Modèles non paramétriques

Davy Paindaveine

2024–2025

Table des matières

Chapitre I

Introduction

I.1 Statistique paramétrique / non paramétrique

Dans les cours d'introduction à la statistique, on considère typiquement des observations aléatoires $X_1, \ldots, X_n$ dont le comportement est décrit par un modèle statistique de la forme

$$\Big(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, \mathcal{P}^{(n)} = \big\{P_\theta^{(n)} : \theta \in \Theta \subset \mathbb{R}^k\big\}\Big),$$

où $\mathcal{X}^{(n)}$ est l'espace dans lequel $(X_1, \ldots, X_n)$ prend ses valeurs (le plus souvent, $\mathcal{X}^{(n)} = \mathbb{R}^n$, mais on peut aussi avoir, par exemple, $\mathcal{X}^{(n)} = (\mathbb{R}^d)^n$ dans le cas d'observations X_i qui seraient à valeurs dans $\mathbb{R}^d$), $\mathcal{A}^{(n)}$ est une σ-algèbre (une tribu) sur $\mathcal{X}^{(n)}$ et où $\mathcal{P}^{(n)}$ est une collection de mesures de probabilité possibles pour la loi de $(X_1, \ldots, X_n)$. Les modèles d'échantillonnage, qui sont le plus classiquement considérés, prévoient que les observations $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées, ce qui se traduit dans le cadre du modèle ci-dessus par le fait que les lois $P_\theta^{(n)}$ sont des lois-produit d'une mesure de probabilité P_θ sur $\mathbb{R}$ représentant la loi commune des X_i.

Dans tous les cas, le modèle ci-dessus sera dit *paramétrique*, parce que la collection de lois $\mathcal{P}^{(n)}$ est indicée par un paramètre fini-dimensionnel θ. Un exemple de tel modèle est le modèle de lois exponentielles pour lequel la loi P_θ admet la densité [1]

$$f_\lambda(x) := \frac{1}{\lambda} e^{-x/\lambda} \mathbb{I}[x > 0],$$

qui fait intervenir le paramètre scalaire $\theta = \lambda \in \Theta = \mathbb{R}_0^+ \subset \mathbb{R}$. Un autre exemple est le

1. Dans ce cours, " :=" signifiera "est égal par définition à" et $\mathbb{I}[A]$ désignera l'indicatrice de la condition A, qui prend la valeur 1 si la condition est satisfaite et la valeur 0 sinon.

modèle de lois gaussiennes pour lequel la loi P_θ admet la densité

$$f_{\mu,\sigma^2}(x) := \phi_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{I.1}$$

pour laquelle le paramètre $\theta = \binom{\mu}{\sigma^2} \in \Theta = \mathbb{R}\times\mathbb{R}_0^+ \subset \mathbb{R}^2$ est bidimensionnel. Dans le cadre paramétrique ci-dessus, estimer la loi de $(X_1,\ldots,X_n)$ revient à estimer le paramètre θ. Un estimateur de θ est une fonction mesurable $\hat\theta_n = \hat\theta_n(X_1,\ldots,X_n)$ à valeurs dans Θ. On dispose de méthodes classiques, comme par exemple la méthode des moments ou la méthode du maximum de vraisemblance, pour construire des estimateurs de θ. Dans une large classe de modèles statistiques paramétriques (les modèles dits *réguliers*, c'est-à-dire ceux dans lesquels la borne de Cramér–Rao est valide), il existe des estimateurs vérifiant [2]

$$\mathrm{MSE}(\hat\theta_n) := \mathrm{E}[(\hat\theta_n-\theta)^2] = O\Big(\frac{1}{n}\Big) \tag{I.2}$$

quand n diverge vers l'infini [3], mais il n'existe pas d'estimateurs tels que

$$\mathrm{MSE}(\hat\theta_n) = o\Big(\frac{1}{n}\Big)$$

quand n diverge vers l'infini. Dans de tels modèles paramétriques, l'*erreur quadratique moyenne* (en anglais, *Mean Square Error* ou *MSE*) d'un estimateur $\hat\theta_n$ tend donc vers zéro au mieux à vitesse $1/n$. On dira que ce taux de convergence est paramétrique (notons que dans certains modèles paramétriques, de meilleurs taux de convergence peuvent être obtenus : par exemple, si $X_1,\ldots,X_n$ sont indépendants et de loi uniforme sur $[0,\theta]$, où $\theta \in \mathbb{R}_0^+$, alors l'estimateur $\hat\theta_n = \max(X_1,\ldots,X_n)$ vérifie $\mathrm{MSE}(\hat\theta_n) = O(\frac{1}{n^2})$). Bien entendu, (I.2) se réécrit aussi sous la forme $\mathrm{E}[\{\sqrt{n}(\hat\theta_n-\theta)\}^2] = O(1)$, ce qui est typiquement associé au fait que $\sqrt{n}(\hat\theta_n-\theta)$ converge en loi, sous $P_\theta^{(n)}$, vers une certaine loi $\mathcal{L}_\theta$ (laquelle est souvent une loi gaussienne de moyenne zéro et de variance σ_θ^2). On dira que $\hat\theta_n$ est un estimateur $\sqrt{n}$-convergent de θ.

Les modèles paramétriques fournissent un cadre commode, dans le sens où, au moins dans les modèles paramétriques réguliers, des résultats précis décrivent le MSE optimal qui peut être obtenu et des méthodes d'estimation permettent de construire des estimateurs qui réalisent ce MSE optimal quand n diverge vers l'infini. Néanmoins, l'hypothèse paramétrique voulant que la vraie loi de $(X_1,\ldots,X_n)$ appartienne à la collec-

2. Dans (I.2), on suppose bien entendu que le paramètre θ est scalaire ($k=1$). Mais le résultat tient aussi pour $k>1$ si on remplace le MSE univarié en (I.2) par un concept adéquat de MSE multivarié.

3. Rappelons qu'on dira que "$a_n = O(b_n)$ quand n diverge vers l'infini" si et seulement si $(|a_n/b_n|)$ est une suite bornée (si les suites (a_n) et (b_n) tendent vers zéro, ceci implique que (a_n) tend vers zéro au moins aussi vite que (b_n)). De même, on dira que "$a_n = o(b_n)$ quand n diverge vers l'infini" si et seulement si a_n/b_n tend vers zéro (si les suites (a_n) et (b_n) tendent vers zéro, ceci implique que (a_n) tend vers zéro plus vite que (b_n)).

tion de lois $\mathcal{P}^{(n)} = \{P_\theta^{(n)}\}$ est une hypothèse extrêmement forte, qui ne représente au mieux qu'une approximation de la réalité : comment être sûr que des observations sont vraiment de loi exponentielle ? Ou vraiment de loi normale ? Comme nous le verrons au début du Chapitre II, se tromper sur la forme analytique de la loi peut avoir des résultats désastreux dans l'estimation de la loi sous-jacente.

Bien entendu, pour minimiser le risque que la vraie loi de $(X_1, \ldots, X_n)$ n'appartienne pas à la collection de lois $\mathcal{P}^{(n)}$ considérée, on peut décider de travailler avec une collection $\mathcal{P}^{(n)}$ très vaste. Un modèle *non paramétrique* est un modèle de la forme

$$\Big(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, \mathcal{P}^{(n)} = \big\{P^{(n)}\big\}\Big),$$

où $\mathcal{P}^{(n)}$ est si vaste qu'on ne peut l'indicer par un paramètre fini-dimensionnel. En particulier, dans un modèle d'échantillonnage, on pourrait considérer $\mathcal{P}^{(n)} = \{P_F^{(n)} : F \in \mathcal{F}\}$, où $P_F^{(n)}$ est la loi de $(X_1, \ldots, X_n)$ lorsque les observations X_i sont indépendantes et partagent la même fonction de répartition $x \mapsto F(x) = P[X \leq x]$ (ici, $\mathcal{F}$ désigne l'ensemble des fonctions de répartition sur $\mathbb{R}$). Le modèle qui en résulte est indicé par la fonction F, qui est un objet de dimension infinie. Puisque $F(x) = 1 \times P[X \leq x] + 0 \times P[X > x] = \mathrm{E}[\mathbb{I}[X \leq x]]$, un estimateur naturel de $F(x)$ est

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[X_i \leq x]. \tag{I.3}$$

La fonction aléatoire $x \mapsto \hat{F}_n(x)$ est appelée *fonction de répartition empirique* ; voir la Figure I.1. Comme on propose de le vérifier dans les exercices, on a [4]

$$\mathrm{E}[\{\hat{F}_n(x) - F(x)\}^2] = \frac{C_{F,x}}{n} \quad \text{et} \quad \sqrt{n}\{\hat{F}_n(x) - F(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, C_{F,x})$$

quand n diverge vers l'infini, où $C_{F,x}$ est une constante ne dépendant pas de n. Ceci montre que $\hat{F}_n(x)$ est un estimateur $\sqrt{n}$-convergent de $F(x)$. Le problème de l'estimation d'une fonction de répartition suggère que les méthodes d'estimation usuelles s'appliqueront aux problèmes non paramétriques et que celles-ci mèneront à des estimateurs $\sqrt{n}$-convergents.

Ceci est en fait loin d'être le cas, comme nous le verrons dans deux problèmes non paramétriques spécifiques, à savoir celui de l'estimation d'une fonction de densité (Chapitre II) et celui de l'estimation d'une fonction de régression non paramétrique (Chapitre III). Ces deux problèmes requièrent de définir des méthodes d'estimation propres, qui ne mèneront malheureusement pas à des estimateurs $\sqrt{n}$-convergents. De nombreuses

4. Ici, $\xrightarrow{\mathcal{L}}$ désigne la convergence en loi et $\mathcal{N}(\mu, \sigma^2)$ est la loi gaussienne de moyenne μ et de variance σ^2.

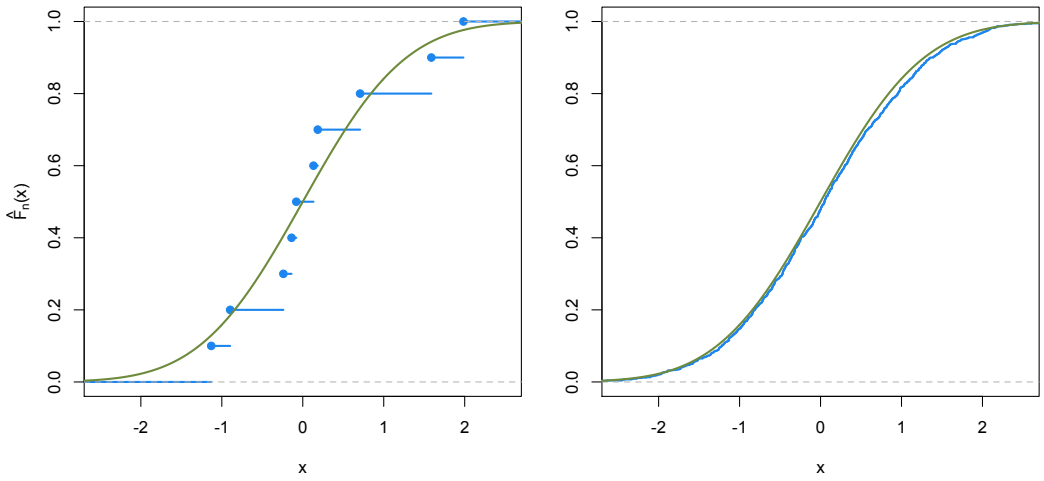

Figure I.1 – *Graphes (en bleu) de la fonction empirique $\hat{F}_n$ calculée sur un échantillon aléatoire de taille $n = 10$ (gauche) et $n = 1000$ (droite) engendré depuis la loi $\mathcal{N}(0, 1)$, c'est-à-dire la loi gaussienne de moyenne 0 et de variance 1. La fonction de répartition F de la loi $\mathcal{N}(0, 1)$ est aussi représentée (en vert).*

questions se posent dans ce cadre, et en particulier, nous nous attacherons à décrire le taux de convergence—que nous verrons être de nature non paramétrique—des différents estimateurs que nous étudierons.

I.2 Code R

Figure I.1

```
set.seed(2)
nvec=c(10,1000)
par(mfrow=c(1,length(nvec)))
par(oma=c(2,1,2,1))
par(mar=c(3,3,1,1))
for (i in (1:length(nvec)))
        {
        n=nvec[i]
        x=rnorm(n)
        plot.ecdf(x,xlim=c(-2.5,2.5),main="",col="dodgerblue2",lwd=2)
        s=seq(-3,3,.01)
        d=pnorm(s)
        lines(s,d,col="darkolivegreen4",lwd=2)
        if (i==1) {mtext(expression(paste(hat(F)[n],"(",x,")",sep="")),
                                side=2,line=2.5,cex=1)}
        mtext("x",side=1,line=3,cex=1)
        }
```

I.3 TD

1. Considérons la fonction de répartition empirique $\hat{F}_n$ définie en (I.3).
   (a) Expliquer pourquoi $n\hat{F}_n(x)$ est de loi $\mathrm{Bin}(n, F(x))$ [5].
   (b) En déduire que $\mathrm{E}[\hat{F}_n(x)] = F(x)$ et $\mathrm{Var}[\hat{F}_n(x)] = F(x)(1-F(x))/n$.
   (c) En utilisant l'identité $\mathrm{Var}[Z] = \mathrm{E}[Z^2] - (\mathrm{E}[Z])^2$, en déduire que

$$\mathrm{E}[\{\hat{F}_n(x) - F(x)\}^2] = \frac{C_{F,x}}{n},$$

   où $C_{F,x} := F(x)(1-F(x))$.
   (d) En appliquant le théorème central-limite, montrer que $\sqrt{n}\{\hat{F}_n(x) - F(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, C_{F,x})$.

2. Fixons $x \in \mathbb{R}$ et $\alpha \in (0,1)$. Soit $z_{\alpha/2}$ le quantile d'ordre $1-(\alpha/2)$ de la loi normale standard.
   (a) Déduire du point (d) de l'exercice précédent que

$$\mathcal{I}_n := \left[\hat{F}_n(x) - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}, \hat{F}_n(x) + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}\right]$$

   est un intervalle de confiance asymptotique pour $F(x)$ au niveau de confiance $1-\alpha$ (obtenir l'intervalle en faisant un raisonnement plutôt qu'en utilisant une formule toute faite).
   (b) Ecrire un programme R qui, pour chaque valeur de $n \in \{10, 20, \ldots, 200\}$, engendre $M = 100\,000$ échantillons aléatoires indépendants de taille n depuis la loi exponentielle de moyenne 1 et qui mesure la proportion p_n des échantillons pour lesquels $F(2)$ (où F est la fonction de répartition de la loi exponentielle de moyenne 1) appartient à l'intervalle de confiance asymptotique pour $F(2)$ au niveau de confiance 95%. Faire le graphe de p_n en fonction de n. Observe-t-on bien la convergence de p_n vers 95% ?

3. Cet exercice est relatif au jeu de données *quakes* de R, qui reprend certaines variables associées à des tremblements de terre (tapez `quakes` pour consulter le jeu de données et `?quakes` pour obtenir des informations).
   (a) Dans R, tracer le graphe de la fonction de répartition empirique associée à la variable *Depth*. Répéter l'exercice pour la variable *Richter magnitude*.
   (b) Pourquoi y a-t-il si peu de plateaux [6] pour cette seconde fonction de répartition en comparaison avec la première ?

5. Pour rappel, la loi $\mathrm{Bin}(n,p)$ est celle du nombre de succès observés lorsqu'on effectue n fois de façon indépendante une expérience aléatoire de type succès-échec dans laquelle un succès se produit avec probabilité p.

6. On appelle ici "plateau" un intervalle maximal où la fonction est constante.

(c) Est-ce un problème, pour l'estimation de la fonction de répartition (ou pour la construction d'un intervalle de confiance sur cette fonction de répartition en un point x fixé), que la variable *Richter magnitude* n'admette pas de densité ?

(d) Notons $x_1, \ldots, x_n$ les valeurs observées de la variable *depth*. Pour chaque $x \in [\min(x_1, \ldots, x_n), \max(x_1, \ldots, x_n)]$, on peut construire un intervalle de confiance asymptotique $[\ell_x, u_x]$ pour $F(x)$ au niveau de confiance 95%. Dans R, tracer le graphe de la fonction empirique de cette variable et y ajouter les graphes de $x \mapsto \ell_x$ et de $x \mapsto u_x$ (travailler avec une grille régulière de pas 0.5 pour x).

Chapitre II

Estimation de densité

II.1 Le problème et les mesures de risque L_2

Considérons une variable aléatoire X qui admet une fonction de densité $f : \mathbb{R} \to \mathbb{R}^+$. Rappelons que ceci signifie que la probabilité que X prenne une valeur dans un borélien B arbitraire peut être évaluée par la relation

$$P[X \in B] = \int_B f(x)\,dx.$$

En particulier, on doit avoir (i) $f(x) \geq 0$ pour tout $x \in \mathbb{R}$ et (ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Toute fonction f satisfaisant les conditions (i)–(ii) est la densité d'une variable aléatoire réelle. Dans ce chapitre, nous considérons le problème de l'estimation de f sur la base de n copies indépendantes de X, c'est-à-dire sur la base d'observations $X_1, \ldots, X_n$ qui sont mutuellement indépendantes et admettent chacune la même densité f que X.

Le plus souvent, on considérera des mesures de *risque* L_2 pour mesurer la qualité d'un estimateur $\hat{f}_n(x)$ de $f(x)$.

Definition II.1 *Le risque L_2 de $\hat{f}_n$ en x est*

$$R_x(\hat{f}_n, f) := \mathrm{E}[\{\hat{f}_n(x) - f(x)\}^2],$$

où l'espérance est calculée en utilisant le fait que $\hat{f}_n(x)$ est basée sur un échantillon d'observations aléatoires $X_1, \ldots, X_n$ indépendantes et admettant la densité f.

Ce risque L_2, qui correspond au MSE décrit dans l'introduction, est typiquement préféré au risque L_1 défini comme $\mathrm{E}[|\hat{f}_n(x) - f(x)|]$. La raison est que, contrairement au risque L_1, le risque L_2 admet une décomposition en termes de *biais* et de *variance*,

qui se révèle capitale en statistique non paramétrique : en utilisant l'identité $\mathrm{Var}[Z] = \mathrm{E}[Z^2] - (\mathrm{E}[Z])^2$, il vient

$$\begin{aligned} R_x(\hat{f}_n, f) &= \mathrm{E}[\{\hat{f}_n(x) - f(x)\}^2] \\ &= (\mathrm{E}[\hat{f}_n(x) - f(x)])^2 + \mathrm{Var}[\hat{f}_n(x) - f(x)] \\ &= \{\mathrm{E}[\hat{f}_n(x)] - f(x)\}^2 + \mathrm{Var}[\hat{f}_n(x)] \\ &= b_x^2(\hat{f}_n, f) + v_x(\hat{f}_n), \end{aligned} \tag{II.1}$$

où $b_x(\hat{f}_n, f) := \mathrm{E}[\hat{f}_n(x)] - f(x)$ est le biais de l'estimateur $\hat{f}_n(x)$ de $f(x)$ et $v_x(\hat{f}_n) := \mathrm{Var}[\hat{f}_n(x)]$ est sa variance. Puisque $b_x^2(\hat{f}_n, f) \geq 0$ et $v_x(\hat{f}_n) \geq 0$, un risque L_2 faible ne peut être obtenu qu'en réalisant un biais et une variance réduits.

