.7)$$

3.4 Borne supérieure du risque intégré

On se propose d'établir une majoration du risque intégré $\mathrm{MISE}_f(\widehat{f}_{m,n})$ pour l'estimateur défini par (3.7). Nous avons, en notant $b_f(\widehat{f}_{m,n}(t, X)) = \mathbb{E}_f[\widehat{f}_{m,n}(t, X)] - f(t)$ le biais de l'estimateur $\widehat{f}_{m,n}$:

$$\begin{aligned}
\mathrm{MISE}_f(\widehat{f}_{m,n}) = \int_I \mathbb{E}_f\left[\left|\widehat{f}_{m,n}(t, X_1, ..., X_n) - f(t)\right|^2\right] dt = \\
\int_I \mathrm{Var}_f(\widehat{f}_{m,n}(t, X_1, ..., X_n))dt \;+\; \int_I \left|b_f(\widehat{f}_{m,n}(t, X_1, ..., X_n))\right|^2 dt \qquad (3.8)
\end{aligned}$$

Etude de la variance : en inversant l'ordre de sommation dans (3.7) et compte-tenu du fait que les X_i sont i.i.d., on a :

$$\begin{aligned}\mathrm{Var}_f(\widehat{f}_{m,n}(t,X_1,...,X_n)) &= \frac{1}{n}\mathrm{Var}_f\left[\sum_{k=0}^{m} g_k(t)g_k(X_1)\right] \\ &\leq \frac{1}{n}\mathbb{E}_f\left[\left|\sum_{k=0}^{m} g_k(t)g_k(X_1)\right|^2\right].\end{aligned}$$

Il s'en suit, par orthonormalité de la base,

$$\begin{aligned}\int_I \mathrm{Var}_f(\widehat{f}_{m,n}(t,X_1,...,X_n)) &\leq \frac{1}{n}\mathbb{E}_f\int_I\left|\sum_{k=0}^{m} g_k(X_1)\,g_k(t)\right|^2 dt \\ &= \frac{1}{n}\mathbb{E}_f\sum_{k=0}^{m}|g_k(X_1)|^2 \\ &\leq \frac{m+1}{n}\,\|f\|_\infty. \qquad (3.9)\end{aligned}$$

Etude du biais :

$$\begin{aligned}b_f(\widehat{f}_{m,n}) &= \frac{1}{n}\sum_{k=-m}^{m} g_k(t)\sum_{i=1}^{n}\mathbb{E}_f\left[g_k(X_i)\right] - f(t) \\ &= \frac{1}{n}\sum_{k=-m}^{m} g_k(t)\sum_{i=1}^{n}\int_I g_k(u)f(u)du - f(t) \\ &= \sum_{k=-m}^{m}\langle f,g_k\rangle\, g_k(t) - f(t) \qquad (3.10) \\ &= P_{V_m}(f)(t) - f(t) \qquad (3.11)\end{aligned}$$

où $P_{V_m}(f)$ est la projection orthogonale de f sur le sous-espace V_m. Par orthonormalité de la base,

$$\int_I\left|b_f(\widehat{f}_{m,n})\right|^2 dt = \left\|\sum_{k=0}^{m}\langle f,g_k\rangle\, g_k - f\right\|_H^2 = \sum_{k>m}|\langle f,g_k\rangle|^2 \qquad (3.12)$$

D'après (3.12), le biais ne dépend que de m et il tend vers zéro pour $m \to \infty$ en tant que reste d'une série sommable. Le lemme 3.1.1 donne de plus des précisions sur la vitesse de cette convergence dans certains cas. Voyons cependant les choses directement, en admetant l' *a priori* suivant : *les conditions de l'expérience permettent de supposer que* $f \in \mathcal{C}(s,L) = D \cap \{g \in \mathcal{W}^s(I) \ : \ \|g\|_{\mathcal{W}^s} \leq L\}$, où D est l'ensemble de toutes les densités de probabilités sur $[0,1]$. Prenons alors $g_0(t) = 1$ et $g_k(t) = \sqrt{2}\cos(2\pi\, k\, t) =$

$(e_k + e_{-k})/\sqrt{2}$ pour $k > 0$ impair et $g_k(t) = \sqrt{2}\sin(2\pi\, k\, t)(e_k - e_{-k})/\sqrt{2}$ pour $k > 0$ pair. Alors

$$\|f\|^2_{\mathcal{W}^s} = |\langle f, g_0\rangle|^2 + \sum_{k\geq 0} |\langle f, g_{2k}\rangle|^2 (1+k)^{2s} + \sum_{k\geq 0} |\langle f, g_{2k+1}\rangle|^2 (1+k)^{2s}.$$

On a aussi que V_{2m+1} est l'intersection de vect$(e_{-m}, \ldots, e_m)$ avec les fonctions réelles. D'où, d'après (1.8),

$$\sum_{k>m} |\langle f, g_k\rangle|^2 = \sum_{|k|>m} |\langle f, e_k\rangle|^2 \leq \frac{L^2}{(2+m)^{2s}}.$$