Bien entendu, il est désirable que le risque L_2 tende vers zéro lorsque la taille d'échantillon n diverge vers l'infini. Par définition, ceci signifie que $\hat{f}_n(x)$ converge vers $f(x)$ en moyenne quadratique, ce qui implique que $\hat{f}_n(x)$ converge vers $f(x)$ en probabilité. Pouvoir apprendre ainsi la convergence d'un estimateur de densité est une autre motivation importante pour contrôler finement le risque L_2 de cet estimateur.

La mesure de risque L_2 ci-dessus, qui mesure la qualité de l'estimation de $f(x)$ par l'estimateur $\hat{f}_n(x)$, est adéquate si on est spécifiquement intéressé par l'estimation de f *au point* x. Le plus souvent, cependant, on désire estimer la densité f dans son ensemble, auquel cas le risque suivant est plus naturel.

Definition II.2 *Le risque L_2 intégré de $\hat{f}_n$ est*

$$R(\hat{f}_n, f) := \mathrm{E}\left[\int_{-\infty}^{\infty} \{\hat{f}_n(x) - f(x)\}^2 \, dx\right],$$

où l'espérance est calculée en utilisant le fait que $\hat{f}_n$ est basée sur un échantillon d'observations aléatoires $X_1, \ldots, X_n$ indépendantes et admettant la densité f.

On parlera parfois d'*erreur quadratique moyenne intégrée* (ou *MISE* en anglais, pour *Mean Integrated Square Error*). En permutant espérance et intégrale, il vient

$$\begin{aligned} R(\hat{f}_n, f) &= \int_{-\infty}^{\infty} \mathrm{E}[\{\hat{f}_n(x) - f(x)\}^2] \, dx \\ &= \int_{-\infty}^{\infty} R_x(\hat{f}_n, f) \, dx \\ &= \int_{-\infty}^{\infty} \{b_x^2(\hat{f}_n, f) + v_x(\hat{f}_n)\} \, dx \\ &= \int_{-\infty}^{\infty} b_x^2(\hat{f}_n, f) \, dx + \int_{-\infty}^{\infty} v_x(\hat{f}_n) \, dx, \end{aligned} \tag{II.2}$$

ce qui montre que le risque intégré $R(\hat{f}_n, f)$ est l'intégrale du risque ponctuel $R_x(\hat{f}_n, f)$ par rapport à x et qu'il hérite donc d'une décomposition en termes de biais et de variance.

A titre d'illustration dans ce chapitre, nous considérerons souvent la densité

$$f(x) = \frac{1}{2}\phi_{0,1}(x) + \frac{1}{10}\sum_{j=0}^{4} \phi_{\frac{j}{2}-1,\frac{1}{100}}(x), \tag{II.3}$$

qui est une "densité de mélange" à six composantes gaussiennes initiales (ici, ϕ_{μ,σ^2} est la densité de la loi gaussienne de moyenne μ et de variance σ^2 ; voir (I.1)). Le graphe de cette densité, appelée dans [6] la *densité de Bart Simpson*, est donné à la Figure II.1.

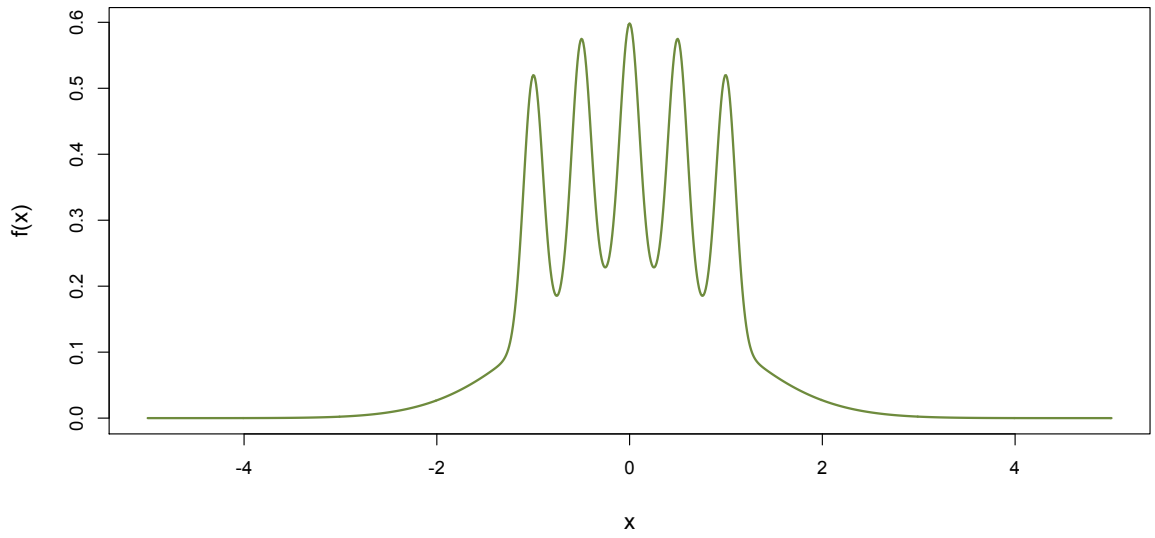

Figure II.1 – *Graphe de la densité de Bart Simpson en (II.3).*

L'approche la plus classique pour estimer f est de nature *paramétrique* et consiste à supposer que f appartient à une classe de fonctions $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ indicée par un paramètre θ de dimension finie. Par exemple, on pourrait supposer que X est de loi gaussienne, auquel cas $\mathcal{F} = \{\phi_{\mu,\sigma^2} : \binom{\mu}{\sigma^2} \in \mathbb{R} \times \mathbb{R}_0^+ \subset \mathbb{R}^2\}$ est la collection de toutes les densités gaussiennes. Bien entendu, ceci ramène l'estimation de f à l'estimation du paramètre θ, ce qui est un problème classique en statistique, pour lequel on dispose d'un large éventail de méthodes standards (méthodes du maximum du vraisemblance, méthode des moments, méthodes des moindres carrés, etc.) Dans le cas du modèle gaussien, la densité estimée par la méthode du maximum de vraisemblance sera

$$\hat{f}_n(x) = \phi_{\hat{\mu}_n,\hat{\sigma}_n^2}(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \exp\left(-\frac{(x-\hat{\mu}_n)^2}{2\hat{\sigma}_n^2}\right), \tag{II.4}$$

où $\hat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $\hat{\sigma}_n^2 = s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ sont respectivement les estimateurs du maximum de vraisemblance de μ et σ^2. Si f est effectivement gaussienne ou est proche d'une loi gaussienne, alors cet estimateur est adéquat. Si, par contre, f s'éloigne fortement d'une densité gaussienne, comme c'est le cas par exemple de la densité de Bart Simpson, alors $\hat{f}_n$ n'est pas un estimateur satisfaisant ; ceci est illustré à la Figure II.2. Si f n'est pas gaussienne, le risque intégré $R(\hat{f}_n, f)$ ne tendra pas vers zéro quand n diverge vers l'infini.

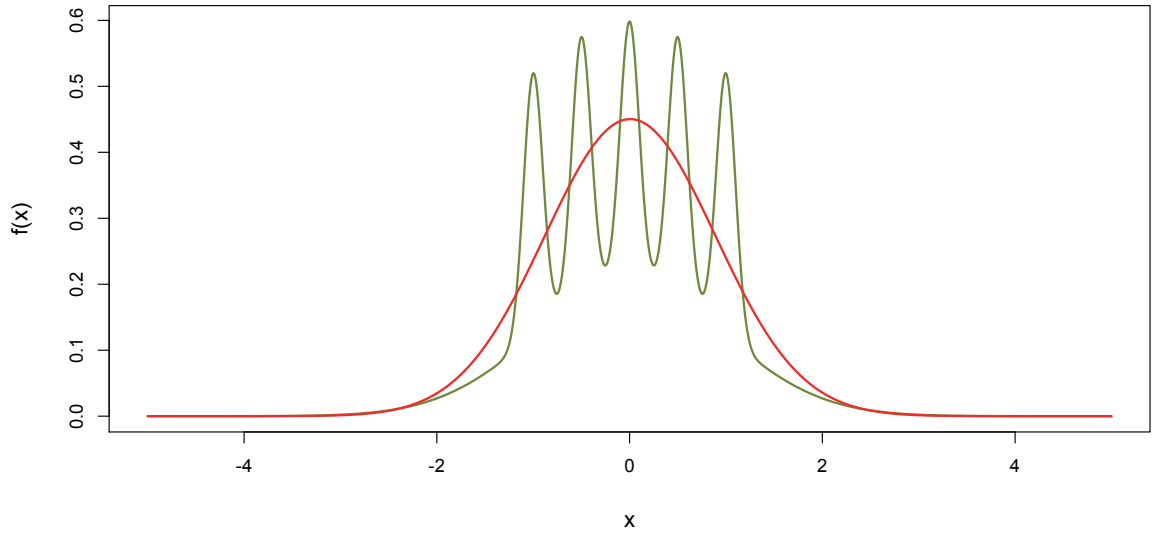

Figure II.2 – *Graphe de l'estimateur paramétrique gaussien en (II.4) (en rouge) calculé sur un échantillon aléatoire de taille* $n = 1000$ *engendré depuis la densité de Bart Simpson (en vert).*

Ceci constitue une motivation importante pour introduire des estimateurs de densité *non paramétriques*, pour lesquels le risque intégré tendra vers zéro sous des hypothèses minimes sur la densité f qu'on estime.

II.2 Estimation par histogramme

Supposons que la densité f à estimer soit nulle en dehors de l'intervalle $\mathcal{I} = [a, b]$ (si un tel intervalle $[a, b]$ n'existe pas, comme c'est le cas pour la densité de Bart Simpson, on prendra l'intervalle suffisamment grand pour que f soit presque nulle en dehors de l'intervalle ; nous renvoyons à la fin de cette section pour une autre approche). Pour un entier strictement positif m fixé, partitionnons $\mathcal{I}$ en les m sous-intervalles ou *cellules*

$$\mathcal{I}_1 = [a, a+h),\ \mathcal{I}_2 = [a+h, a+2h), \ldots,\ \mathcal{I}_m = [a+(m-1)h, a+mh = b],$$

où $h = (b-a)/m$ est la *taille de la cellule*.

Definition II.3 *L'estimateur par histogramme de f associé au nombre de cellules m est donné par*

$$\hat{f}_n(x) = \frac{1}{h} \sum_{j=1}^{m} \hat{p}_j \mathbb{I}[x \in \mathcal{I}_j], \tag{II.5}$$

où $\hat{p}_j = Y_j/n$ est basé sur le nombre $Y_j = \sum_{i=1}^{n} \mathbb{I}[X_i \in \mathcal{I}_j]$ d'observations appartenant à $\mathcal{I}_j$.

L'estimateur par histogramme en (II.5) se réécrit

$$\hat{f}_n(x) = \begin{cases} \hat{p}_1/h & \text{si } x \in \mathcal{I}_1 \\ \quad \vdots & \\ \hat{p}_m/h & \text{si } x \in \mathcal{I}_m, \end{cases}$$

ce qui implique que $\hat{f}_n$ est une fonction constante par morçeaux, et donc discontinue. La Figure II.3 montre l'estimateur $\hat{f}_n$ associé à $a = -5$, $b = 5$ et $m = 100$ calculé sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson. Clairement, pour ces valeurs des paramètres a, b et m, l'histogramme fournit une reconstruction assez satisfaisante de f.

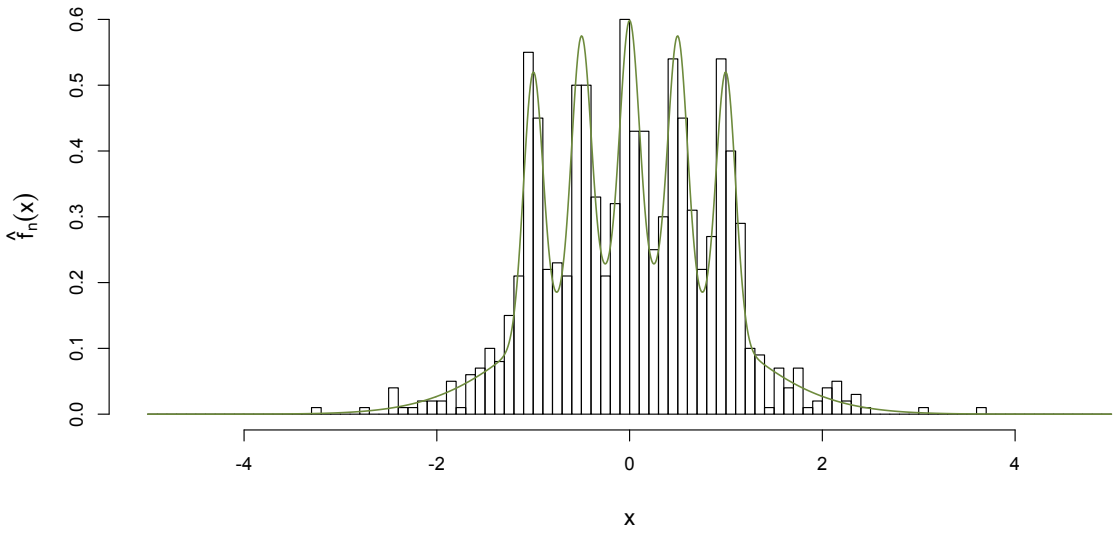

Figure II.3 – *Graphe de l'estimateur par histogramme $\hat{f}_n$ associé à $a = -5$, $b = 5$ et $m = 100$ calculé sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson (en vert).*

Nous donnons maintenant un argument heuristique suggérant que l'estimateur par histogramme est une procédure pour laquelle les risques L_2 introduits ci-dessus pour-

raient effectivement tendre vers zéro quand n diverge vers l'infini. Notons d'abord que la loi des grands nombres assure que $\hat{p}_j$ converge presque sûrement vers

$$p_j := \mathrm{E}[\mathbb{I}[X_1 \in \mathcal{I}_j]] = P[X_1 \in \mathcal{I}_j] = \int_{\mathcal{I}_j} f(x)\,dx.$$

Par ailleurs, si f est continue, alors le théorème de la moyenne (Théorème A.1) montre que, pour un certain $\theta_j \in \mathcal{I}_j$, on a $p_j/h = f(\theta_j)$, de sorte que si m diverge vers l'infini, on a $(p_j/h)\mathbb{I}[x \in \mathcal{I}_j] \to f(x)\mathbb{I}[x \in \mathcal{I}_j]$. Ceci suggère que, pour n grand,

$$\hat{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h}\mathbb{I}[x \in \mathcal{I}_j] \approx \sum_{j=1}^{m} \frac{p_j}{h}\mathbb{I}[x \in \mathcal{I}_j] \approx \sum_{j=1}^{m} f(x)\mathbb{I}[x \in \mathcal{I}_j] = f(x),$$

ce qui tendrait à montrer que $\hat{f}_n$ est un estimateur convergent de f, pour peu que $m = m_n \to \infty$. Nous insistons sur le caractère heuristique de cet argument (en particulier, l'usage de la loi des grands nombres ci-dessus requiert que m soit fixé, alors que la suite de l'argument demande que m diverge vers l'infini).

Le résultat principal pour cette section est le théorème ci-dessous, qui établit un résultat précis sur le risque intégré de l'estimateur par histogramme. Son importance réside aussi dans le fait qu'il nous permettra de préciser l'impact du paramètre-clé à choisir quand on utilise cet estimateur, à savoir le nombre m de cellules.

Théorème II.1 *Supposons que f soit nulle en dehors de $[a, b]$ et soit de classe C^2 sur $[a, b]$. Soit $m = m_n$ une suite d'entiers strictement positifs qui diverge vers l'infini et soit $h_n = (b - a)/m_n$. Alors, le risque intégré de l'estimateur par histogramme satisfait*

$$R(\hat{f}_n, f) = \frac{h_n^2}{12}\|f'\|_2^2 + \frac{1}{nh_n} + o(h_n^2) + o\Big(\frac{1}{nh_n}\Big) \tag{II.6}$$

quand n diverge vers l'infini, où $\|g\|_2 = (\int_a^b (g(x))^2\,dx)^{1/2}$ est la norme L_2 de g sur $[a, b]$.

Preuve du Théorème II.1. Dans cette preuve, on écrira respectivement h et m pour h_n et m_n pour alléger la notation. Rappelons que

$$R(\hat{f}_n, f) = \int_{-\infty}^{\infty} b_x^2(\hat{f}_n, f)\,dx + \int_{-\infty}^{\infty} v_x(\hat{f}_n)\,dx,$$

où $b_x(\hat{f}_n, f) = \mathrm{E}[\hat{f}_n(x)] - f(x)$ est le biais de l'estimateur $\hat{f}_n(x)$ et $v_x(\hat{f}_n) = \mathrm{Var}[\hat{f}_n(x)]$ est sa variance ; voir (II.2). Commençons par étudier le terme de biais.

Puisque $n\hat{p}_j = Y_j \sim \text{Bin}(n, p_j)$, on a $\text{E}[Y_j] = np_j$, donc $\text{E}[\hat{p}_j] = p_j$. Il en découle que

$$\text{E}[\hat{f}_n(x)] = \text{E}\left[\frac{1}{h}\sum_{j=1}^{m}\hat{p}_j\mathbb{I}[x \in \mathcal{I}_j]\right] = \frac{1}{h}\sum_{j=1}^{m}p_j\mathbb{I}[x \in \mathcal{I}_j],$$

ce qui fournit

$$b_x(\hat{f}_n, f) = \text{E}[\hat{f}_n(x)] - f(x) = \sum_{j=1}^{m}\Big(\frac{p_j}{h} - f(x)\Big)\mathbb{I}[x \in \mathcal{I}_j]$$

et donc

$$\int_a^b b_x^2(\hat{f}_n, f)\,dx = \sum_{j=1}^{m}\int_{\mathcal{I}_j} b_x^2(\hat{f}_n, f)\,dx = \sum_{j=1}^{m}\int_{\mathcal{I}_j}\Big(\frac{p_j}{h} - f(x)\Big)^2\,dx.$$

Si $x, y \in \mathcal{I}_j$, on a (Théorème A.3)

$$f(y) = f(x) + (y-x)f'(x) + \frac{1}{2}f''(\lambda_j)(y-x)^2$$

pour un certain $\lambda_j \in \mathcal{I}_j$, ce qui fournit (toujours pour $x \in \mathcal{I}_j$)

$$\begin{aligned}
\frac{p_j}{h} - f(x) &= \frac{1}{h}\int_{\mathcal{I}_j}(f(y) - f(x))\,dy \\
&= \frac{1}{h}\int_{\mathcal{I}_j}\Big((y-x)f'(x) + \frac{1}{2}f''(\lambda_j)(y-x)^2\Big)\,dy \\
&= \frac{f'(x)}{h}\int_{\mathcal{I}_j}(y-x)\,dy + \frac{1}{2h}\int_{\mathcal{I}_j}f''(\lambda_j)(y-x)^2\,dy \\
&= f'(x)\left(\frac{1}{h}\int_{\mathcal{I}_j}y\,dy - x\right) + O(h^2) \\
&= f'(x)(c_j - x) + O(h^2),
\end{aligned}$$

où

$$c_j = \frac{1}{2}\{(a + (j-1)h) + (a + jh)\} = a + h\Big(j - \frac{1}{2}\Big)$$

est le milieu de la cellule $\mathcal{I}_j$ et où on a utilisé le fait que

$$\left|\int_{\mathcal{I}_j}f''(\lambda_j)(y-x)^2\,dy\right| \le \int_{\mathcal{I}_j}|f''(\lambda_j)|(y-x)^2\,dy \le \int_{\mathcal{I}_j}\|f''\|_\infty h^2\,dy = \|f''\|_\infty h^3 = O(h^3)$$

($\|f''\|_\infty := \sup_{x\in[a,b]}|f''(x)| < \infty$ puisque, par hypothèse, f'' est continue sur $[a,b]$).

Puisque ceci implique que (encore pour $x \in \mathcal{I}_j$)

$$\begin{aligned}\Big(\frac{p_j}{h} - f(x)\Big)^2 &= (f'(x))^2(c_j - x)^2 + 2f'(x)(c_j - x)O(h^2) + O(h^4) \\ &= (f'(x))^2(c_j - x)^2 + O(h^3)\end{aligned}$$

(où on a utilisé le fait que $|f'(x)| \leq \|f'\|_\infty < \infty$ et que $c_j - x = O(h)$), nous obtenons, en appliquant le théorème de la moyenne (voir le Théorème A.1),

$$\begin{aligned}\int_{\mathcal{I}_j} \Big(\frac{p_j}{h} - f(x)\Big)^2 dx &= \int_{\mathcal{I}_j} (f'(x))^2(c_j - x)^2\, dx + O(h^4) \\ &= (f'(\xi_j))^2 \int_{\mathcal{I}_j} (c_j - x)^2\, dx + O(h^4) \\ &= h(f'(\xi_j))^2 \Bigg(\underbrace{\frac{1}{h}\int_{\mathcal{I}_j} (c_j - x)^2\, dx}_{=h^2/12} \Bigg) + O(h^4) \\ &= \frac{h^3}{12}(f'(\xi_j))^2 + O(h^4)\end{aligned}$$

pour un certain $\xi_j \in \mathcal{I}_j$. Pour le terme de biais, nous pouvons donc conclure que

$$\begin{aligned}\int_a^b b_x^2(\hat{f}_n, f)\, dx &= \sum_{j=1}^m \int_{\mathcal{I}_j} \Big(\frac{p_j}{h} - f(x)\Big)^2 dx \\ &= \sum_{j=1}^m \frac{h^3}{12}(f'(\xi_j))^2 + mO(h^4) \\ &= \frac{h^2}{12}\sum_{j=1}^m h(f'(\xi_j))^2 + O(h^{-1})O(h^4) \\ &= \frac{h^2}{12}\Big(\int_a^b (f'(x))^2\, dx + o(1) \Big) + O(h^3) \\ &= \frac{h^2}{12}\|f'\|_2^2 + o(h^2), \qquad \text{(II.7)}\end{aligned}$$

où on a utilisé le fait que

$$\sum_{j=1}^m h(f'(\xi_j))^2 = \sum_{j=1}^m (f'(\xi_j))^2\{(a + jh) - (a + (j-1)h)\}$$

est une somme de Riemann pour l'intégrale $\int_a^b (f'(x))^2\, dx$ et converge donc vers cette intégrale quand h tend vers zéro.

Il nous reste à traiter la variance intégrée. Notons tout d'abord que

$$v_x(\hat{f}_n) = \mathrm{Var}[\hat{f}_n(x)] = \mathrm{Var}\left[\frac{1}{h}\sum_{j=1}^{m}\hat{p}_j\mathbb{I}[x\in\mathcal{I}_j]\right] = \frac{1}{h^2}\sum_{j=1}^{m}\mathrm{Var}[\hat{p}_j]\mathbb{I}[x\in\mathcal{I}_j],$$

où on a utilisé le fait que les variables aléatoires $\hat{p}_j\mathbb{I}[x\in\mathcal{I}_j]$, $j=1,\ldots,m$ sont non-corrélées (contrairement aux variables aléatoires $\hat{p}_j$, $j=1,\ldots,m$). Puisque $n\hat{p}_j = Y_j \sim \mathrm{Bin}(n,p_j)$ implique $\mathrm{Var}[Y_j] = np_j(1-p_j)$, et donc $\mathrm{Var}[\hat{p}_j] = p_j(1-p_j)/n$, on a

$$v_x(\hat{f}_n) = \frac{1}{nh^2}\sum_{j=1}^{m}p_j(1-p_j)\mathbb{I}[x\in\mathcal{I}_j].$$

La variance intégrée s'écrit donc

$$\begin{aligned}\int_a^b v_x(\hat{f}_n)\,dx &= \sum_{j=1}^{m}\int_{\mathcal{I}_j} v_x(\hat{f}_n)\,dx = \sum_{j=1}^{m}\int_{\mathcal{I}_j}\frac{p_j(1-p_j)}{nh^2}\,dx = \sum_{j=1}^{m}\frac{p_j(1-p_j)}{nh}\\ &= \frac{1}{nh}\sum_{j=1}^{m}p_j - \frac{1}{nh}\sum_{j=1}^{m}p_j^2 = \frac{1}{nh}\underbrace{\int_a^b f(x)\,dx}_{=1} - \frac{1}{nh}\sum_{j=1}^{m}p_j^2 = \frac{1}{nh} - R_n,\end{aligned}$$

où, puisque le théorème de la moyenne assure l'existence d'un $\eta_j\in\mathcal{I}_j$ tel que

$$p_j = \int_{\mathcal{I}_j} f(x)\,dx = hf(\eta_j),$$

on a

$$R_n = \frac{1}{nh}\sum_{j=1}^{m}p_j^2 = \frac{1}{n}\sum_{j=1}^{m}hf^2(\eta_j) = \frac{1}{n}\Big(\int_a^b f^2(x)\,dx + o(1)\Big) = O\Big(\frac{1}{n}\Big) = o\Big(\frac{1}{nh}\Big).$$

Nous avons donc prouvé que

$$\int_a^b v_x(\hat{f}_n)\,dx = \frac{1}{nh} + o\Big(\frac{1}{nh}\Big),$$

ce qui, conjointement avec (II.7), établit le résultat. □

Le Théorème II.1 a de nombreuses conséquences intéressantes. Tout d'abord, il montre que, pour une densité f de classe C^2, le risque intégré $R(\hat{f}_n, f)$ de l'estimateur par histogramme tend vers zéro si et seulement si $m = m_n$ mène à

$$\text{(i) } h_n\to 0 \quad \text{et} \quad \text{(ii) } \frac{1}{nh_n}\to 0$$

(ou, formulé en termes de la suite (m_n), si et seulement si (i) $m_n \to \infty$ et (ii) $m_n/n \to 0$). Ceci indique que (i) le nombre de cellules doit bien entendu diverger vers l'infini pour obtenir un estimateur convergent, mais que (ii) ce nombre de cellules ne peut pas diverger trop vite vers l'infini. Le point (ii) est également raisonnable : pour que l'estimation de la masse de probabilité dans chaque cellule se fasse de façon convergente (par le mécanisme de la loi des grands nombres), il faut que le nombre d'observations par cellule diverge vers l'infini plus vite que le nombre m_n de cellules (si l'inverse se produit, les cellules seront finalement vides quand n diverge vers l'infini, ce qui exclut qu'on puisse estimer la densité dans chaque cellule).

La preuve du Théorème II.1 révèle que

$$R(\hat{f}_n, f) = \underbrace{\frac{h_n^2}{12}\|f'\|_2^2 + o(h_n^2)}_{=\int_a^b b_x^2(\hat{f}_n, f)\,dx} + \underbrace{\frac{1}{nh_n} + o\Big(\frac{1}{nh_n}\Big)}_{=\int_a^b v_x(\hat{f}_n)\,dx},$$

ce qui indique quels termes du risque intégré contribuent au (carré du) biais (intégré) et quels termes contribuent à la variance (intégrée). Le terme dominant dans le biais est nul si et seulement si $f'(x) = 0$ pour tout $x \in [a, b]$, c'est-à-dire si et seulement si f est constante sur $[a, b]$, ce qui s'interprète aisément. De façon plus importante, le biais, qui est de la forme $c_f h_n^2 + o(h_n^2)$, peut être rendu petit en choisissant h_n petit (de façon équivalente, m_n grand), mais ceci rendra grande la variance, qui est en $d_f/(nh_n) + o(1/(nh_n))$. Au contraire, la variance peut être rendue petite en prenant h_n grand (de façon équivalente, m_n petit), mais ceci rendra le biais grand. Minimiser le risque intégré doit donc se faire en réalisant un *équilibre entre biais et variance* (en anglais, *bias-variance trade-off*).

Pour illustrer cette dépendance en h_n (ou m_n) des contributions de type biais et de type variance au risque intégré, la Figure II.4 graphe, pour le même échantillon de taille $n = 1000$ que celui ayant mené à la Figure II.3, les estimateurs par histogramme associés à $m = 20$, $m = 100$ et $m = 500$. Clairement, l'estimateur associé à $m = 20$ montre un biais plus grand et une variance plus petite que l'estimateur associé à $m = 100$, tandis que l'estimateur associé à $m = 500$ montre au contraire un biais plus petit et une variance plus grande que l'estimateur associé à $m = 100$. Ceci est en adéquation avec les résultats théoriques obtenus ci-dessus.

Pour montrer que le risque intégré peut être minimisé en choisissant $m = m_n$ de façon adéquate, nous avons conduit l'expérience empirique suivante. Pour chaque taille d'échantillon $n \in \{300, 600, 1000, 2000\}$ et chaque nombre de cellules $m \in \{10, 20, 30, \ldots, 590, 600\}$, nous avons estimé le risque intégré $R(\hat{f}_n, f)$, où $\hat{f}_n$ est l'estimateur de f fondé sur m cellules et f est la densité de Bart Simpson (nous avons pris $a = -5$ et $b = 5$).

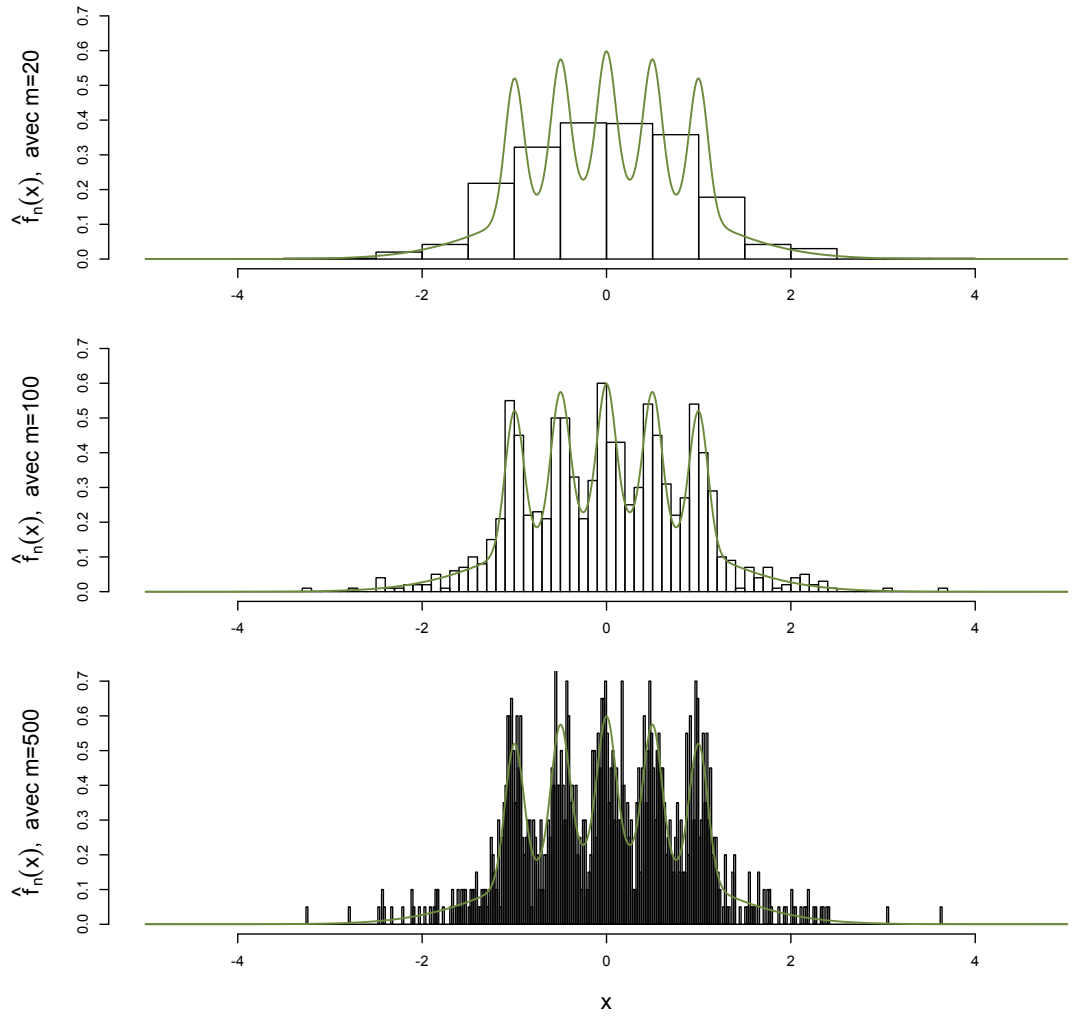

Figure II.4 – *Graphes de l'estimateur par histogramme $\hat{f}_n$ associé à $a = -5$, $b = 5$ et $m = 20$ (haut), $m = 100$ (milieu), et $m = 500$ (bas) calculé sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson (en vert).*

Plus précisément, sur la base de $S = 10000$ échantillons indépendants $(X_{1,s}, \ldots, X_{n,s})$, $s = 1, \ldots, S$, engendrés depuis f, le risque estimé que nous avons considéré est

$$R_{\text{est}}(\hat{f}_n, f) := \frac{1}{S} \sum_{s=1}^{S} \left[\frac{1}{200} \sum_{r=1}^{200} \left\{ \hat{f}_{n,s}\Big(-5 + \frac{r-1}{100} \Big) - f\Big(-5 + \frac{r-1}{100} \Big) \right\}^2 \right], \qquad \text{(II.8)}$$

où $\hat{f}_{n,s}$ désigne l'estimateur par histogramme de f calculé sur $X_{1,s}, \ldots, X_{n,s}$ (la somme en r est une somme de Riemann, qui approxime l'intégrale $\int_{-5}^{5} \{\hat{f}_{n,s}(x) - f(x)\}^2 \, dx$, tandis que la somme sur s approxime l'espérance mathématique de cette intégrale, qui est bien le risque intégré). La Figure II.5 montre, pour les différentes tailles d'échantillon n considérées, le graphe de $R_{\text{est}}(\hat{f}_n, f)$ en fonction de m. La figure met clairement en évidence l'existence d'un $m = m_n$ minimisant le risque intégré en réalisant un équilibre

biais-variance.

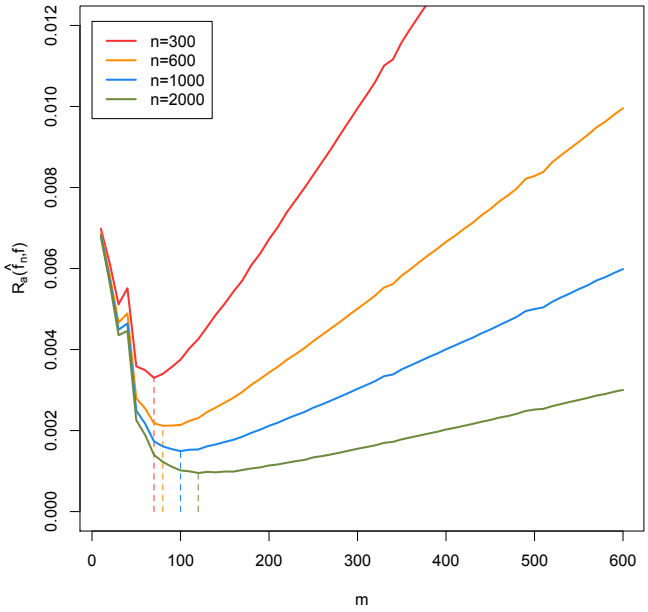

Figure II.5 – *Pour différentes tailles d'échantillon n, graphes, en fonction de m, de l'estimation $R_{\text{est}}(\hat{f}_n, f)$ du risque intégré associé à l'estimateur par histogramme fondé sur m cellules lorsqu'on estime, sur base d'un échantillon de taille n, la densité de Bart Simpson. Pour chaque taille d'échantillon, la valeur de m fournissant le risque minimal est mise en évidence par un trait en pointillés.*

Bien entendu, nous ne pouvons pas, en pratique, identifier la valeur optimale de m sur la base d'un graphe tel que celui de la Figure II.5, car celui-ci ne peut être réalisé que si f est connue (ce qui n'est évidemment pas le cas). Par contre, on peut tenter d'identifier la valeur de m qui minimise le risque intégré apparaissant dans le Théorème II.1. Ceci fournit le résultat suivant (la démonstration est proposée dans les exercices du cours).

Théorème II.2 *Sous les conditions du Théorème II.1, la valeur de $h = h_n$ qui minimise le risque intégré $R(\hat{f}_n, f)$ est*

$$h_n^{\text{opt}} = C_f \, n^{-1/3},$$

où $C_f := (\frac{1}{6}\|f'\|_2^2)^{-1/3}$ (la valeur optimale correspondante de $m = m_n$ est donc $m_n^{\text{opt}} = (b-a)/h_n^{\text{opt}}$). Le risque intégré minimal associé est de la forme

$$R^{\text{opt}}(\hat{f}_n, f) = D_f \, n^{-2/3} + o(n^{-2/3}),$$

quand n diverge vers l'infini, où $D_f := (\frac{9}{16}\|f'\|_2^2)^{1/3}$.

Ce résultat montre en particulier que, lorsqu'il est fondé sur la valeur optimale de m_n

(de façon équivalente, de h_n), le risque intégré de l'estimateur par histogramme converge vers zéro à la vitesse $n^{-2/3}$. Ceci est à comparer à la vitesse de convergence du risque (c'est-à-dire l'erreur quadratique moyenne) obtenu lorsqu'on estime un paramètre de dimension finie, qui, comme on l'a rappelé dans l'introduction, tend typiquement vers zéro à vitesse n^{-1}. La perte de vitesse de convergence (de n^{-1} à $n^{-2/3}$) peut être vue comme le coût d'estimer une densité de façon non paramétrique.