En conclusion, nous avons, pour tout $f \in \mathcal{C}(s, L)$,

$$\mathrm{MISE}_f(\widehat{f}_{2m+1,n}) \leq L^2\,(2+m)^{-2s} + \frac{2m+1}{n}\|f\|_\infty. \tag{3.13}$$

On notera l'analogie avec l'inégalité (2.11) obtenue pour les estimateurs à noyaux. On constatera notamment que le paramètre m^{-1} joue ici le même rôle que la fenêtre h dans (2.11) et qu'il y a lieu de réaliser le même type de compromis que dans (2.11) entre les vitesses de convergence de m^{-1} vers zéro d'une part et de n vers l'infini, d'autre part. A titre d'exemple, la suite d'estimateurs $\widehat{f}_{[cn^{1+2s}],n}$, avec $c > 0$ donnerait la *bonne* (au sens minimax) vitesse pour le MISE dés que $s > 1/2$. En effet, remarquons que, par inégalité de Cauchy-Schwarz, pour tout $s > 1/2$ et tout $f \in \mathcal{W}(s, L)$,

$$\sum_{|k|\leq m} |\langle f, e_k\rangle| = \sum_{|k|\leq m} |\langle f, e_k\rangle|(1+|k|)^s(1+|k|)^{-s} \leq$$
$$\left(\sum_{|k|\leq m} |\langle f, e_k\rangle|^2(1+|k|)^{2s} \sum_{|k|\leq m} (1+|k|)^{-2s}\right)^{1/2} \leq L\sum_{k\in\mathbb{Z}}(1+|k|)^{-2s}.$$

Le cas $s > 1/2$ est celui pour lequel $\|f\|_\infty$ est donc borné dans $\mathcal{C}(s, L)$.

Voyons directement comment, dans le cas présent, on détermine le $m(n)$ optimal par le principe d'estimation sans biais du risque (cf. section 2.4).

3.5 Estimation sans biais du risque

On se propose d'appliquer le **principe d'estimation sans biais du risque** à l'estimateur (3.7). Reprenons la démarche de la section 2.4. On partira de nouveau de la relation (2.20) (qui, en toute rigueur, doit être écrite explicitement "en module", pour tenir compte des estimateurs à valeurs complexes qui interviennent ici) :

$$\begin{aligned}\mathrm{MISE}_f(\widehat{f}_{2m+1,n}) &= \mathbb{E}_f\left[\int |\widehat{f}_{m,n}(t,X) - f(t)|^2\,dt\right] \\ &= \mathbb{E}_f\left[\int |\widehat{f}_{m,n}(t,X)|^2\,dt - 2\int \widehat{f}_{m,n}(t,X)f(t)\,dt\right] \\ &\qquad + \int (f(t))^2\,dt\end{aligned}$$

ou encore, en notations de l'espace de Hilbert $L^2(I)$:

$$\mathrm{MISE}_f(\widehat{f}_{2m+1,n}) = \mathbb{E}_f\left[\|f - \widehat{f}_{2m+1,n}\|_H^2\right] = \|f\|_H^2 + \mathbb{E}_f\left[\|\widehat{f}_{2m+1,n}\|_H^2\right] - 2\langle f, \mathbb{E}_f[\widehat{f}_{2m+1,n}]\rangle$$

On cherche maintenant (cf. section 2.4) un estimateur sans biais de $\mathbb{E}_f\left[\|\widehat{f}_{2m+1,n}\|_H^2\right] - 2\langle f, \mathbb{E}_f[\widehat{f}_{2m+1,n}]\rangle$. De façon évidente, $U(X_1,\cdots,X_n) = \int |\widehat{f}_{2m+1,n}(t,X_1,\cdots,X_n)|^2\,dt$ en est un pour $\mathbb{E}_f\left[\|\widehat{f}_{2m+1,n}\|_H^2\right]$. On vérifie ensuite que, comme dans le cas des estimateurs à noyau (section 2.4), *l'estimateur construit ici à partir de $\widehat{f}_{2m+1,n}$ en appliquant le principe de validation croisée est bien un estimateur sans biais de $\langle f, \mathbb{E}_f[\widehat{f}_{2m+1,n}]\rangle$* . En effet, le principe de validation croisée conduit ici à la construction, pour $j = 1, ..., n$, des estimateurs :

$$\widehat{f}_{2m+1,n,j} = \frac{1}{n-1}\sum_{k=-m}^{m}\left(\sum_{i\neq j}^{n} g_k(X_i)\right) g_k(X_j)$$

et, du fait que la suite (X_i) est i.i.d, on a pour $i \neq j$:

$$\mathbb{E}_f\left[g_k(X_i)g_k(X_j)\right] = \int\int_{I\times I} g_k(u)g_k(v)f(u)f(v)dudv = |\langle f, g_k\rangle|^2 .$$

Ainsi :

$$\mathbb{E}_f\left[\widehat{f}_{2m+1,n,j}\right] = \sum_{k=-m}^{m} |\langle f, g_k\rangle|^2$$