Malheureusement, le Théorème II.2 ne permet pas non plus de déterminer le nombre de cellules optimal m_n^{opt}, car la constante C_f dépend de f et est donc inconnue. Il existe cependant des méthodes complexes, telles que les méthodes de *validation croisée*, qui permettent de déterminer une valeur de m adéquate. Nous n'entrerons pas dans ces détails pour l'estimateur par histogramme car nous présenterons dans la section suivante un autre estimateur qui domine l'estimateur par histogramme en termes de risque intégré, au sens où le risque intégré de ce nouvel estimateur tendra vers zéro plus vite que $n^{-2/3}$.

Pour terminer cette section, nous revenons sur l'hypothèse que la densité f est nulle en dehors de l'intervalle $[a, b]$ choisi pour construire l'histogramme. Cette hypothèse est requise par les Théorèmes II.1–II.2, qui ne s'appliquent donc pas, par exemple, pour la densité de Bart Simpson. La théorie peut être étendue relativement aisément à une partition de la droite réelle du type

$$\ldots,\ \mathcal{I}_{-1} = [x_0 - 2h, x_0 - h),\ \mathcal{I}_0 = [x_0 - h, x_0),\ \mathcal{I}_1 = [x_0, x_0 + h),\ \mathcal{I}_2 = [x_0 + h, x_0 + 2h), \ldots$$

faisant encore intervenir une taille de cellule h mais un nombre infini de cellules. Si le paramètre-clé est encore le paramètre de lissage h, il faut ici choisir aussi le paramètre x_0, qui peut également avoir un impact sur l'estimation par histogramme ; nous renvoyons à [2] pour une discussion de ce point et pour certaines solutions possibles.

II.3 Estimation par noyau

En dépit de son utilisation courante, l'estimateur par histogramme souffre d'un certain nombre de défauts importants. En particulier, il prend comme valeur une fonction discontinue, et ce même si la densité f qu'il estime est très lisse. Par ailleurs, comme annoncé ci-dessus, il ne fournit pas le plus petit risque intégré possible, même lorsque le nombre de cellules utilisé a été choisi de façon optimale. Enfin, il ne permet pas de construire aisément des intervalles de confiance pour $f(x)$. Ceci fournit une motivation importante pour présenter et étudier les *estimateurs à noyau*, qui ne souffriront pas des défauts précités.

II.3.1 Définition

Les estimateurs à noyau sont construits en exploitant le fait que, en tout point x, la fonction de densité f est la dérivée de la fonction de répartition correspondante $x \mapsto F(x) = P[X \leq x]$ au point x, et satisfait donc

$$f(x) = F'(x) = \lim_{h \searrow 0} \frac{F(x+h) - F(x-h)}{2h}. \tag{II.9}$$

L'idée naturelle consiste à définir un estimateur $\hat{f}_n(x)$ de $f(x)$ en remplaçant F dans (II.9) par la fonction de répartition empirique $\hat{F}_n$, mais ceci est exclu car $\hat{F}_n$ n'est pas dérivable (ceci mènerait à un estimateur $\hat{f}_n$ qui serait nul partout sauf en un nombre fini de valeurs de x pour lesquelles $\hat{f}_n(x)$ ne serait pas défini).

Par contre, (II.9) implique que, pour $h > 0$ petit, on a

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}, \tag{II.10}$$

ce qui suggère d'estimer $f(x)$ par

$$\begin{aligned}
\hat{f}_n(x) &:= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\
&= \frac{1}{2h}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \leq x+h] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \leq x-h]\right) \\
&= \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{I}[x-h < X_i \leq x+h] \\
&= \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big),
\end{aligned}$$

où h est petit et où $z \mapsto K(z) := \frac{1}{2}\mathbb{I}[-1 \leq z < 1]$ est la fonction de densité de la loi uniforme sur $[-1, 1)$ (ou $[-1, 1]$). Si cet estimateur $\hat{f}_n$ va hériter du caractère discontinu de la fonction de densité K ci-dessus, on peut penser à remplacer K par une autre densité *continue*, ce qui rendra continu l'estimateur $\hat{f}_n$ associé.

Definition II.4 *Un estimateur à noyau de f est un estimateur de la forme*

$$\hat{f}_n(x) := \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \tag{II.11}$$

où K est un noyau, c'est-à-dire une fonction de $\mathbb{R}$ dans $\mathbb{R}$ telle que (i) $\int_{-\infty}^{\infty} K(z)\,dz = 1$, (ii) $\int_{-\infty}^{\infty} zK(z)\,dz = 0$ et (iii) $\|K\|_2^2 := \int_{-\infty}^{\infty} K^2(z)\,dz < \infty$.

Outre le *noyau uniforme* ci-dessus, les noyaux les plus communément utilisés sont présentés à la Figure II.6. On vérifie directement que, quelles que soient les valeurs prises par les variables aléatoires $X_1, \ldots, X_n$, on a

$$\int_{-\infty}^{\infty} \hat{f}_n(x)\, dx = 1.$$

Par ailleurs, si $K(z) \geq 0$ pour tout z, alors on a trivialement $\hat{f}_n(x) \geq 0$ pour tout x, auquel cas $\hat{f}_n$ est donc une fonction de densité (ce qui est naturel, car on estime alors une fonction de densité par une fonction de densité). Comme on le verra par la suite, utiliser un noyau K négatif par endroit peut cependant présenter certains avantages en termes d'efficacité. Finalement, comme indiqué ci-dessus, si le noyau K est continu, alors $\hat{f}_n$ l'est également ; plus généralement, $\hat{f}_n$ hérite de la régularité (C^k, C^∞, etc.) du noyau K utilisé. Ceci est illustré à la Figure II.7, qui représente, pour le noyau gaussien et le noyau uniforme, les estimateurs à noyau obtenus pour un échantillon de taille $n = 1000$ engendré depuis la densité de Bart Simpson. Le noyau gaussien fournit un estimateur à noyau lisse, tandis que le noyau uniforme fournit un estimateur à noyau discontinu.

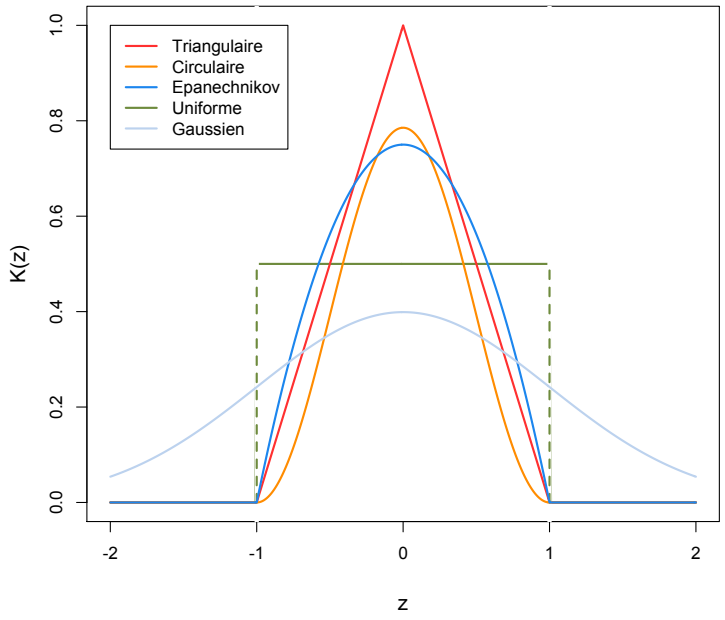

Figure II.6 – *Graphes des noyau triangulaire* $(1 - |z|)\mathbb{I}[-1 \leq z \leq 1]$, *noyau circulaire* $\frac{\pi}{4}\cos(\frac{\pi}{2}z)\mathbb{I}[-1 \leq z \leq 1]$, *noyau d'Epanechnikov* $\frac{3}{4}(1 - z^2)\mathbb{I}[-1 \leq z \leq 1]$, *noyau uniforme* $\frac{1}{2}\mathbb{I}[-1 \leq z < 1]$, *et noyau gaussien* $K(z) = (2\pi)^{-1/2}\exp(-z^2/2)$.

Il devrait être clair que pour que $\hat{f}_n$ soit un estimateur convergent, il faut non seulement que le nombre d'observations n diverge vers l'infini, mais aussi qu'on utilise une *taille de fenêtre* $h = h_n$ qui tende vers zéro (par valeurs plus grandes, puisque h doit

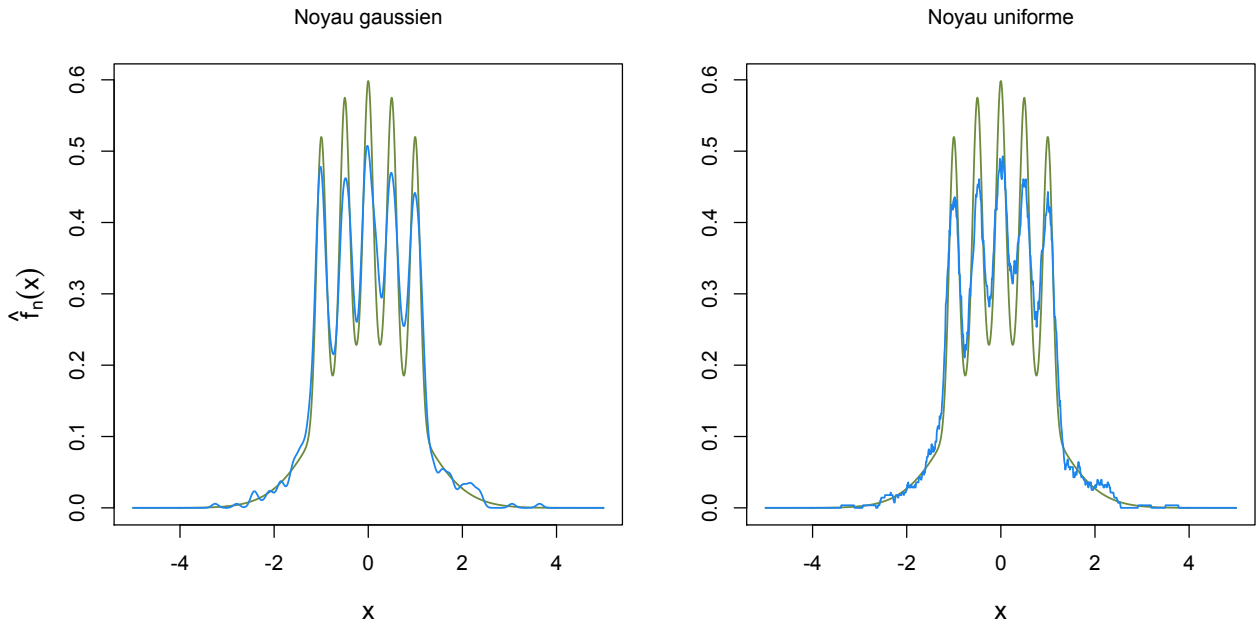

Figure II.7 – *Graphes de l'estimateur à noyau $\hat{f}_n$ calculé sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson, en utilisant le noyau gaussien et $h = .07$ (gauche) ou le noyau uniforme et $h = .14$ (droite).*

être strictement positif). Cette convergence ne peut pas avoir lieu de manière arbitraire, comme le suggère l'argument heuristique suivant. Supposons, pour simplifier, que le noyau utilisé soit à support compact, et, plus spécifiquement, que $\{z \in \mathbb{R} : K(z) > 0\} = (-c, c)$, disons. Dans l'estimation de $f(x)$ par $\hat{f}_n(x)$, l'observation X_i va alors apporter une contribution (au sens où le $i$ème terme du membre de droite de (II.11) va prendre une valeur strictement positive) si et seulement si

$$\left|\frac{x - X_i}{h_n}\right| < c,$$

c'est-à-dire si et seulement si $X_i \in (x - ch_n, x + ch_n)$. Une observation X_i est donc "active" si et seulement si elle prend sa valeur dans une "fenêtre" centrée en x et de largeur $2ch_n$, ce qui explique la terminologie *taille de la fenêtre*. Si f est continue en x, alors le nombre moyen d'observations actives est

$$\mathrm{E}\left[\sum_{i=1}^{n} \mathbb{I}[X_i \in (x - ch_n, x + ch_n)]\right] = \sum_{i=1}^{n} P[X_i \in (x - ch_n, x + ch_n)]$$

$$= nP[X_1 \in (x - ch_n, x + ch_n)] = n \int_{x-ch_n}^{x+ch_n} f(z)\, dz = 2cnh_n(f(x) + o(1)),$$

quand n diverge vers l'infini (la dernière égalité est obtenue en utilisant le théorème de la

moyenne puis la continuité de f en x). Clairement, une estimation convergente requiert que ce nombre moyen d'observations actives diverge vers l'infini, ce qui impose que nh_n diverge vers l'infini. La taille de la fenêtre doit ainsi converger vers zéro, mais pas trop vite. Que ceci est une condition nécessaire pour la convergence de $\hat{f}_n(x)$ sera confirmé théoriquement ci-dessous.

II.3.2 Risques ponctuel et intégré

Contrairement à l'estimateur par histogramme, l'estimateur à noyau permet d'obtenir de façon simple une expression du risque L_2 ponctuel.

Théorème II.3 *Supposons que f soit de classe C^3 sur $\mathbb{R}$ et que $\|f'\|_\infty$ et $\|f'''\|_\infty$ soient finies (où $\|g\|_\infty := \sup_{x\in\mathbb{R}} |g(x)|$). Soit K un noyau tel que $\int_{-\infty}^{\infty} |z|^3 |K(z)|\, dz < \infty$ et $\int_{-\infty}^{\infty} |z| K^2(z)\, dz < \infty$. Soit $h = h_n$ une suite qui est $o(1)$ et telle que nh_n diverge vers l'infini. Soit $\hat{f}_n$ l'estimateur à noyau de f associé à K et à h_n. Posons $d_K := \int_{-\infty}^{\infty} z^2 K(z)\, dz$ et $c_K := \|K\|_2^2$. Alors, pour tout $x \in \mathbb{R}$, on a*

$$b_x(\hat{f}_n, f) = \frac{h_n^2}{2} f''(x) d_K + o(h_n^2) \quad \textit{et} \quad v_x(\hat{f}_n) = \frac{1}{nh_n} f(x) c_K + o\Big(\frac{1}{nh_n}\Big),$$

quand n diverge vers l'infini, de sorte que

$$R_x(\hat{f}_n, f) = \frac{h_n^4}{4} (f''(x))^2 d_K^2 + \frac{1}{nh_n} f(x) c_K + o(h_n^4) + o\Big(\frac{1}{nh_n}\Big), \tag{II.12}$$

quand n diverge vers l'infini.

Preuve du Théorème II.3. Puisque les observations X_i sont indépendantes et admettent la densité f, on obtient

$$\begin{aligned}
\mathrm{E}[\hat{f}_n(x)] &= \mathrm{E}\Big[\frac{1}{nh_n} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_n}\Big)\Big] = \frac{1}{nh_n} \sum_{i=1}^{n} \mathrm{E}\Big[K\Big(\frac{x - X_i}{h_n}\Big)\Big] \\
&= \frac{1}{h_n} \mathrm{E}\Big[K\Big(\frac{x - X_1}{h_n}\Big)\Big] = \frac{1}{h_n} \int_{-\infty}^{\infty} K\Big(\frac{x - z}{h_n}\Big) f(z)\, dz = \int_{-\infty}^{\infty} K(y) f(x - h_n y)\, dy,
\end{aligned}$$

où on a effectué le changement de variable $z = x - h_n y$. Par conséquent, le biais de $\hat{f}_n(x)$ est

$$b_x(\hat{f}_n, f) = \mathrm{E}[\hat{f}_n(x)] - f(x) = \int_{-\infty}^{\infty} K(y) \{f(x - h_n y) - f(x)\}\, dy, \tag{II.13}$$

où on a utilisé le fait que $\int_{-\infty}^{\infty} K(z)\,dz = 1$. Un développement de Taylor incluant la formule de Cauchy pour le reste (Théorème A.3) fournit

$$f(x - h_n y) = f(x) - h_n y f'(x) + \frac{h_n^2}{2} y^2 f''(x) - \frac{h_n^3}{6} y^3 f'''(\theta_{n,x,y}),$$

pour un certain $\theta_{n,x,y}$ entre x et $x - h_n y$. En utilisant l'identité $\int_{-\infty}^{\infty} yK(y)\,dy = 0$, ceci fournit

$$\begin{aligned} b_x(\hat{f}_n, f) &= \int_{-\infty}^{\infty} K(y)\Big\{ - h_n y f'(x) + \frac{h_n^2}{2} y^2 f''(x) - \frac{h_n^3}{6} y^3 f'''(\theta_{n,x,y})\Big\}\,dy \\ &= -h_n f'(x) \int_{-\infty}^{\infty} yK(y)\,dy + \frac{h_n^2}{2} f''(x) \int_{-\infty}^{\infty} y^2 K(y)\,dy - \frac{h_n^3}{6} \int_{-\infty}^{\infty} y^3 f'''(\theta_{n,x,y}) K(y)\,dy \\ &= \frac{h_n^2}{2} f''(x) d_K + o(h_n^2), \end{aligned} \tag{II.14}$$

puisque

$$\left| \frac{h_n^3}{6} \int_{-\infty}^{\infty} y^3 f'''(\theta_{n,x,y}) K(y)\,dy \right| \le \frac{h_n^3}{6} \|f'''\|_\infty \int_{-\infty}^{\infty} |y|^3 |K(y)|\,dy = O(h_n^3) = o(h_n^2).$$

Passons à la variance de $\hat{f}_n(x)$. On a

$$\begin{aligned} v_x(\hat{f}_n) = \mathrm{Var}[\hat{f}_n(x)] &= \frac{1}{n^2 h_n^2} \mathrm{Var}\Big[\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_n}\Big)\Big] = \frac{1}{n h_n^2} \mathrm{Var}\Big[K\Big(\frac{x - X_1}{h_n}\Big)\Big] \\ &= \frac{1}{n h_n^2} \left\{ \mathrm{E}\Big[K^2\Big(\frac{x - X_1}{h_n}\Big)\Big] - \Big(\mathrm{E}\Big[K\Big(\frac{x - X_1}{h_n}\Big)\Big]\Big)^2 \right\} \\ &= \frac{1}{n h_n^2} \mathrm{E}\Big[K^2\Big(\frac{x - X_1}{h_n}\Big)\Big] - \frac{1}{n} \Big(\frac{1}{h_n} \mathrm{E}\Big[K\Big(\frac{x - X_1}{h_n}\Big)\Big]\Big)^2 \\ &= \frac{1}{n h_n^2} \mathrm{E}\Big[K^2\Big(\frac{x - X_1}{h_n}\Big)\Big] + O\Big(\frac{1}{n}\Big), \end{aligned}$$

où a utilisé le fait que

$$\frac{1}{n} \Big(\frac{1}{h_n} \mathrm{E}\Big[K\Big(\frac{x - X_1}{h_n}\Big)\Big]\Big)^2 = \frac{1}{n} (\mathrm{E}[\hat{f}_n(x)])^2 = \frac{1}{n} (f(x) + o(1))^2 = \frac{1}{n} O(1) = O\Big(\frac{1}{n}\Big).$$