Or, nous savons que $\mathbb{E}_f\left[\widehat{f}_{2m+1,n}(t,X)\right] = \sum_{k=-m}^{m}\langle f, g_k\rangle\, g_k(t)$ (cf. p.ex. (3.10) et, par conséquent :

$$\langle f, \mathbb{E}_f[\widehat{f}_{2m+1,n}]\rangle = \sum_{k=-m}^{m} |\langle f, g_k\rangle|^2 \tag{3.14}$$

Les estimateurs $\widehat{f}_{2m+1,n,j}$ sont donc bien des estimateurs sans biais de $\langle f, \mathbb{E}_f[\widehat{f}_{2m+1,n}]\rangle$. Ainsi, l'estimateur CV, qui prend la forme (2.24) dans le cas de la validation croisée pour un estimateur à noyau, sera défini dans le cas présent par :

$$\mathrm{CV}(m, X_1, .., X_n) := \int_I |\widehat{f}_{2m+1,n}(t)|^2 \, dt \; - \; \frac{2}{n} \left(\sum_{j=1}^{n} \widehat{f}_{2m+1,n,j} \right) \tag{3.15}$$

et l'estimateur (plutôt : pseudo-estimateur) sans biais du risque $\mathrm{MISE}_f(\widehat{f}_{m,n})$ s'écrit

$$\widehat{\mathrm{MISE}}(m, X) = \mathrm{CV}(m, X) + \|f\|_H^2.$$

L'estimateur adaptatif $\widehat{f}_{2\widehat{m}+1,n}$ est donc défini par le choix de $\widehat{m}$ correspondant à :

$$\widehat{m}(X_1, .., X_n) = \operatorname*{argmin}_{m \geq 0} \widehat{\mathrm{MISE}}(m, X_1, .., X_n) = \operatorname*{argmin}_{m \geq 0} \mathrm{CV}(m, X_1, .., X_n).$$

3.6 Estimateur de projection pour la régression

On reprend le modèle (2.27). Nous avons ici $\mathbb{P}_{f,n} = \mathbb{P}^n_{f,\mathbb{P}_Y,\mathbb{P}_\epsilon}$ et $\phi(f, \mathbb{P}_Y, \mathbb{P}_\epsilon) = f$ est la fonction inconnue du paramètre à estimer. On notera $\mathbb{P}^f_{(Y_1,Z_1)}$ la loi commune des (Y_i, Z_i) sous f. Nous savons que, dans les conditions du modèle (2.27), on a $f(y) = \mathbb{E}_f[Z|Y = y]$. On écrira donc :

$$f = \operatorname*{argmin}_{g \in H} \int |z - g(y)|^2 \, \mathbb{P}^f_{(Y_1,Z_1)}(dy, dz) =: \operatorname*{argmin}_{g \in H} \mathcal{C}(g, \mathbb{P}^f_{(Y_1,Z_1)}) \tag{3.16}$$

(tout repose donc ici sur l'hypothèse que nous connaissons l'espace $H = L^2(\mathbb{P}_Y)$, même si nous ignorons la loi $\mathbb{P}_Y$ elle -même). Considèrons maintenant une suite croissante (V_m) de sous-espaces vectoriels de H satisfaisant (3.3) et construisons l'estimateur :

$$\widehat{f}_{m,n}((Y_1, Z_1), ..., (Y_n, Z_n)) = \operatorname*{argmin}_{g \in V_m} \mathcal{C}(g, P_n) \tag{3.17}$$

avec $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{(Y_i,Z_i)}$. Alors, on a bien sous chaque f, en vertu de la loi forte des grands nombres, presque sûrement,

$$\mathcal{C}(g, P_n) = \frac{1}{n} \sum_{j=1}^{n} |g(Y_j) - Z_j|^2 \to \mathbb{E}_f \left[|Z_1 - g(Y_1)|^2 \right] = \mathcal{C}(g, \mathbb{P}^f_{(Y_1,Z_1)})$$

(cf. (3.16). Nous nous trouvons donc dans le cadre défini à la section (3.3) avec :

$$\widehat{f}_{m,n} = \underset{g \in V_m}{\operatorname{argmin}} \sum_{j=1}^{n} |g(Y_j) - Z_j|^2 . \tag{3.18}$$

Comme dans le cas de l'estimation de densité, la forme de cet estimateur, ainsi que la facilité du calcul dépendront fortement du choix des espaces V_m. Les calculs sont fortement simplifiés en considérant le cas où V_m est engendré par une base orthogonale.

3.7 Un exemple de critère pénalisé pour la régression

Nous allons examiner cela plus en détails dans le cas de *l'approximation polynomiale locale* associée un *critère pénalisé*.

Pour un estimateur à noyau (resp. estimateur de projection (3.7)), le *paramètre de lissage* est h (resp. m). Une autre variante d'estimateurs découle de l' utilisation d' un critère pénalisé :

$$\widehat{\phi(f)}_{\lambda,n} = \underset{g \in H}{\operatorname{argmin}} \left(\mathcal{C}(g, P_n) + \lambda \ \mathrm{PEN}(g) \right) \quad \text{avec} \quad \lambda > 0.$$

où $\mathrm{PEN}(g)$ est une *fonction de pénalisation* et λ jouera le rôle de paramètre de lissage analogue à m et h dans les exemples précédents. En fait, le critère pénalisé ci-dessus est souvent équivalent à une projection sur un sous-espace de dimension finie de H, sous-espace entièrement paramétré par λ.

Nous allons voir, pour un critère particulier, que ce problème conduit à des calculs relativement faciles dans le cadre de la régression non-paramétrique.

Nous supposerons que pour le problème de regression non-paramétrique auquel nous nous intéressons, on peut admettre un *a priori* de la forme $g \in \mathcal{W}^2(I)$, I étant ici un intervalle (a, b) avec $-\infty \leq a < b \leq \infty$ et nous allons considérer l'estimateur obtenu par la pénalisation suivante :

$$\widehat{f}_{\lambda,n} = \underset{g \in \mathcal{W}^2}{\operatorname{argmin}} \left(\sum_{j=1}^{n} |g(Y_j) - Z_j|^2 + \lambda \int_I |g''(t)|^2 \, dt \right) \tag{3.19}$$

Ce problème de minimisation possède une solution numérique simple grâce à l'utilisation des *spline cubiques*.

Définition 3.7.1 (Espaces spline)
Soient $a = t_0 < t_1 < \ldots < t_{p-1} < t_p = b$, des points choisis arbitrairement dans l'intervalle I. On note $\mathcal{S}((t_i)_{i=0,\ldots,p})$ l'ensemble des fonctions 2 fois continument dérivables sur I dont les restrictions aux intervalles (t_i, t_{i+1})

sont des polynômes de degré au plus 3 *si* $0 < i < p-1$ *et de degré au plus* 1 *sinon.*

Il est clair que $\mathcal{S}((t_i)_{i=0,\dots,p})$ est un sous espace vectoriel de $\mathcal{W}^2$. Nous allons exposer maintenant quelques propriétés de cet espace qui en font un outil trés intéressant pour la résolution de(3.19).

Nous avons le résultat suivant.

Théorème 3.7.2
L'espace $\mathcal{S}((t_i)_{i=0,\dots,p})$ *est de dimension* $p-1$ *et les éléments suivants forment une base de cet espace, appelée* base spline *:*

$$\psi_0,\, \psi_1,\; f_j \;=\; \xi_j \;-\; \xi_{p-2} \qquad (j = 1, \dots, p-3)$$

où :

$$\psi_i(t) \;=\; t^i, \quad \xi_j = \frac{\phi_j - \phi_{p-1}}{t_j - t_{p-1}} \qquad (j = 1, \dots, p-2)$$

avec :

$$\phi_j(t) \;=\; \max((t-t_j)^3, 0) \qquad (j = 1, \dots, p-1)$$

De plus, pour toute suite de nombres (réels ou complexes) $z_1, \dots, z_{p-1}$*, il existe une unique fonction interpolante* $\tilde{g} \in \mathcal{S}((t_i)_{i=0,\dots,p})$ *telle que* $\tilde{g}(t_i) = z_i$ *pour* $i = 1, \dots, p-1$*.*

Démonstration Il est aisé de vérifier que la base proposée appartient bien à l'espace spline. La dimension est facilement obtenue en comptant les degrés de liberté imposées aux polynômes par morceaux.

Pour le problème d'interpolation, on va construire, de proche en proche sur les intervalles (t_i, t_{i+1}), une fonction $\tilde{g} \in \mathcal{S}((t_i)_{i=0,\dots,p})$ que l'on "ajustera" ensuite pour satisfaire toutes les conditions requises. Notons $\tilde{g}_i$ la restriction de $\tilde{g}$ à l'intervalle (t_i, t_{i+1}) et posons sur (t_0, t_1) : $\tilde{g}_0(t) = u_0 t + v_0$, avec $\tilde{g}_0(t_1) = z_1$.

A partir de cette donnée, la fonction $\tilde{g}_1(t) = a_0^1 t^3 + a_1^1 t^2 + a_2^1 t + a_3^1$ sur (t_1, t_2) sera déterminée de façon unique par le système de quatre équations linéaires (aux inconnues $a_0^1, a_1^1, a_2^1, a_3^1$) suivantes : $\tilde{g}_1(t_1) = z_1, \tilde{g}_1'(t_1) = u_0$, $\tilde{g}_1''(t) = 0$ et $\tilde{g}_1(t_2) = z_2$. *On remarquera que les solutions* a_j^1 *de ce système sont des fonctions linéaires de* u_0 *(la pente choisie pour* $\tilde{g}_0$*) dont les coefficients dépendent uniquement de* $z_1, z_2, t_1^k, t_2^k, k = 1, 2, 3$.