En procédant comme pour le biais, on obtient alors

$$\begin{aligned}\frac{1}{nh_n^2}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)\Big] &= \frac{1}{nh_n^2}\int_{-\infty}^{\infty} K^2\Big(\frac{x-z}{h_n}\Big)f(z)\,dz = \frac{1}{nh_n}\int_{-\infty}^{\infty} K^2(y)f(x-h_ny)\,dy \\ &= \frac{1}{nh_n}\int_{-\infty}^{\infty} K^2(y)f(x)\,dy + \frac{1}{nh_n}\int_{-\infty}^{\infty} K^2(y)\{f(x-h_ny)-f(x)\}\,dy \\ &= \frac{1}{nh_n}f(x)c_K - \frac{1}{n}\int_{-\infty}^{\infty} yK^2(y)f'(\lambda_{n,x,y})\,dy \\ &= \frac{1}{nh_n}f(x)c_K + O\Big(\frac{1}{n}\Big),\end{aligned}$$

puisque

$$\left|\frac{1}{n}\int_{-\infty}^{\infty} yK^2(y)f'(\lambda_{n,x,y})\,dy\right| \leq \frac{1}{n}\|f'\|_\infty\int_{-\infty}^{\infty}|y|K^2(y)\,dy = O\Big(\frac{1}{n}\Big).$$

On peut donc conclure que

$$v_x(\hat{f}_n) = \frac{1}{nh_n}f(x)c_K + O\Big(\frac{1}{n}\Big) = \frac{1}{nh_n}f(x)c_K + o\Big(\frac{1}{nh_n}\Big),$$

ce qui, en utilisant (II.1) et (II.14), fournit finalement

$$R_x(\hat{f}_n,f) = b_x^2(\hat{f}_n,f) + v_x(\hat{f}_n) = \frac{h_n^4}{4}(f''(x))^2d_K^2 + o(h_n^4) + \frac{1}{nh_n}f(x)c_K + o\Big(\frac{1}{nh_n}\Big).$$

Le résultat est démontré. □

Le Théorème II.3 permet d'appréhender l'impact de la taille de la fenêtre h_n sur les propriétés de biais et de variance de l'estimateur à noyau. Clairement, il faut choisir h_n petit pour obtenir un petit biais, mais cela conduira à une grande variance ; au contraire, il faut prendre h_n grand pour obtenir une petite variance, mais cela mènera à un biais important. Ce comportement en h_n, qui est illustré à la Figure II.8, suggère qu'il ne faut choisir h_n ni trop petit ni trop grand pour obtenir un risque optimal, réalisant un équilibre entre biais et variance. Ceci est confirmé à la Figure II.9, qui graphe un risque (intégré) estimé en fonction de h_n pour les estimateurs à noyau fondés sur un noyau gaussien et un noyau uniforme. Il est à noter que le risque n'est pas nécessairement convexe en h_n, comme le montre l'exemple du noyau uniforme.

Un calcul direct montre que le risque ponctuel en (II.12) est minimisé lorsque

$$h_n = \left(\frac{f(x)c_K}{(f''(x))^2d_K^2}\right)^{1/5}n^{-1/5}, \tag{II.15}$$

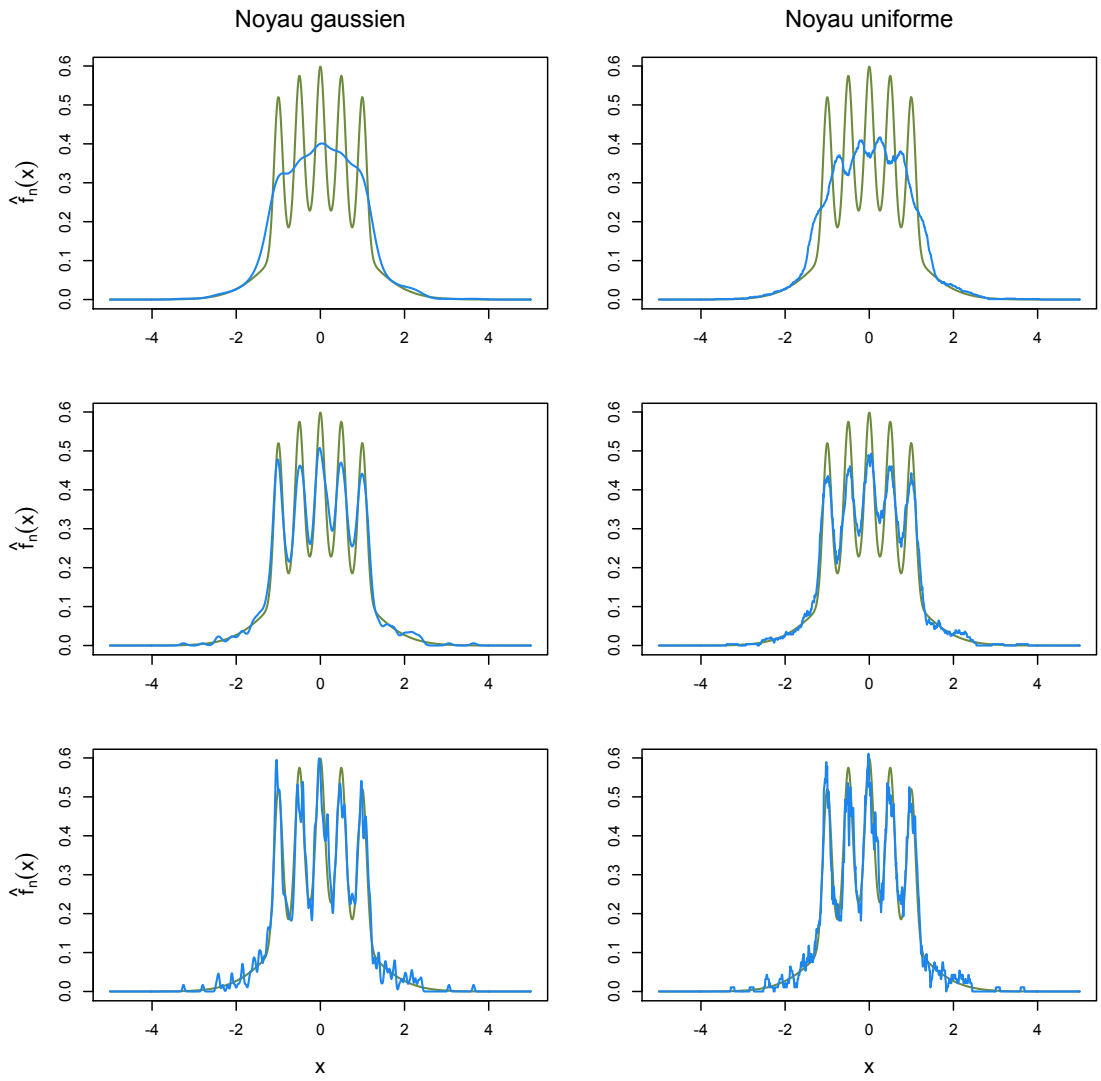

Figure II.8 – *(Gauche :) Graphes des estimateurs à noyau $\hat{f}_n$ calculés sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson, en utilisant le noyau gaussien et, pour $h_0 = .07$, une taille de fenêtre $h = 3h_0$ (haut), $h = h_0$ (milieu) et $h = h_0/3$ (bas). (Droite :) Les graphes correspondants obtenus en utilisant le noyau uniforme et, pour $h_0 = .14$, une taille de fenêtre $h = 3h_0$ (haut), $h = h_0$ (milieu) et $h = h_0/3$ (bas).*

ce qui mène au risque ponctuel minimal

$$R_x(\hat{f}_n, f) = \frac{5}{4}\Big(f^2(x)f''(x)c_K^2 d_K\Big)^{2/5} n^{-4/5} + o(n^{-4/5}). \tag{II.16}$$

Bien entendu, la formule (II.15) ne permet pas de choisir h_n en pratique puisque $f(x)$ et $f''(x)$ sont inconnus. Le problème de la sélection de h_n sera discuté à la Section II.3.3.

Nous considérons maintenant le risque intégré de l'estimateur à noyau, qui nous permettra de comparer les performances de cet estimateur avec celles de l'estimateur par histogramme.

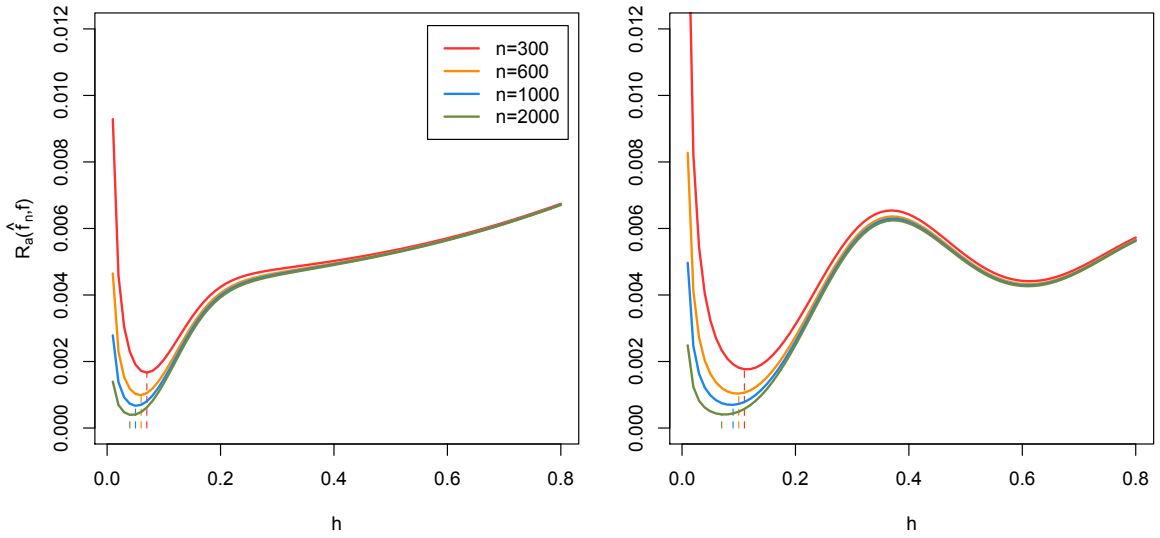

Figure II.9 – *Pour différentes tailles d'échantillon n, graphes, en fonction de h, de l'estimation $R_{\text{est}}(\hat{f}_n, f)$ (voir (II.8)) du risque intégré associé à l'estimateur à noyau utilisant la taille de fenêtre h et le noyau gaussien (gauche) ou le noyau uniforme (droite) lorsqu'on estime, sur base d'un échantillon de taille n, la densité de Bart Simpson. Pour chaque taille d'échantillon, la valeur de h fournissant le risque minimal est mise en évidence par un trait en pointillés.*

Théorème II.4 *Supposons que f soit de classe C^2 sur $\mathbb{R}$ et que $\|f''\|_2 < \infty$. Soit K un noyau à valeurs non-négatives et tel que $\int_{-\infty}^{\infty} z^2 K(z)\,dz < \infty$. Soit $\hat{f}_n$ l'estimateur à noyau de f associé à K et à h. Posons $d_K := \int_{-\infty}^{\infty} z^2 K(z)\,dz$ et $c_K := \|K\|_2^2$. Alors, (i) on a*

$$R(\hat{f}_n, f) \leq \frac{h^4}{3}\|f''\|_2^2 d_K^2 + \frac{1}{nh} c_K,$$

(ii) Par conséquent, si $h = h_n = \xi n^{-1/5}$, on a, avec $\mathcal{F}_M := \{f : \|f''\|_2 \leq M\}$,

$$\sup_{f \in \mathcal{F}_M} R(\hat{f}_n, f) \leq C n^{-4/5}, \tag{II.17}$$

pour une constante C qui ne dépend que de ξ, de K et de M.

Preuve du Théorème II.4. (i) Commençons par la variance. On a

$$v_x(\hat{f}_n) = \mathrm{Var}[\hat{f}_n(x)] = \frac{1}{nh^2}\mathrm{Var}\left[K\Big(\frac{x - X_1}{h}\Big)\right] \leq \frac{1}{nh^2}\mathrm{E}\left[K^2\Big(\frac{x - X_1}{h}\Big)\right]$$
$$= \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\Big(\frac{x - z}{h}\Big) f(z)\,dz = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(y) f(x - hy)\,dy.$$

Donc le théorème de Fubini fournit

$$\begin{aligned}\int_{-\infty}^{\infty} v_x(\hat{f}_n)\,dx &\le \frac{1}{nh}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K^2(y) f(x-hy)\,dy\,dx \\ &= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(y)\underbrace{\int_{-\infty}^{\infty} f(x-hy)\,dx}_{=1}\,dy \\ &= \frac{1}{nh}c_K. \end{aligned} \tag{II.18}$$

Pour le biais, on a, en procédant comme en (II.13), puis en appliquant le Théorème A.2(ii),

$$\begin{aligned} b_x(\hat{f}_n, f) = \mathrm{E}[\hat{f}_n(x)] - f(x) &= \int_{-\infty}^{\infty} K(y)\{f(x-hy)-f(x)\}\,dy \\ &= \int_{-\infty}^{\infty} K(y)\Big\{-hyf'(x) + h^2y^2\int_0^1 (1-s)f''(x-shy)\,ds\Big\}\,dy \\ &= h^2\int_{-\infty}^{\infty}\int_0^1 y^2K(y)(1-s)f''(x-shy)\,dsdy, \end{aligned}$$

où on a utilisé l'identité $\int_{-\infty}^{\infty} yK(y)\,dy = 0$. En utilisant l'inégalité de Cauchy–Schwarz, il vient alors

$$\begin{aligned} b_x^2(\hat{f}_n, f) &= h^4\left(\int_{-\infty}^{\infty}\int_0^1 \Big\{y\sqrt{K(y)}\Big\}\Big\{y\sqrt{K(y)}(1-s)f''(x-shy)\Big\}\,dsdy\right)^2 \\ &\le h^4\left(\int_{-\infty}^{\infty}\int_0^1 y^2K(y)\,dsdy\right)\left(\int_{-\infty}^{\infty}\int_0^1 y^2K(y)(1-s)^2(f''(x-shy))^2\,dsdy\right) \\ &= h^4 d_K\left(\int_{-\infty}^{\infty}\int_0^1 y^2K(y)(1-s)^2(f''(x-shy))^2\,dsdy\right). \end{aligned}$$

Donc, en intégrant sur $\mathbb{R}$ par rapport à x et en utilisant le théorème de Fubini, on obtient

$$\begin{aligned}\int_{-\infty}^{\infty} b_x^2(\hat{f}_n, f)\,dx &\le h^4 d_K\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^1 y^2K(y)(1-s)^2(f''(x-shy))^2\,ds\,dy\,dx \\ &= h^4 d_K\int_{-\infty}^{\infty} y^2K(y)\int_0^1(1-s)^2\underbrace{\int_{-\infty}^{\infty}(f''(x-shy))^2\,dx}_{=\|f''\|_2^2}\,ds\,dy \\ &= h^4 d_K\|f''\|_2^2\left(\int_{-\infty}^{\infty} y^2K(y)\,dy\right)\left(\int_0^1(1-s)^2\,ds\right) \\ &= \frac{h^4}{3}d_K^2\|f''\|_2^2. \end{aligned}$$

Conjointement avec (II.18), ceci établit donc

$$R(\hat{f}_n, f) = \int_{-\infty}^{\infty} b_x^2(\hat{f}_n, f)\, dx + \int_{-\infty}^{\infty} v_x(\hat{f}_n)\, dx \leq \frac{h^4}{3} d_K^2 \|f''\|_2^2 + \frac{1}{nh} c_K.$$

(ii) Pour $h = h_n = \xi n^{-1/5}$ et pour tout $f \in \mathcal{F}_M$, le risque intégré satisfait donc

$$R(\hat{f}_n, f) \leq \Big(\frac{\xi^4}{3} d_K^2 \|f''\|_2^2 + \frac{c_K}{\xi}\Big) n^{-4/5} \leq \Big(\frac{\xi^4}{3} d_K^2 M^2 + \frac{c_K}{\xi}\Big) n^{-4/5},$$

ce qui démontre le résultat. □

Le Théorème II.4(ii) indique que le risque intégré de l'estimateur à noyau tend vers zéro à la vitesse $n^{-4/5}$, qui est plus rapide que la vitesse $n^{-2/3}$ réalisée par l'estimateur par histogramme (voir le Théorème II.2). L'estimateur à noyau jouit donc d'un taux de convergence meilleur. Une question naturelle est celle de l'optimalité de ce taux de convergence. En d'autres termes, existe-t-il un estimateur de densité pour lequel le risque intégré convergerait plus vite vers zéro que $n^{-4/5}$? Comme l'indique le résultat suivant, la réponse est négative.

Théorème II.5 *Il existe une constante $D = D_M$ telle que, avec $\mathcal{F}_M := \{f : \|f''\|_2 \leq M\}$, on ait*

$$\sup_{f \in \mathcal{F}_M} R(\tilde{f}_n, f) \geq D n^{-4/5},$$

quel que soit l'estimateur de densité $\tilde{f}_n$.

Nous ne présentons pas la preuve de ce résultat, qui est très technique (le lecteur intéressé pourra trouver une preuve dans [5] ; voir le Théorème 24.4). Ce résultat, qui garantit que les estimateurs à noyau sont optimaux en termes de taux de convergence, explique, bien entendu, le succès important de ces estimateurs.

Le Théorème II.5 peut essentiellement être lu en disant que si on se restreint à la collection des densités de classe C^2, alors le taux de convergence optimal est $n^{-4/5}$. On peut montrer que si on se restreint à des collections de densités plus régulières, le taux de convergence optimal s'améliore. On a le résultat suivant.

Théorème II.6 *Supposons que f soit de classe C^k sur $\mathbb{R}$ et que $\|f^{(k)}\|_2 < \infty$ (où $f^{(k)}$ désigne la $k$ème dérivée de f). Soit K un noyau tel que $\int_{-\infty}^{\infty} z^r K(z)\, dz = 0$ pour tout $r = 1, \ldots, k-1$ et $\int_{-\infty}^{\infty} |z|^k K(z)\, dz < \infty$. Soit $\hat{f}_n$ l'estimateur à noyau de f associé à K et à h. Alors, (i) on a*

$$R(\hat{f}_n, f) \leq C\Big(h^{2k} + \frac{1}{nh}\Big),$$

pour une certain constante C, de sorte que si $h_n = \xi n^{-1/(2k+1)}$, on a, avec $\mathcal{F}_{k,M} := \{f : \|f^{(k)}\|_2 \leq M\}$,

$$\sup_{f \in \mathcal{F}_{k,M}} R(\hat{f}_n, f) \leq C n^{-2k/(2k+1)}. \tag{II.19}$$

(ii) Il existe une constante $D = D_M$ telle qu'on ait

$$\sup_{f \in \mathcal{F}_{k,M}} R(\tilde{f}_n, f) \geq D n^{-2k/(2k+1)},$$

quel que soit l'estimateur de densité $\tilde{f}_n$.