De proche en proche, si l'on a construit $\tilde{g}_i$ pour $i < p-3$, $\tilde{g}_{i+1}(t) = a_0^{i+1} t^3 + a_1^{i+1} t^2 + a_2^{i+1} t + a_3^{i+1}$ sur (t_{i+1}, t_{i+2}) sera déterminée de façon unique à partir des équations $\tilde{g}_{i+1}(t_{i+1}) = z_{i+1}$, $\tilde{g}_{i+1}'(t_{i+1}) = \tilde{g}_i'(t_{i+1})$, $\tilde{g}_{i+1}''(t_{i+1}) = \tilde{g}_i''(t_{i+1})$ et $\tilde{g}_{i+1}(t_{i+2}) = z_{i+2}$. Et les solutions a_j^{i+1} de ces équations seront de nouveau des fonctions linéaires de u_0 (avec des coefficients dépendants uniquement de $z_j, t_j^k, j = 1, \dots, i+2, k = 1, 2, 3$).

De sorte que, une fois parvenus à avoir $\tilde{g}_{p-2}$ sur (t_{p-2}, t_{p-1}), nous construirons $\tilde{g}_{p-1}(t) = u_1 t + v_1$ sur (t_{p-1}, t_p) à partir des conditions $\tilde{g}_{p-1}(t_{p-1}) = z_{p-1}$ et $\tilde{g}'_{p-1}(t_{p-1}) = \tilde{g}'_{p-2}(t_{p-1})$. La fonction $\tilde{g}$ construite ainsi sur (t_0, t_p) satisfaira toutes les conditions requises, sauf peut-être celle d'être de classe $\mathcal{C}^2$ au point t_{p-1}. En effet, rien ne garantissait, dans la construction de $\tilde{g}_{p-2}$, d'avoir $\tilde{g}''_{p-2}(t_{p-1}) = 0 = \tilde{g}''_{p-1}(t_{p-1})$.

La dépendance de ces quantités en u_0 étant linéaire, il existe un unique u_0 satisfaisant cette contrainte. ∎

Proposition 3.7.3
Supposons que l'on ait observé $Y_1, \ldots, Y_n$ distincts dans (a, b). Alors, pour tout $\lambda > 0$, la solution de (3.19) appartient à l'espace $\mathcal{S}(a, Y_{(1)}, \ldots, Y_{(n)}, b)$, où $(Y_{(1)}, \ldots, Y_{(n)})$ est la suite des valeurs $Y_1, \ldots, Y_n$ rangées dans l'ordre croissant.

Démonstration Supposons que l'on ait observé $Y_1, \ldots, Y_n$. Prenons $g \in \mathcal{W}^2$, posons

$$J(g) = \sum_{j=1}^{n} |g(Y_j) - Z_j|^2 + \lambda \int_I |g''(t)|^2 \, dt$$

et notons $\tilde{g}$ l'unique élément de $\mathcal{S}((0, Y_{(1)}, \ldots, Y_{(n)}, 1))$ vérifiant $g(Y_i) = \tilde{g}(Y_i)$ pour $i = 1, \ldots, n$. On va établir que

$$J(g) = J(\tilde{g}) + \int_I |g''(t) - \tilde{g}''(t)|^2 \, dt,$$

ce qui prouvera bien que le minimum de J ne peut être atteint que dans $\mathcal{S}((0, Y_{(1)}, \ldots, Y_{(n)}, 1))$.

Par définition de $\tilde{g}$, on a $\sum_{j=1}^{n} |g(Y_j) - Z_j|^2 = \sum_{j=1}^{n} |\tilde{g}(Y_j) - Z_j|^2$. Reste à comparer $\int_I |g''(t)|^2 \, dt$ avec $\int_I |\tilde{g}''(t)|^2 \, dt$.

Pour simplifier, notons nos observations y_i et supposons-les déjà rangées dans l'ordre croissant : $y_1 < \ldots < y_{n-1} < y_n$. Nous avons :

$$\int_I |g''(t)|^2 \, dt = \int_I |g''(t) - \tilde{g}''(t)|^2 \, dt + \int_I |\tilde{g}''(t)|^2 \, dt + 2\Re \int_I \overline{\tilde{g}''}(t)(g''(t) - \tilde{g}''(t)) \, dt$$

et on aura établi le résultat, si l'on montre que $K = \int_I \overline{\tilde{g}''}(t)(g''(t) - \tilde{g}'') \, dt = 0$. On écrit

$$
\begin{aligned}
K &= \int_a^b \tilde{g}''(t)(g''(t) - \tilde{g}''(t))\,dt \\
&= \int_{y_1}^{y_n} \tilde{g}''(t)(g''(t) - \tilde{g}''(t))\,dt \\
&= \left[\overline{\tilde{g}''}(t)(g'(t) - \tilde{g}'(t))\right]_{y_1}^{y_n} - \int_{y_1}^{y_n} \overline{\tilde{g}'''}(t)(g'(t) - \tilde{g}'(t))\,dt \\
&= -\int_{y_1}^{y_n} \overline{\tilde{g}'''}(t)(g'(t) - \tilde{g}'(t))\,dt
\end{aligned}
$$

où nous avons utilisés successivement que $\tilde{g}''$ est continue sur (a, b), nulle entre a et y_1, ainsi qu'entre y_n et b, ainsi qu'aux points y_1 et y_n. L'intégration par partie est valide car $\tilde{g}''$ est continue et C^1 par morceaux.