Pour les densités de classe C^k, le taux de convergence optimal est donc $n^{-2k/(2k+1)}$, qui devient arbitrairement proche du taux de convergence paramétrique n^{-1} quand k diverge vers l'infini. Pour $k > 2$, réaliser le taux de convergence $n^{-2k/(2k+1)}$ requiert un noyau *d'ordre supérieur*, au sens où K doit annuler des intégrales du type $\int_{-\infty}^{\infty} z^r K(z)\, dz = 0$ pour $r \geq 2$. Ceci nécessite bien entendu que K soit négatif par endroit, ce qui pourrait, dans certains cas, donner lieu à des estimateurs à noyau $\hat{f}_n$ qui sont aussi négatifs par endroit.

II.3.3 Choix de K et de h_n

Estimer f par un estimateur à noyau demande de choisir deux quantités : le noyau K et la taille de la fenêtre $h = h_n$. Dans cette section, nous décrivons des procédures qui permettent d'effectuer ces choix, en particulier pour le paramètre crucial h_n.

Choix de K

Considérons d'abord le choix du noyau K. Comme on l'a vu en (II.16), le risque ponctuel associé à l'estimateur à noyau utilisant le noyau K et la taille de fenêtre h_n optimale est donné par

$$R_x(\hat{f}_n, f) = \frac{5}{4}\Big(f^2(x) f''(x) c_K^2 d_K\Big)^{2/5} n^{-4/5} + r_n, \tag{II.20}$$

où r_n est un terme de reste. Il est important de noter que quand on construit un estimateur à noyau, les noyaux $K(z)$ et $K_\sigma(z) := \frac{1}{\sigma} K(\frac{z}{\sigma})$ sont équivalents pour tout $\sigma > 0$: si on prend $h_\sigma = h/\sigma$, on a en effet

$$\begin{aligned}
\hat{f}_n^{K_\sigma, h_\sigma}(x) &= \frac{1}{n h_\sigma} \sum_{i=1}^{n} K_\sigma\Big(\frac{x - X_i}{h_\sigma}\Big) \\
&= \frac{1}{n h_\sigma \sigma} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_\sigma \sigma}\Big) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) = \hat{f}_n^{K,h}(x),
\end{aligned}$$

ce qui montre que l'estimateur à noyau $\hat{f}_n$ fondé sur (K, h) coïncide avec celui fondé sur (K_σ, h_σ). Plutôt qu'un noyau K, c'est donc une classe de noyaux K, définis à une échelle σ près, qu'il convient de choisir. En ligne avec ceci, on vérifiera facilement que le facteur dépendant du noyau dans le terme principal du risque en (II.20) satisfait $c_{K_\sigma}^2 d_{K_\sigma} = c_K^2 d_K$. Par conséquent, il n'y a pas de perte de généralité à supposer que l'échelle de K a été fixée de sorte que $d_K = 1$. En se restreignant à de tels noyaux, le risque ci-dessus s'écrit

$$R_x(\hat{f}_n, f) = \frac{5}{4}\Big(f^2(x)f''(x)c_K^2\Big)^{2/5} n^{-4/5} + r_n.$$

Choisir le noyau de façon optimale consiste donc à minimiser c_K dans la classe des noyaux satisfaisant $d_K = 1$, ce qui mène au problème variationnel visant à minimiser $(c_K =) \int_{-\infty}^{\infty} K^2(z)\,dz$ dans la classe des fonctions $K : \mathbb{R} \to \mathbb{R}$ satisfaisant $\int_{-\infty}^{\infty} K(z)\,dz = 1$, $\int_{-\infty}^{\infty} zK(z)\,dz = 0$ et $(d_K =) \int_{-\infty}^{\infty} z^2 K(z)\,dz = 1$. Les outils du calcul variationnel permettent de montrer que le noyau optimal est donné par

$$K(z) = \frac{3}{4\sqrt{5}}\Big(1 - \frac{z^2}{5}\Big)\mathbb{I}[-\sqrt{5} \le z \le \sqrt{5}],$$

qui, à un facteur d'échelle près, est le noyau d'Epanechnikov. On peut donc conclure que le noyau d'Epanechnikov est optimal.

Au vu de ce résultat d'optimalité, il peut sembler étonnant, à première vue, qu'il soit de pratique courante d'utiliser d'autres noyaux (comme le noyau gaussien par exemple). La raison est que, si le noyau d'Epanechnikov fournira effectivement le plus petit risque, le gain en termes de risque reste minimal. Pour le montrer, revenons à (II.20), qui rend clair que l'impact du noyau sur le risque ponctuel est mesuré par le facteur $(c_K^2 d_K)^{2/5}$. Par conséquent, le bénéfice à utiliser le noyau d'Epanechnikov $K = K_{\text{Epan}}$ plutôt qu'un noyau K peut être mesuré par le ratio

$$\frac{(c_{K_{\text{Epan}}}^2 d_{K_{\text{Epan}}})^{2/5}}{(c_K^2 d_K)^{2/5}}.$$

Ce ratio prend les valeurs 0.989 pour le noyau triangulaire, 0.999 pour le noyau circulaire, 0.943 pour le noyau uniforme, et 0.961 pour le noyau gaussien. Le gain en termes de risque à utiliser le noyau d'Epanechnikov plutôt que le noyau gaussien est donc de moins de 4%, ce qui explique qu'en pratique, on préfèrera souvent les estimateurs très lisses qui résultent de l'utilisation du noyau gaussien.

Choix de h_n

Si le noyau d'Epanechnikov est optimal, le gain par rapport aux autres noyaux est marginal, et le choix du noyau n'est donc pas prépondérant. Au contraire, le choix de h_n est crucial car il a un impact énorme sur l'estimateur à noyau $\hat{f}_n$. Il est donc capital de disposer d'une méthode permettant de faire ce choix en pratique.

La plupart des méthodes disponibles dans la littérature visent à choisir h_n de façon à minimiser le risque intégré $R(\hat{f}_n, f)$. Si le Théorème II.4 ne donne qu'une borne supérieure sur ce risque, il est facile de montrer que, sous les hypothèses du Théorème II.3,

$$R(\hat{f}_n, f) = \frac{h_n^4}{4}\|f''\|_2^2 d_K^2 + \frac{1}{nh_n}c_K + o(h_n^4) + o\Big(\frac{1}{nh_n}\Big), \tag{II.21}$$

quand n diverge vers l'infini. On vérifie directement que la taille de fenêtre qui minimise ce risque intégré est donnée par

$$h_n = \left(\frac{c_K}{\|f''\|_2^2 d_K^2}\right)^{1/5} n^{-1/5}. \tag{II.22}$$

Le fait que f soit inconnue empêche d'utiliser cette formule en pratique. Nous présentons deux méthodes permettant de faire le choix de h_n.

(A) La première méthode consiste à se comporter comme si la densité f inconnue était gaussienne, de moyenne μ et de variance σ^2, disons. Dans ce cas, on vérifie que $\|f''\|_2^2 = 3/(8\sqrt{\pi}\sigma^5)$, ce qui mène à estimer (II.22) par

$$h_n = \hat{\sigma}\left(\frac{8\sqrt{\pi}c_K}{3d_K^2}\right)^{1/5} n^{-1/5},$$

où $\hat{\sigma}$ est un estimateur de σ. Pour le noyau gaussien, ceci donne

$$h_n \approx 1.06\hat{\sigma}n^{-1/5}.$$

Il est naturel d'estimer σ par l'écart-type empirique $\hat{\sigma}$, où $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, avec $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Néanmoins, cet estimateur est peu robuste, au sens où des valeurs extrêmes vont résulter en une sur-estimation de σ, et donc en une taille de fenêtre trop grande. Il est alors courant d'utiliser

$$\tilde{\sigma} = \min\Big(\hat{\sigma}, \frac{\mathrm{IR}}{1.34}\Big),$$

où IR désigne la longueur de l'intervale inter-quartile, c'est-à-dire la différence entre les quantiles d'ordre 75% et d'ordre 25% de $X_1, \ldots, X_n$ (le facteur 1.34 assure que IR/1.34

estime bien l'écart-type population si la loi sous-jacente est effectivement normale). La taille de fenêtre finale qui en résulte est donc

$$h_n \approx 1.06 \min\Big(\hat{\sigma}, \frac{\mathrm{IR}}{1.34}\Big) n^{-1/5}. \tag{II.23}$$

Cette procédure repose explicitement sur le fait que f est normale, ce qui est incompatible avec l'approche non paramétrique qui est adoptée dans ce cours. Elle donnera de bons résultats si la densité inconnue n'est pas trop différente d'une loi normale. Les densités visées par cette méthode sont donc typiquement les densités unimodales, symétriques et dont les queues sont relativement légères. Des raffinements classiques de cette méthode visent à adopter une approche de type *plug-in*, ce qui consiste à remplacer dans (II.22) la quantité $\|f''\|_2$ par un estimateur non paramétrique convergent.

(B) La seconde méthode est connue sous le nom de *validation croisée* (en anglais, *cross-validation*). En notant par $\hat{f}_{n,h}$ l'estimateur à noyau utilisant la taille de fenêtre h, elle vise à choisir la valeur de h qui minimise l'*erreur quadratique intégrée* (en anglais, *Integrated Square Error*)

$$\begin{aligned}
\mathrm{ISE}_n(h) &= \int_{-\infty}^{\infty} \{\hat{f}_{n,h}(x) - f(x)\}^2 \, dx \\
&= \int_{-\infty}^{\infty} \hat{f}^2_{n,h}(x) \, dx - 2\int_{-\infty}^{\infty} \hat{f}_{n,h}(x) f(x) \, dx + \int_{-\infty}^{\infty} f^2(x) \, dx,
\end{aligned}$$

ou, de manière équivalente (puisque le dernier terme ne dépend pas de h),

$$\mathrm{ISE}_{n0}(h) = \int_{-\infty}^{\infty} \hat{f}^2_{n,h}(x) \, dx - 2\int_{-\infty}^{\infty} \hat{f}_{n,h}(x) f(x) \, dx =: T_{n1} - 2T_{n2}. \tag{II.24}$$

En principe, T_{n1} peut être évalué sur la base des observations $X_1, \ldots, X_n$, mais il sera plus commode de réécrire ce terme sous une forme qui ne demande pas d'évaluer une intégrale. Pour ce faire, on écrit

$$\begin{aligned}
T_{n1} &= \int_{-\infty}^{\infty} \hat{f}^2_{n,h}(x) \, dx = \int_{-\infty}^{\infty} \Big\{ \frac{1}{(nh)^2} \sum_{i,j=1}^{n} K\Big(\frac{x - X_i}{h}\Big) K\Big(\frac{x - X_j}{h}\Big) \Big\} \, dx \\
&= \frac{1}{(nh)^2} \sum_{i,j=1}^{n} \int_{-\infty}^{\infty} K\Big(\frac{x - X_i}{h}\Big) K\Big(\frac{x - X_j}{h}\Big) \, dx,
\end{aligned}$$

ce qui, en faisant le changement de variable $x = X_i - hz$ donne

$$T_{n1} = \frac{1}{n^2 h} \sum_{i,j=1}^{n} \int_{-\infty}^{\infty} K\Big(\frac{X_i - X_j}{h} - z\Big) K(-z) \, dz = \frac{1}{n^2 h} \sum_{i,j=1}^{n} (K \star \bar{K})\Big(\frac{X_i - X_j}{h}\Big),$$

où $(K \star L)(x) := \int_{-\infty}^{\infty} K(x-z)L(z)\,dz$ est la convolution des fonctions K, L, et où $\bar{K}(z) := K(-z)$ (on peut vérifier que pour le noyau gaussien, on a $(K \star \bar{K})(z) = \frac{1}{2\sqrt{\pi}} \exp(-z^2/4)$, tandis que pour le noyau uniforme, on a $(K \star \bar{K})(z) = \frac{1}{4}\{(2+z)\mathbb{I}[-2 < z < 0] + (2-z)\mathbb{I}[0 \leq z < 2]\}$). La quantité T_{n2} fait par contre intervenir la densité f inconnue et doit donc être estimée. Pour ce faire, notons que

$$T_{n2} = \int_{-\infty}^{\infty} \hat{f}_{n,h}(x) f(x)\,dx = \mathrm{E}[\hat{f}_{n,h}(X) | X_1, \ldots, X_n],$$

où X est indépendant de $X_1, \ldots, X_n$ et admet aussi la densité f. Par conséquent, en notant $\hat{f}_{n,h}^{(-i)}$ l'estimateur à noyau fondé sur $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ et la taille de fenêtre h, on peut estimer cette quantité par

$$\begin{aligned} \hat{T}_{n2} := \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{n,h}^{(-i)}(X_i) &= \frac{1}{n}\sum_{i=1}^{n} \Big\{ \frac{1}{(n-1)h} \sum_{\substack{j \neq i \\ j=1}}^{n} K\Big(\frac{X_i - X_j}{h}\Big) \Big\} \\ &= \frac{1}{n(n-1)h} \sum_{\substack{i \neq j \\ i,j=1}}^{n} K\Big(\frac{X_i - X_j}{h}\Big). \end{aligned}$$

La principe de validation croisée consiste donc à choisir la taille de fenêtre

$$\hat{h}_n^{\mathrm{CV}} = \arg\min_{h>0} \widehat{\mathrm{ISE}}_{n0}(h), \tag{II.25}$$

où

$$\widehat{\mathrm{ISE}}_{n0}(h) := \frac{1}{n^2 h}\sum_{i,j=1}^{n} (K \star \bar{K})\Big(\frac{X_i - X_j}{h}\Big) - \frac{2}{n(n-1)h} \sum_{\substack{i \neq j \\ i,j=1}}^{n} K\Big(\frac{X_i - X_j}{h}\Big).$$

De façon remarquable, il a été prouvé (voir [3]) que, si la densité f est bornée, alors

$$\frac{\widehat{\mathrm{ISE}}_n(\hat{h}_n^{\mathrm{CV}})}{\min_{h>0} \mathrm{ISE}_n(h)} \to 1$$

presque sûrement quand n diverge vers l'infini. Ceci indique que la procédure de validation croisée livre une taille de fenêtre qui fournit une erreur quadratique intégrée qui coïncide asymptotiquement avec l'erreur quadratique intégrée minimale.

La Figure II.10 illustre la procédure pour le noyau gaussien et pour le noyau uniforme. Pour deux tailles d'échantillon différentes, les graphes de $h \mapsto \widehat{\mathrm{ISE}}_{n0}(h)$ sont tracés et les minima correspondants sont mis en évidence. On notera que ces minima sont assez proches de ceux obtenus à la Figure II.9. Pour la plus grande des deux tailles d'échantillon considérées, les estimateurs à noyau associés à ces tailles de fenêtre obtenues par validation croisée sont également représentés à la Figure II.10.

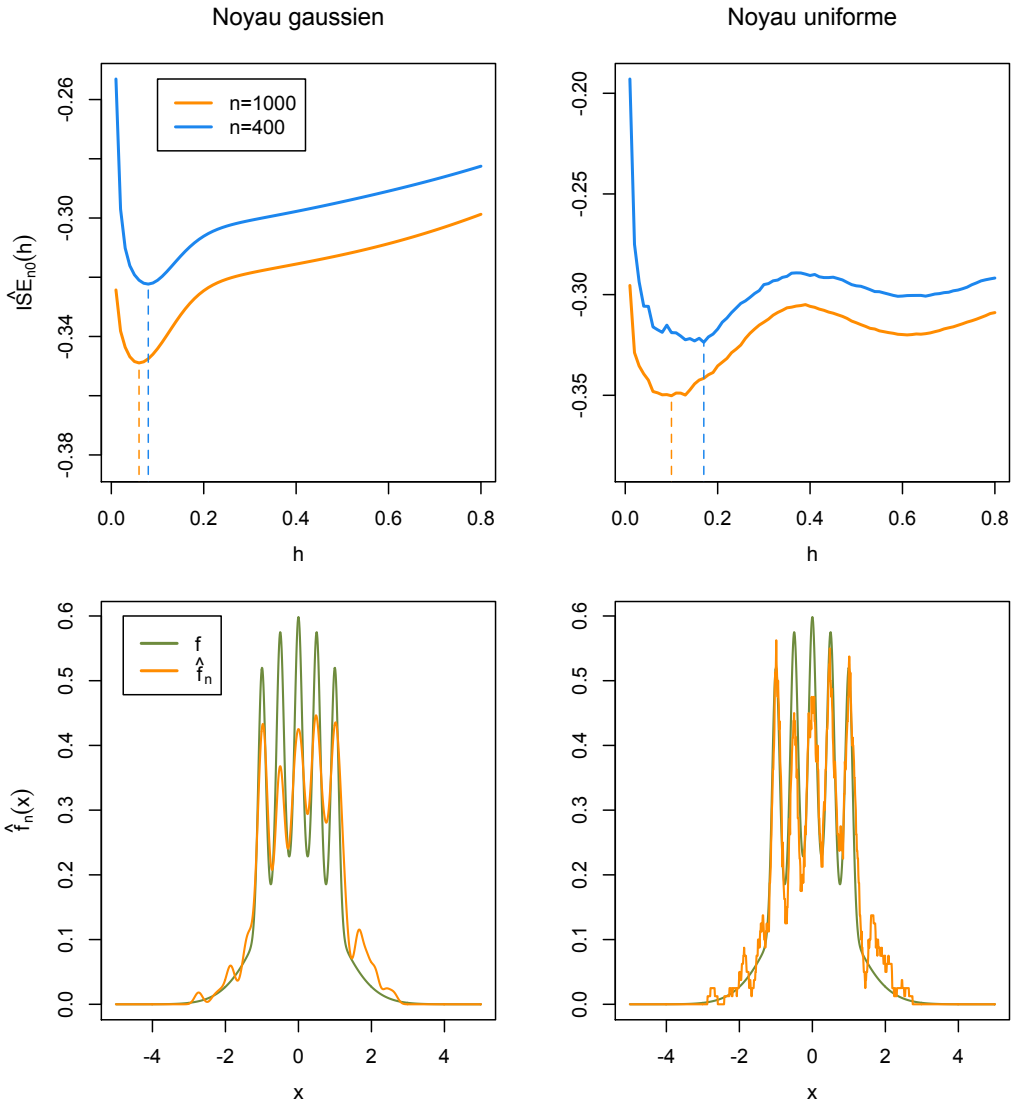

Figure II.10 – *(Haut, gauche :) Pour deux tailles d'échantillon n, graphes de $h \mapsto \widehat{\mathrm{ISE}}_{n0}(h)$ (avec le noyau gaussien) calculé sur la base d'un échantillon de taille n engendré depuis la densité de Bart Simpson. Pour chaque taille d'échantillon, la valeur de $\hat{h}_n^{\mathrm{CV}}$ est mise en évidence par un trait en pointillés. (Bas, gauche :) graphe de l'estimateur à noyau qui en résulte pour $n = 1000$ et graphe de la densité de Bart Simpson. (Haut et bas, droite :) les résultats correspondants pour le noyau uniforme.*