Il suffit maitenant de remarquer que sur chaque intervalle (y_i, y_{i+1}), la fonction $\tilde{g}'''$ est constante, de sorte que la dernière intégrale sur un tel intervalle est proportionnelle à $[g(y_{i+1}) - \tilde{g}(y_{i+1})] - [g(y_i) - \tilde{g}(y_i)] = 0$ (par construction de $\tilde{g}$). Ceci prouve le résultat. ∎

La méthode d'approximations polynomiales locales en bases spline consiste à calculer l'élément de l'espace $\mathcal{S}((0, Y_{(1)}, \ldots, Y_{(n)}, 1))$ qui résoud (3.19), par exemple en l'exprimant dans la base spline indiquée dans le Théorème 3.7.2.

Chapitre 4

Exercices

4.1 Des rappels autour de la régression linéaire

4.1 Exercice:
On considère le modèle de régression de dimension N

$$x_k = \theta_k + \epsilon_k, \quad k = 1, \dots, N,$$

où $\mathbf{x} = (x_k)_{k=1,\dots,N}$ est l'observation dans $\mathbb{R}^N$, $\boldsymbol{\theta} = (\theta_k)_{k=1,\dots,N}$ le paramètre de $\mathbb{R}^N$ et les ϵ_k sont des variables i.i.d. centrées de variance connue σ^2.

1. On cherche un estimateur de $\boldsymbol{\theta}$ de la forme $\tilde{\boldsymbol{\theta}}(\lambda) = \lambda \mathbf{x}$ avec $\lambda \in [0, 1]$. Calculer son biais $\mathbb{E}[\tilde{\boldsymbol{\theta}}(\lambda)] - \boldsymbol{\theta}$ et sa variance sommée $\sum_{k=1}^N \mathrm{Var}(\tilde{\theta}_k(\lambda))$. Quel est l'effet de λ sur le biais et la variance ?
2. Calculer $\lambda^* \in [0, 1]$ qui minimise le risque d'oracle de cet estimateur :

$$R(\lambda) = \mathbb{E}\left[\sum_{k=1}^N (\theta_k - \lambda x_k)^2\right] .$$

3. Proposer un estimateur sans biais de θ_k^2 à partir de x_k et σ^2.
4. En déduire un estimateur sans biais $\mathcal{C}(\lambda)$ de $R(\lambda) - \sum_{k=1}^N \theta_k^2$.
5. Quel est l'estimateur $\widehat{\boldsymbol{\theta}} = \lambda(\mathbf{x})\,\mathbf{x}$, avec $\lambda : \mathbb{R}^N \to [0, 1]$, obtenu par minimisation du critère $\mathcal{C}$ sur $[0, 1]$? ♢

4.2 Exercice:
Notation : Dans cet exercice, on pourra utiliser les fonctions quantile et de répartition des lois usuelles : on notera Q et F celles de la gaussienne, Qt_d et Ft_d celles de la loi de Student à d degrés de liberté, $Q\chi_d$ et $F\chi_d$ celles de la loi du χ^2 à d degrés de liberté, Qf_{d_1,d_2} et Ff_{d_1,d_2} celles de la loi de Fisher à $(d_1, d_2$ degrés de liberté.

On considère un modèle linéaire

$$\mathbf{Y} = \mathbf{X}\beta + \sigma\boldsymbol{\epsilon} ,$$

où $\mathbf{Y}$ est un vecteur de $\mathbb{R}^n$ d'observations, $\mathbf{X}$ est une matrice déterministe (connue) de taille $n \times p$ et β est un paramètre inconnu de $\mathbb{R}^p$, $\boldsymbol{\epsilon}$ est un vecteur gaussien de $\mathbb{R}^n$ de moyenne nulle et de matrice de covariance identité I_n et $\sigma^2 > 0$ est un paramètre inconnu. On note I_p la matrice identité $p \times p$. On note r le rang de $\mathbf{X}$. On note, pour tout $\lambda \geq 0$,

$$\widehat{\beta}_\lambda = \arg\min_{u \in \mathbb{R}^p} \left\{ \|\mathbf{Y} - \mathbf{X}u\|^2 + \lambda \|u\|^2 \right\} \ ,$$

quand il y a existence et unicité du minimum.