II.3.4 Intervalles de confiance ponctuels

On peut construire un intervalle de confiance pour $f(x)$ en obtenant un résultat de normalité asymptotique pour $\hat{f}_n(x)$. Ecrivons

$$\sqrt{nh_n}\{\hat{f}_n(x) - f(x)\} = \sqrt{nh_n v_x(\hat{f}_n)}\, Z_n + \sqrt{nh_n}\, b_x(\hat{f}_n, f), \tag{II.26}$$

où

$$Z_n := \frac{\hat{f}_n(x) - \mathrm{E}[\hat{f}_n(x)]}{\sqrt{v_x(\hat{f}_n)}}.$$

En appliquant le théorème central-limite de Levy–Lindeberg, on peut montrer que Z_n est asymptotiquement normal standard. Si on considère l'ordre optimal de la taille de fenêtre, c'est-à-dire si on prend $h_n = c_n n^{-1/5}$ avec $c_n \to c > 0$ (que se passe-t-il si on prend une taille de fenêtre tendant vers zéro plus vite ? Ou tendant vers zéro moins vite ?), alors (II.26) livre, en utilisant le Théorème II.3,

$$\sqrt{c_n} n^{2/5} \{\hat{f}_n(x) - f(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(\frac{c^{5/2}}{2} f''(x) d_K, f(x) c_K\Big),$$

ou, de façon équivalente,

$$n^{2/5} \{\hat{f}_n(x) - f(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(\frac{c^2}{2} f''(x) d_K, \frac{f(x) c_K}{c}\Big).$$

Comme on peut montrer que

$$\hat{f}_n''(x) = \frac{1}{n h_n^3} \sum_{i=1}^{n} K''\Big(\frac{x - X_i}{h_n}\Big)$$

est un estimateur convergent de $f''(x)$, on en déduit que

$$\frac{n^{2/5} \{\hat{f}_n(x) - f(x)\} - \frac{c^2}{2} \hat{f}_n''(x) d_K}{\sqrt{\frac{\hat{f}_n(x) c_K}{c}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

En procédant comme d'habitude, on obtient alors qu'un intervalle de confiance asymptotique pour $f(x)$ au niveau de confiance $1 - \alpha$ est l'intervalle dont les extrémités sont

$$\hat{f}_n^{\pm\alpha}(x) := \hat{f}_n(x) \pm n^{-2/5} \left\{ \frac{c^2}{2} \hat{f}_n''(x) d_K + z_{\alpha/2} \sqrt{\frac{\hat{f}_n(x) c_K}{c}} \right\},$$

où z_β est le quantile d'ordre $1-\beta$ de la loi normale standard. Une illustration est donnée à la Figure II.11. Nous insistons sur le fait que ces intervalles de confiance sont *ponctuels en chaque x* ; il est donc incorrect de considérer qu'il y a une probabilité asymptotiquement égale à $1 - \alpha$ que le graphe de la densité f soient compris dans la bande délimitée par les graphes des fonctions $\hat{f}_n^{\pm\alpha}$. La construction d'un tel intervalle de confiance *global* est beaucoup plus compliquée.

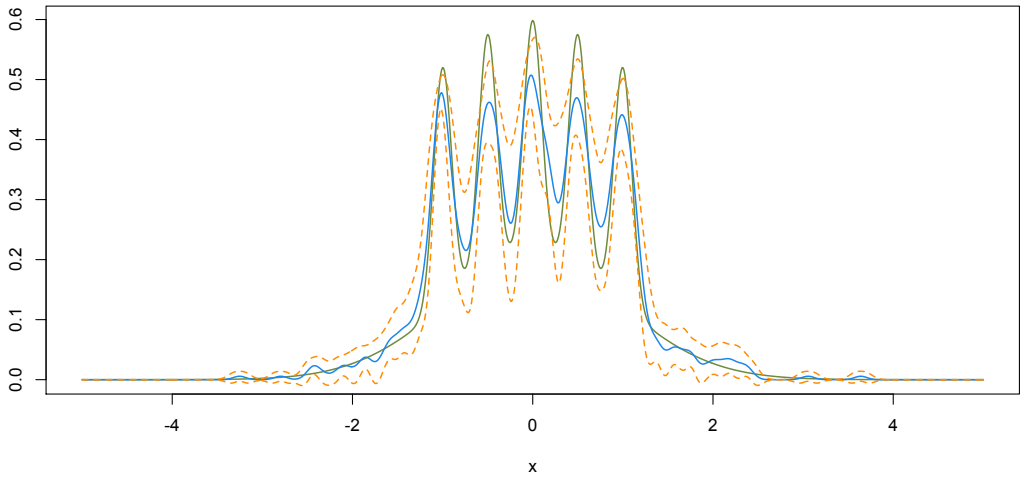

Figure II.11 – *Graphes de l'estimateur à noyau $\hat{f}_n$ (bleu) calculé sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson (vert), en utilisant le noyau gaussien et $h = .07$, ainsi que les graphes des extrémités supérieure et inférieure $\hat{f}_n^{\pm\alpha}$ (orange) déterminant les intervalles des confiance ponctuels au niveau de confiance* 95% *(la valeur de $c = c_n$ fournissant la taille de la fenêtre $h_n = cn^{-1/5} = .07$ a été utilisée ici).*

II.4 Estimation par projection

Dans cette section, nous décrivons une troisième méthode d'estimation de densité, pour laquelle nous rentrerons moins dans les détails que pour les deux premières. Le cadre fonctionnel pour cette méthode est l'espace $L_2([a,b])$ des fonctions de carré intégrable ($\int_a^b f^2(x)\,dx < \infty$) définies sur $[a,b]$. On rappelle que, muni du produit scalaire

$$\langle f, g\rangle := \int_a^b f(x)g(x)\,dx$$

(et donc de la norme associée $\|f\|_2 := \sqrt{\langle f, f\rangle}$), cet ensemble de fonctions est un espace de Hilbert (c'est-à-dire un espace vectoriel complet qui est équipé d'un produit scalaire). Considérons un sous-espace $\mathcal{S}_m$ de dimension m (finie) de $L_2([a,b])$ et soit $\{g_1, \ldots, g_m\}$ une base orthonormée de $\mathcal{S}_m$. Par définition, tout élément ℓ de $\mathcal{S}_m$ s'écrit donc

$$\ell = \sum_{r=1}^{m} c_r g_r$$

et on a $\langle g_r, g_s\rangle = \delta_{rs}$, où δ_{rs} est le delta de Kronecker (qui vaut 1 si ses deux indices sont égaux et 0 sinon).

II.4.1 Principe général

Considérons le problème de l'estimation d'une densité $f \in L_2([a,b])$. La méthode d'estimation par projection procède en deux étapes : (1) approximer f par sa projection

$$f_m := \arg\min_{h \in \mathcal{S}_m} \|f - h\|_2$$

dans $\mathcal{S}_m$; f_m est la projection orthogonale de f sur $\mathcal{S}_m$, c'est-à-dire l'élément de $\mathcal{S}_m$ le plus proche de f au sens de la distance L_2 (la théorie des espaces de Hilbert garantit l'existence et l'unicité de f_m et montre que f_m satisfait $\langle f - f_m, h\rangle = 0$ pour tout $h \in \mathcal{S}_m$). (2) Etant un élément de $\mathcal{S}_m$, cette approximation f_m de f s'écrit sous la forme

$$f_m = \sum_{r=1}^{m} c_r g_r, \tag{II.27}$$

pour certains coefficients réels $c_1, \ldots, c_m$. L'estimateur par projection $\hat{f}_n$ est alors

$$\hat{f}_n = \sum_{r=1}^{m} \hat{c}_r g_r,$$

où $\hat{c}_r$ est un estimateur de c_r fondé sur les observations $X_1, \ldots, X_n$ disponibles.

Pour mettre cette méthode en oeuvre, on doit décrire les estimateurs $\hat{c}_r$, $r = 1, \ldots, m$. En utilisant le caractère orthonormé de la base de $\mathcal{S}_m$ considérée, on obtient

$$\begin{aligned} c_r = \sum_{s=1}^{m} c_s \delta_{rs} = \sum_{s=1}^{m} c_s \langle g_r, g_s\rangle = \Big\langle g_r, \sum_{s=1}^{m} c_s g_s \Big\rangle = \langle g_r, f_m\rangle \\ = \langle g_r, f\rangle - \langle g_r, f - f_m\rangle = \langle g_r, f\rangle = \int_a^b g_r(x) f(x)\, dx = \mathrm{E}[g_r(X)], \end{aligned}$$

où X est une variable aléatoire admettant la densité f. Par la loi des grands nombres, on peut donc estimer c_r de façon convergente par

$$\hat{c}_r := \frac{1}{n} \sum_{i=1}^{n} g_r(X_i). \tag{II.28}$$

L'estimateur de f final est donc

$$\hat{f}_n = \sum_{r=1}^{m} \hat{c}_r g_r,$$

fondé sur les $\hat{c}_r$ en (II.28).

Bien entendu, cet estimateur requiert de choisir la base $\{g_1, \ldots, g_m\}$ utilisée (et donc

le sous-espace $\mathcal{S}_m$). Deux (familles de) bases classiques sont les *bases d'histogrammes réguliers* et les *bases trigonométriques*. La base d'histogramme est régulier est

$$g_1(x) = \frac{1}{\sqrt{h}}\mathbb{I}[a \le x < a+h],\ g_2(x) = \frac{1}{\sqrt{h}}\mathbb{I}[a+h \le x < a+2h], \ldots,$$

$$g_m(x) = \frac{1}{\sqrt{h}}\mathbb{I}[a+(m-1)h \le x \le a+mh = b],$$

où $h = (b-a)/m$. On vérifiera qu'il s'agit effectivement de la base orthonormée d'un sous-espace de dimension m de $L_2([a,b])$ et que l'estimateur par projection $\hat{f}_n$ associé à cette base est l'estimateur par histogrammes en (II.5). La base trigonométrique, qui requiert que $m = 2k+1$ soit un entier impair ($k \in \mathbb{N}$), est

$$g_1(x) = \sqrt{\frac{1}{b-a}},\ g_{2j}(x) = \sqrt{\frac{2}{b-a}}\cos\Big(\frac{2\pi j(x-a)}{b-a}\Big),\ g_{2j+1}(x) = \sqrt{\frac{2}{b-a}}\sin\Big(\frac{2\pi j(x-a)}{b-a}\Big),$$

où $j = 1, \ldots, k$. On vérifiera encore qu'il s'agit effectivement de la base orthonormée d'un sous-espace de dimension m de $L_2([a,b])$. L'estimateur par projection qui en résulte n'a pas encore été considéré dans ce cours. On pourrait aussi considérer d'autres bases, comme par exemple des bases d'ondelettes (voir [4])

II.4.2 Risque intégré

Le risque intégré associé à l'estimateur par projection $\hat{f}_n$ est

$$R(\hat{f}_n, f) = \mathrm{E}\bigg[\int_{-\infty}^{\infty}\{\hat{f}_n(x) - f(x)\}^2\,dx\bigg] = \mathrm{E}[\|\hat{f}_n - f\|_2^2].$$

Puisque $\hat{f}_n \in \mathcal{S}_m$ presque sûrement, on a $\langle f_m - f, \hat{f}_n\rangle = 0$, ce qui, puisqu'on a aussi $\langle f_m - f, f_m\rangle = 0$, livre

$$\begin{aligned}\|\hat{f}_n - f\|_2^2 &= \|\{f_m - f\} + \{\hat{f}_n - f_m\}\|_2^2 \\ &= \langle\{f_m - f\} + \{\hat{f}_n - f_m\}, \{f_m - f\} + \{\hat{f}_n - f_m\}\rangle \\ &= \|f_m - f\|_2^2 + \|\hat{f}_n - f_m\|_2^2,\end{aligned}$$

de sorte que le risque intégré s'écrit

$$R(\hat{f}_n, f) = \mathrm{E}[\|\hat{f}_n - f\|_2^2] = \|f_m - f\|_2^2 + \mathrm{E}[\|\hat{f}_n - f_m\|_2^2].$$

Ceci peut être vu comme une nouvelle décomposition du risque intégré en termes de biais et de variance : le terme $\|f_m - f\|_2^2$ représente le coût d'approximation de la fonction f

à estimer par sa projection f_m dans $\mathcal{S}_m$. Pour une collection de sous-espaces emboîtés ($\mathcal{S}_m \subset \mathcal{S}_{m+1}$ pour tout m), ce coût est d'autant moins élevé que m est grand. Au contraire, le terme $\mathrm{E}[\|\hat{f}_n - f_m\|_2^2]$ est un terme de variance, qui représente la difficulté à estimer la projection f_m de f. Plus m est grand, plus le nombre de coefficients c_r à estimer est élevé, ce qui augmente la variabilité de l'estimateur $\hat{f}_n$. Ceci est illustré à la Figure II.12, qui montre l'impact du choix de m sur l'estimation de la densité de Bart Simpson par un estimateur à projection fondé sur la base trigonométrique.

Une minimisation du risque intégré s'obtient donc en réalisant à nouveau un équilibre entre biais et variance, à travers un choix adéquat de la valeur d'un paramètre de lissage (qui est ici le paramètre m). Comme pour l'estimation par noyau, il existe des méthodes qui permettent de choisir en pratique la valeur du paramètre de lissage. Par ailleurs, dans des classes adéquates de densité "de régularité k" (les détails dépendent malheureusement de la base choisie), les estimateurs par projection offrent des taux de convergence optimaux en $n^{-2k/(2k+1)}$, comme c'était le cas pour les estimateurs à noyau.

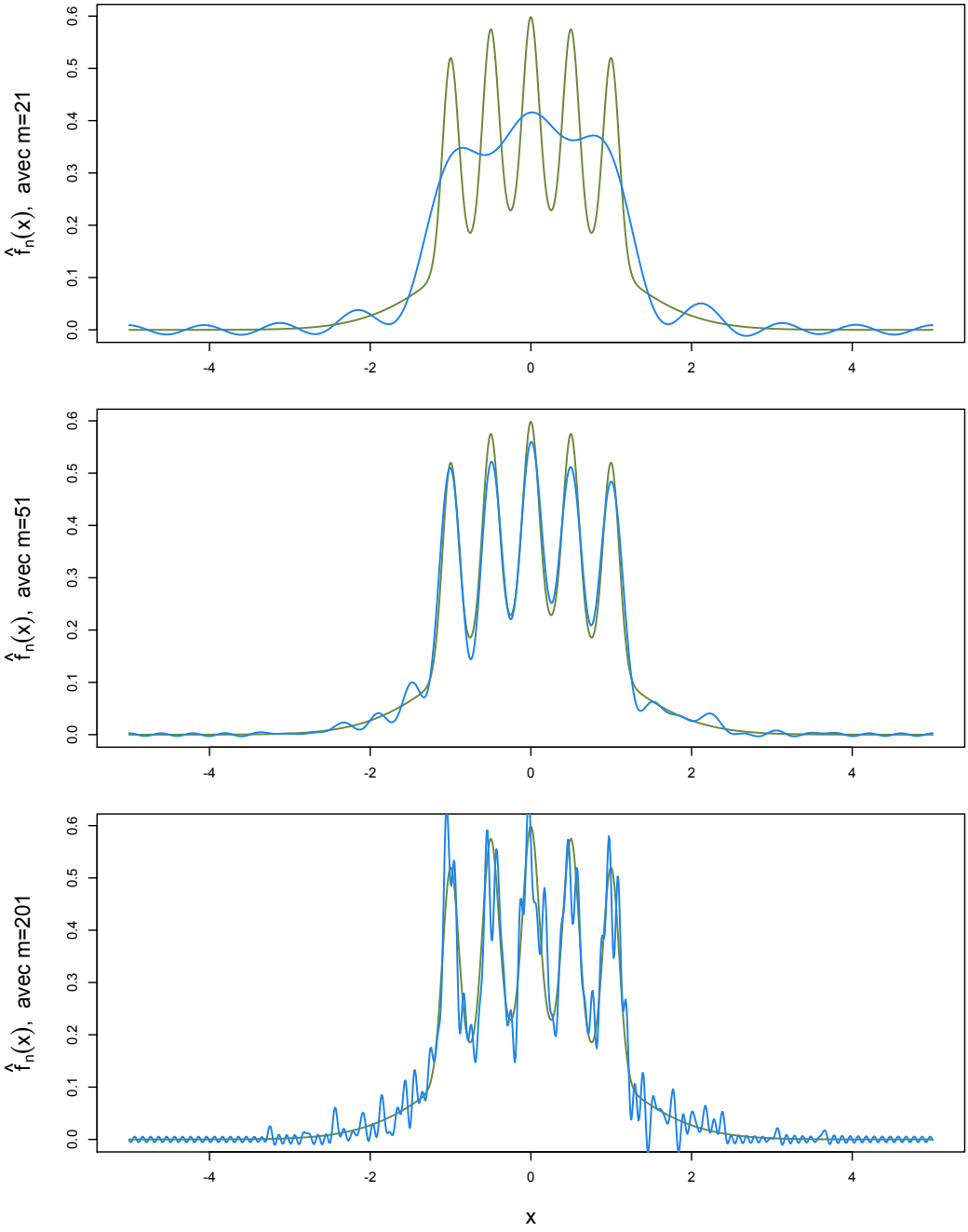

Figure II.12 – *Graphes de l'estimateur par projection* $\hat{f}_n(x)$ *associé à* $a = -5$, $b = 5$ *et la base trigonométrique avec* $m = 21$ *(haut),* $m = 51$ *(milieu), et* $m = 201$ *(bas) calculé sur un échantillon aléatoire de taille* $n = 1000$ *engendré depuis la densité de Bart Simpson (en vert).*

II.5 Codes R

Figure II.1

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
#########
a=-5
b=5
par(mfrow=c(1,1))
par(oma=c(2,2,2,1))
par(mar=c(3,3,1,1))
z=seq(a,b,by=.01)
plot(z,fSimpson(z),typ="l",lwd=2,col="darkolivegreen4",xlab="",ylab="")
mtext("x",side=1,line=3.5,cex=1.2)
mtext("f(x)",side=2,line=3.5,cex=1.2)
```

Figure II.2

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
        }
    }
  return(temp)
}
#########
a=-5
b=5
par(mfrow=c(1,1))
par(oma=c(2,2,2,1))
par(mar=c(3,3,1,1))
z=seq(a,b,by=.01)
plot(z,fSimpson(z),typ="l",lwd=2,col="darkolivegreen4",xlab="",ylab="")
        mtext("x",side=1,line=3.5,cex=1.2)
        mtext("f(x)",side=2,line=3.5,cex=1.2)
```

```
set.seed(1)
n=1000
x=rSimpson(n)
muhat=mean(x)
sigma2hat=mean((x-muhat)^2)
fhat=dnorm(z,mean=muhat,sd=sqrt(sigma2hat))
points(z,fhat,typ="l",lwd=2,col="firebrick2",xlab="",ylab="")
```