1. Comment s'appelle cet estimateur pour $\lambda = 0$? Donner son expression explicite lorsque $r = p$.
2. Comment s'appelle ce type de critère pour définir un estimateur quand $\lambda > 0$? Montrer que, pour tout $\lambda \geq 0$, $\widehat{\beta}_\lambda$ est donné de façon équivalente par l'équation
$$(\mathbf{X}^T\mathbf{X} + \lambda \mathrm{I}_p)\widehat{\beta}_\lambda = \mathbf{X}^T Y \ .$$
3. Montrer que le noyau de $\mathbf{X}$ est le même que celui de $\mathbf{X}^T\mathbf{X}$.
4. L'estimateur $\widehat{\beta}_\lambda$ est-il correctement défini si $\lambda = 0$ et $r \leq p - 1$?
5. Montrer que la matrice $\mathbf{X}^T\mathbf{X} + \lambda \mathrm{I}_p$ est toujours inversible pour $\lambda > 0$ quelque soit le rang de $\mathbf{X}$.
6. Dans quel cas pratique d'utilisation de ce modèle peut-il être utile de prendre $\lambda > 0$?

On suppose désormais $\lambda > 0$.

7. Montrer que le biais de cet estimateur s'écrit :
$$\mathrm{biais}_\lambda(\beta) = -\lambda(\mathbf{X}^T\mathbf{X} + \lambda \mathrm{I}_p)^{-1}\beta \ .$$
8. Calculer la matrice d'autocovariance $\Sigma(\sigma^2, \lambda)$ de $\widehat{\beta}_\lambda$.

On écrit $\mathbf{X}^T\mathbf{X} = P^T \Delta P$ avec P matrice orthogonale et Δ diagonale à élément diagonaux $\delta_1, \ldots, \delta_r, 0, \ldots, 0$.

9. Montrer que l'on peut définir une matrice $p \times p$ $U(\sigma^2, \lambda)$ telle que $U(\sigma^2, \lambda)\Sigma(\sigma^2, \lambda)U(\sigma^2, \lambda)^T = \mathrm{I}_{r,p}$, où $\mathrm{I}_{r,p}$ est la matrice diagonale ayant les r premières valeurs de la diagonale égales à 1 et le reste à zero.
10. Quelle est la loi de $\widehat{\beta}_\lambda$?
11. On suppose σ^2 connu. Peut-on construire une fonction pivotale de β ?
12. En déduire un test de l'hypothèse $H_0 = \{\beta = 0\}$ contre $H_1 = \{\beta \neq 0\}$ de niveau α. Exprimer la p-valeur associée à ce test.
13. On trouve une p-valeur égale à 0.78. Quelle interprétation donnée ? Accepteriez–vous ou rejetteriez–vous H_0 pour un niveau égale à 0.05 ?
14. Construire un intervalle de confiance pour la première composante $\beta_1$de probabilité de couverture $1 - \alpha$. On supposera $r = p$ pour cette question.
15. Comment se comporte $\mathrm{biais}_\lambda(\beta)$ quand $\lambda \downarrow 0$?
16. Même question pour $\Sigma(\sigma^2, \lambda)$. $\diamondsuit$

4.2 Estimation d'une densité par l'histogramme

4.3 Exercice:
On souhaite estimer "globalement" une densité de probabilité inconnue sur un intervalle donné, disons $[0,1]$ pour simplifier, à partir de l'observation de la réalisation d'un n-échantillon $(X_1,\dots,X_n)$. Cela signifie que les variables aléatoires réelles $X_1, X_2, \dots, X_n$ sont indépendantes et identiquement distribuées. Dans toute la suite, on notera $x \mapsto f(x)$ leur densité de probabilité commune définie sur $[0,1]$. Le but est d'estimer les valeurs $\big(f(x), x \in [0,1]\big)$ simultanément.

Soit $m \geq 1$ un entier. On définit les *boites* $B_1, B_2, \dots, B_m$ en posant :

$$B_1 = \big[0, \tfrac{1}{m}\big)\,, \quad B_2 = \big[\tfrac{1}{m}, \tfrac{2}{m}\big)\,, \dots, \quad B_m = \Big[\tfrac{(m-1)}{m}, 1\Big] \ .$$

On appelle *largeur de bande* associée aux boites B_j le nombre $h = 1/m$. Pour $j = 1, \dots, m$, on définit

$$\widehat{p_j} = \frac{1}{n} \#\left\{X_i \in B_j,\ i = 1, \dots, n\right\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{B_j}(X_i) \ .$$

1. Montrer que $\widehat{p_j}$ est un estimateur sans biais de $p_j = \int_{B_j} f(u)\mathrm{d}u$. Quelle est sa variance ?

L'estimateur par histogramme de la densité est alors défini par la formule

$$\widehat{f_n}(x) = \sum_{j=1}^{m} \frac{\widehat{p_j}}{h}\ \mathbb{1}_{B_j}(x) \qquad \text{pour } \ x \in [0,1] \ .$$

2. Soit j l'indice de la boite contenant x. Montrer que

$$\mathbb{E}[\widehat{f_n}(x)] = \frac{p_j}{h} \quad \text{et} \quad \mathrm{Var}\Big[\widehat{f_n}(x)\Big] = \frac{p_j(1-p_j)}{nh^2}.$$

3. Supposons que f est continue en x. Que dire du biais de $\widehat{f_n}(x)$ pour estimer $f(x)$ quand h tend vers 0 ?

On définit *l'erreur quadratique moyenne intégrée* de l'estimateur $\widehat{f_n}$ de f en posant

$$\mathcal{R}(\widehat{f_n}, f) = \mathbb{E}\left[\int_0^1 \big(\widehat{f_n}(u) - f(u)\big)^2 \mathrm{d}u\right] \ .$$

On suppose désormais que f est 2 fois continûment dérivable sur $[0,1]$.

On note $b(x) = \mathbb{E}[\widehat{f_n}(x)] - f(x)$ le biais de l'estimateur $\widehat{f_n}(x)$ et $v(x) = \mathrm{Var}\Big[\widehat{f_n}(x)\Big]$ sa variance.

4. Montrer que

$$b(x) = f'(x)\big(h(j - \tfrac{1}{2}) - x\big) + O(h^2) \ ,$$

où $x \in B_j$.

5. Montrer que

$$\int_0^1 b(x)^2 \mathrm{d}x = \frac{h^2}{12} \int_0^1 \big(f'(x)\big)^2 \mathrm{d}x + o(h^2) \; .$$

On pourra utiliser que $f'(x) = f'(x_j) + O(h)$ pour $x \in B_j$, où x_j désigne le centre de la boite B_j.

6. Comment varie le biais en fonction de h ?
7. En reproduisant les arguments précédents, montrer que

$$v(x) = \frac{f(x) + O(h)}{nh} \; ,$$

puis que

$$\int_0^1 v(x) \mathrm{d}x = \frac{1}{nh} + O(1/n) \; .$$

8. Comment varie la variance en fonction de h ?
9. Déduire des questions précédentes que

$$\mathcal{R}(\widehat{f}_n, f) = \frac{h^2}{12} \int_0^1 f'(u)^2 \mathrm{d}u + \frac{1}{nh} + o(h^2) + o(1/(nh)) \; . \qquad (4.1)$$

10. On note $\widehat{f}_n = \widehat{f}_{n,h}$ pour mettre en évidence la dépendance en h de l'estimateur. Montrer que

$$\lim_{n\to\infty} n^{2/3} \inf_h \mathcal{R}^0(\widehat{f}_{n,h}, f) = (3/4)^{2/3} \left(\int_0^1 f'(u)^2 \mathrm{d}u \right)^{1/3} \; ,$$

où $\mathcal{R}^0$ est l'approximation du risque obtenue en négligeant les termes en $o(\dots)$ dans (4.1). ♢

Nous avons vu que la taille de fenêtre optimale $h_n^\star = h_n^\star(f)$ dépend de f, qui est inconnu. On s'intéresse maintenant au problème du choix automatique de la fenêtre pour l'estimation de la densité par histogramme. Nous allons donc chercher un choix de h dicté par l'observation $X_1, \dots, X_n$ uniquement, et dont l'erreur imite le mieux possible l'erreur idéale fournie par le choix de $h_n^\star$. Nous considérons la méthode de **validation croisée** de type *leave one out*. Nous écrivons désormais $\widehat{f_n(x)} = \widehat{f_{n,h}(x)}$ et définissions

$$L_n(h) = \int_0^1 \big(\widehat{f_{n,h}(u)} - f(u)\big)^2 du = \int_0^1 (\widehat{f_{n,h}(u)})^2 du - 2 \int_0^1 \widehat{f_{n,h}(u)} f(u) du + \int_0^1 f(u)^2 du.$$

Définition 4.2.1
L'estimateur du risque par validation croisée est

$$\widehat{J_n}(h) = \int_0^1 (\widehat{f_{n,h}(u)})^2 du - \frac{2}{n}\sum_{i=1}^{n} \widehat{f_{n,h,i}}(X_i) \ ,$$

où $\widehat{f_{n,h,i}}(x)$ *est l'estimateur de* f *au point* x *obtenu en ignorant la donnée* X_i.

4.4 Exercice:

1. Montrer que minimiser $\mathcal{R}(\widehat{f_{n,h}}, f)$ est équivalent à minimiser l'espérance de
$$J_n(h) = \int_0^1 (\widehat{f_{n,h}(u)})^2 du - 2\int_0^1 \widehat{f_{n,h}(u)} f(u) du.$$
2. Comparer $\mathbb{E}[\widehat{J_n}(h)]$ et $\mathbb{E}[J_n(h)]$. ♢

En principe, pour minimiser $h \mapsto \widehat{J_n}(h)$, on doit reconstruire n histogrammes pour chaque valeur de h. Heureusement, on dispose du raccourci suivant.

3. Montrer que :
$$\widehat{J_n}(h) = \frac{2}{(n-1)h} - \frac{1}{h}\frac{n+1}{n-1}\sum_{j=1}^{m} \widehat{p_j}^2.$$

Bibliographie

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95284-5. Data mining, inference, and prediction.

Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004. ISBN 3-540-40592-5.