Figure II.3

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
        }
    }
  return(temp)
}
#########
a=-5
b=5
n=1000
m=100
set.seed(1)
nbreaks=a+(b-a)*(0:m)/m
x=rSimpson(n)
hist(x,freq=FALSE,breaks=nbreaks,main="",xlab="",ylab="")
z=seq(a,b,by=.01)
points(z,fSimpson(z),typ="l",lwd=1.4,col="darkolivegreen4",xlab="",ylab="")
        mtext("x",side=1,line=3.5,cex=1.2)
        mtext(expression(hat(f)[n](x)),side=2,line=3,cex=1.2)
```

Figure II.4

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
```

```
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
        }
    }
  return(temp)
}
#########
a=-5
b=5
n=1000
mvec=c(20,100,500)
set.seed(1)
x=rSimpson(n)
z=seq(a,b,by=.01)
par(mfrow=c(length(mvec),1))
par(oma=c(2,2,0,1))
par(mar=c(3,4,1,1))
for (i in (1:length(mvec)))
        {
        m=mvec[i]
        nbreaks=a+(b-a)*(0:m)/m
        hist(x,freq=FALSE,breaks=nbreaks,main="",xlab="",ylab="",ylim=c(0,.7))
        points(z,fSimpson(z),typ="l",lwd=1.4,col="darkolivegreen4",xlab="",ylab="")
        if (i==length(mvec)) {mtext("x",side=1,line=3.5,cex=1)}
        mtext(substitute(paste(hat(f)[n](x),",␣␣avec␣m=",mm,sep=""),list(mm=mvec[i])),
                        side=2,line=3.5,cex=1)
        }
```

Figure II.5

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
            }
    }
  return(temp)
```

```
}
#########
a=-5
b=5
nvec=c(300,600,1000,2000)
M=10000
mvec=seq(10,600,by=10)
set.seed(1);
xgrid=seq(a,b,by=0.01)
xgrid=xgrid[1:(length(xgrid)-1)]
f=fSimpson(xgrid)
risk=array(0,dim<-c(length(nvec),length(mvec)))
time=proc.time()[3]
for (r in (1:M))
        {
        for (i in 1:length(nvec))
                {
                n=nvec[i]
                x=rSimpson(n)
                x=pmax(x,a)
                x=pmin(x,b)
                for (j in 1:length(mvec))
                        {
                        m=mvec[j]
                        h=(b-a)/m
                        nbreaks=a+(b-a)*(0:m)/m
                        counts=hist(x,breaks=nbreaks,plot=FALSE)$counts
                        fhat=rep(0,length=length(xgrid))
                        for (s in (1:(length(xgrid))))
                                {
                                fhat[s]=counts[min(which( (xgrid[s]>=nbreaks)
                                                & (xgrid[s]<nbreaks+h) ))]/(n*h)
                                }
                        risk[i,j]=risk[i,j]+sum((f-fhat)^2)/(length(xgrid)-1)
                        }
                }
        print(c(r,M,round((proc.time()[3]-time)/r*( (M-r) )/60,1)))
        }
risk=risk/M

colnvec=c("firebrick1","darkorange","dodgerblue2","darkolivegreen4")
par(mfrow=c(1,1))
par(oma=c(2,1,2,1))
par(mar=c(3,3,1,1))
plot(mvec,risk[1,],typ="l",lwd=2,col=colnvec[1],xlab="",ylab="",ylim=c(0,.012))
        i=1
        xi=mvec[which(risk[i,]==min(risk[i,]))]
        yi=min(risk[i,])
        segments(xi,0,xi,yi,lwd=1,col=colnvec[i],lty=2)
for (i in 2:length(nvec))
        {
        points(mvec,risk[i,],typ="l",lwd=2,col=colnvec[i],xlab="",ylab="")
        xi=mvec[which(risk[i,]==min(risk[i,]))]
        yi=min(risk[i,])
        segments(xi,0,xi,yi,lwd=1,col=colnvec[i],lty=2)
        }
legend(0,.0121,legend=c(paste("n=",nvec[1],sep=""),paste("n=",nvec[2],sep=""),
        paste("n=",nvec[3],sep=""),paste("n=",nvec[4],sep="")),col=colnvec,lwd=2,lty=1,cex=1)
mtext("m",side=1,line=3,cex=1)
```

```
mtext(expression(paste(R[a],"(",hat(f)[n],",f",")",sep="")),side=2,line=2.5,cex=1)
```

Figure II.6

```
rm(list = ls())
#########
KUnif<-function(z){return( (1/2)*(z<=1)*(z>=-1) )}
KTriangulaire<-function(z){return( (1-abs(z))*(z<=1)*(z>=-1) )}
KEpanech<-function(z){return( (3/4)*(1-z^2)*(z<=1)*(z>=-1) )}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
Kcirc<-function(z){return( (pi/4)*cos(pi/2*z)^2*(z<=1)*(z>=-1) )}
#########
a=-2
b=2
colvec=c("firebrick1","darkorange","dodgerblue2","darkolivegreen4","lightsteelblue2")
par(mfrow=c(1,1))
par(oma=c(2,1,2,1))
par(mar=c(3,3,1,1))
z=seq(a,b,by=.01)
plot(z,KUnif(z),typ="l",lwd=2,col=colvec[4],xlab="",ylab="",ylim=c(0,1.0))
        abline(v=-1,col="white",lwd=4)
        segments(-1,0,-1,1/2,col=colvec[4],lwd=2,lty=2)
        abline(v=1,col="white",lwd=4)
        segments(1,0,1,1/2,col=colvec[4],lwd=2,lty=2)
points(z,KTriangulaire(z),typ="l",lwd=2,col=colvec[1],xlab="",ylab="")
points(z,Kcirc(z),typ="l",lwd=2,col=colvec[2],xlab="",ylab="")
points(z,KEpanech(z),typ="l",lwd=2,col=colvec[3],xlab="",ylab="")
points(z,KGauss(z),typ="l",lwd=2,col=colvec[5],xlab="",ylab="")
        mtext("z",side=1,line=3.5,cex=1.2)
        mtext("K(z)",side=2,line=2.8,cex=1.2)
legend(-2,1.0,legend=c("Triangulaire","Circulaire","Epanechnikov","Uniforme","Gaussien"),
        col=colvec,lwd=2,lty=1,cex=1)
```

Figure II.7

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
            }
    }
  return(temp)
```

```
}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
KUnif<-function(z){return( (1/2)*(z<=1)*(z>=-1) )}
#########
a=-5
b=5
n=1000
hGauss=.07
hUnif=.14
set.seed(1)
x=rSimpson(n)
z=seq(a,b,by=.01)
fhatGauss=rep(0,length=length(z))
fhatUnif=rep(0,length=length(z))
for (r in (1:length(z)))
        {
        fhatGauss[r]=(1/(n*hGauss))*sum( KGauss((z[r]-x)/hGauss) )
        fhatUnif[r]=(1/(n*hUnif))*sum( KUnif((z[r]-x)/hUnif) )
        }
par(mfrow=c(1,2))
par(oma=c(2,1,2,1))
par(mar=c(3,4,1,1))
plot(z,fSimpson(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
points(z,fhatGauss,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
        mtext("Noyau␣gaussien",side=3,line=1.4,cex=1)
        mtext("x",side=1,line=3,cex=1.2)
        mtext(expression(hat(f)[n](x)),side=2,line=2.6,cex=1.2)
plot(z,fSimpson(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
points(z,fhatUnif,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
        mtext("Noyau␣uniforme",side=3,line=1.4,cex=1)
        mtext("x",side=1,line=3,cex=1.2)
```

Figure II.8

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
            }
    }
  return(temp)
}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
KUnif<-function(z){return( (1/2)*(z<=1)*(z>=-1) )}
```

```
#########
a=-5
b=5
n=1000
hGaussvec=.07*c(3,1,1/3)
hUnifvec=.14*c(3,1,1/3)
set.seed(1)
x=rSimpson(n)
z=seq(a,b,by=.01)
fhatGauss=rep(0,length=length(z))
fhatUnif=rep(0,length=length(z))
par(mfrow=c(length(hGaussvec),2))
par(oma=c(2,1,2,1))
par(mar=c(4,4,1,1))
for (i in (1:length(hGaussvec)))
        {
        hGauss=hGaussvec[i]
        hUnif=hUnifvec[i]
        for (r in (1:length(z)))
                {
                fhatGauss[r]=(1/(n*hGauss))*sum( KGauss((z[r]-x)/hGauss) )
                fhatUnif[r]=(1/(n*hUnif))*sum( KUnif((z[r]-x)/hUnif) )
                }
        plot(z,fSimpson(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        points(z,fhatGauss,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
                if (i==1) {mtext("Noyau␣gaussien",side=3,line=1.4,cex=1)}
                if (i==length(hGaussvec)) {mtext("x",side=1,line=3,cex=.85)}
                mtext(expression(hat(f)[n](x)),side=2,line=2.7,cex=.85)

        plot(z,fSimpson(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        points(z,fhatUnif,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
                if (i==1) {mtext("Noyau␣uniforme",side=3,line=1.4,cex=1)}
                if (i==length(hGaussvec)) {mtext("x",side=1,line=3,cex=.85)}
        }
```

Figure II.9

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
            }
    }
  return(temp)
```

```
}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
KUnif<-function(z){return( (1/2)*(z<=1)*(z>=-1) )}
#########
a=-5
b=5
nvec=c(300,600,1000,2000)
M=20
hGaussvec=seq(.01,.8,by=.01)
hUnifvec=seq(.01,.8,by=.01)
set.seed(1)
xgrid=seq(a,b,by=0.01)
xgrid=xgrid[1:(length(xgrid)-1)]
f=fSimpson(xgrid)
fhatGauss=rep(0,length=length(f))
fhatUnif=rep(0,length=length(f))
riskGauss=array(0,dim<-c(length(nvec),length(hGaussvec)))
riskUnif=array(0,dim<-c(length(nvec),length(hUnifvec)))
time=proc.time()[3]
for (r in (1:M))
        {
        for (i in 1:length(nvec))
                {
                n=nvec[i]
                x=rSimpson(n)
                x=pmax(x,a)
                x=pmin(x,b)
                for (j in 1:length(hGaussvec))
                        {
                        hGauss=hGaussvec[j]
                        hUnif=hUnifvec[j]
                        for (s in (1:(length(xgrid)-1)))
                                {
                                fhatGauss[s]=(1/(n*hGauss))*sum( KGauss((xgrid[s]-x)/hGauss) )
                                fhatUnif[s]=(1/(n*hUnif))*sum( KUnif((xgrid[s]-x)/hUnif) )
                                }
                        riskGauss[i,j]=riskGauss[i,j]+sum((f-fhatGauss)^2)/(length(xgrid)-1)
                        riskUnif[i,j]=riskUnif[i,j]+sum((f-fhatUnif)^2)/(length(xgrid)-1)
                        }
                }
        print(c(r,M,round((proc.time()[3]-time)/r*( (M-r) )/60,1)))
        }
riskGauss=riskGauss/M
riskUnif=riskUnif/M
colnvec=c("firebrick1","darkorange","dodgerblue2","darkolivegreen4");
par(mfrow=c(1,2))
par(oma=c(2,1,2,1))
par(mar=c(3,3,1,1))
plot(hGaussvec,riskGauss[1,],typ="l",lwd=2,col=colnvec[1],xlab="",ylab="",ylim=c(0,.012))
        i=1
        xi=hGaussvec[which(riskGauss[i,]==min(riskGauss[i,]))]
        yi=min(riskGauss[i,])
        segments(xi,0,xi,yi,lwd=1,col=colnvec[i],lty=2)
for (i in 2:length(nvec))
        {
        points(hGaussvec,riskGauss[i,],typ="l",lwd=2,col=colnvec[i],xlab="",ylab="")
        xi=hGaussvec[which(riskGauss[i,]==min(riskGauss[i,]))]
        yi=min(riskGauss[i,])
        segments(xi,0,xi,yi,lwd=1,col=colnvec[i],lty=2)
```

```
        }
mtext(expression(paste(R[a],"(",hat(f)[n],",f",")",sep="")),side=2,line=2.5,cex=1)
mtext("h",side=1,line=3,cex=1)
legend(0.565,.0121,legend=c(paste("n=",nvec[1],sep=""),paste("n=",nvec[2],sep=""),
        paste("n=",nvec[3],sep=""),paste("n=",nvec[4],sep="")),col=colnvec,lwd=2,lty=1,cex=1)
plot(hUnifvec,riskUnif[1,],typ="l",lwd=2,col=colnvec[1],xlab="",ylab="",ylim=c(0,.012))
        i=1
        xi=hUnifvec[which(riskUnif[i,]==min(riskUnif[i,]))]
        yi=min(riskUnif[i,])
        segments(xi,0,xi,yi,lwd=1,col=colnvec[i],lty=2)
for (i in 2:length(nvec))
        {
        points(hUnifvec,riskUnif[i,],typ="l",lwd=2,col=colnvec[i],xlab="",ylab="")
        xi=hUnifvec[which(riskUnif[i,]==min(riskUnif[i,]))]
        yi=min(riskUnif[i,])
        segments(xi,0,xi,yi,lwd=1,col=colnvec[i],lty=2)
        }
mtext("h",side=1,line=3,cex=1)
```

Figure II.10

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
            }
    }
  return(temp)
}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
KGaussConv<-function(z){return( (1/(2*sqrt(pi)))*exp(-z^2/4) )}
KUnif<-function(z){return( (1/2)*(z<=1)*(z>=-1) )}
KUnifConv<-function(z){return( ((2+z)/4)*(z>-2)*(z<0) + ((2-z)/4)*(z>=0)*(z<2) )}
#########
a=-5
b=5
nvec=c(1000,400)
hgrid=seq(.01,.8,by=.01)
set.seed(1)
colnvec=c("darkorange","dodgerblue2")
ISEGauss=array(0,dim<-c(length(nvec),length(hgrid)))
ISEUnif=array(0,dim<-c(length(nvec),length(hgrid)))
for (r in (1:length(nvec))){
        n=nvec[r];
```

```
        x=rSimpson(n);
        MGaussConv=array(0,dim<-c(n,n))
        MGauss=array(0,dim<-c(n,n))
        MUnifConv=array(0,dim<-c(n,n))
        MUnif=array(0,dim<-c(n,n))
        for (s in (1:length(hgrid)))
                {
                print(c(r,length(nvec),s,length(hgrid)))
                h=hgrid[s];

                for (i in (1:n)) {
                        for (j in (1:n)) {
                                MGaussConv[i,j]=KGaussConv((x[i]-x[j])/h)
                                MGauss[i,j]=KGauss((x[i]-x[j])/h)
                                MUnifConv[i,j]=KUnifConv((x[i]-x[j])/h)
                                MUnif[i,j]=KUnif((x[i]-x[j])/h)
                                }
                        }
                ISEGauss[r,s]=sum( MGaussConv )/(n^2*h)-2/(n*(n-1)*h)*sum( MGauss )
                        + 2/((n-1)*h)*KGauss(0)
                ISEUnif[r,s]=sum( MUnifConv )/(n^2*h)-2/(n*(n-1)*h)*sum( MUnif )
                        + 2/((n-1)*h)*KUnif(0)
                }
         }
```

Figure II.11

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
            }
    }
  return(temp)
}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
KGausspp<-function(z){return( (1/sqrt(2*pi))*(z^2-1)*exp(-z^2/2) )}
#########
a=-5
b=5
n=1000
alpha=.05
hGauss=.07
cGauss=hGauss*n^(1/5)
```

```
set.seed(1)
x=rSimpson(n)
z=seq(a,b,by=.01)
dKGauss=1
cKGauss=1/sqrt(2*pi)
fhatGauss=rep(0,length=length(z))
fhatGausspp=rep(0,length=length(z))
fhatGaussalphaplus=rep(0,length=length(z))
fhatGaussalphamoins=rep(0,length=length(z))
zalphademi=qnorm(1-alpha/2)
for (r in (1:length(z)))
        {
        fhatGauss[r]=(1/(n*hGauss))*sum( KGauss((z[r]-x)/hGauss) )
        fhatGausspp[r]=(1/(n*hGauss^3))*sum( KGausspp((z[r]-x)/hGauss) )
        fhatGaussalphaplus[r]=fhatGauss[r]+n^{-2/5}*( (cGauss^2/2)*dKGauss*fhatGausspp[r]
                + zalphademi*sqrt(fhatGauss[r]*cKGauss/cGauss) )
        fhatGaussalphamoins[r]=fhatGauss[r]-n^{-2/5}*( (cGauss^2/2)*dKGauss*fhatGausspp[r]
                + zalphademi*sqrt(fhatGauss[r]*cKGauss/cGauss) )
        }
par(mfrow=c(1,1))
par(oma=c(2,1,2,1))
par(mar=c(3,2,1,1))
plot(z,fSimpson(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
points(z,fhatGauss,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
points(z,fhatGaussalphaplus,typ="l",lty=2,lwd=1.4,main="",col="darkorange",xlab="",ylab="")
points(z,fhatGaussalphamoins,typ="l",lty=2,lwd=1.4,main="",col="darkorange",xlab="",ylab="")
        mtext("x",side=1,line=3,cex=1)
```

Figure II.12

```
rm(list = ls())
#########
fSimpson<-function(z){
  temp=(1/2)*dnorm(z,mean=0,sd=1)
  for (j in (0:4)) {temp=temp+(1/10)*dnorm(z,mean=(j/2)-1,sd=1/10)}
  return(temp)
}
rSimpson<-function(n){
  u=runif(n)
  temp=rep(0,n)
  for (i in (1:n))
    {
    if (u[i]>1/2) {temp[i]=rnorm(1,mean=0,sd=1)}
    for (j in (0:4))
        {
        if ((u[i]>(j/10)) & (u[i]<=(j+1)/10))
                {temp[i]=rnorm(1,mean=(j/2)-1,sd=1/10)}
            }
    }
  return(temp)
}
grtrig<-function(z,r,a,b){
  if (r==1) {return(1/sqrt(b-a))}
  if ((r>1)&((r%%2)==0)) {return( sqrt(2/(b-a)) * cos(2*pi*(r/2)*(z-a)/(b-a)) )}
  if ((r>1)&((r%%2)==1)) {return( sqrt(2/(b-a)) * sin(2*pi*((r-1)/2)*(z-a)/(b-a)) )}
}
#########
```

```
a=-5
b=5
n=1000
mvec=c(21,51,201) # m doit etre impair
set.seed(1)
x=rSimpson(n)
z=seq(a,b,by=.01)
par(mfrow=c(length(mvec),1))
par(oma=c(2,2,0,1))
par(mar=c(3,4,1,1))
for (i in (1:length(mvec)))
        {
        m=mvec[i]
        fhatproj=rep(0,length=length(z))
        for (s in (1:length(z)))
                {
                for (r in (1:m)) {fhatproj[s]=fhatproj[s]+mean(grtrig(x,r,a,b))*grtrig(z[s],r,a,b)}
                }
        plot(z,fSimpson(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        points(z,fhatproj,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
        if (i==length(mvec)) {mtext("x",side=1,line=3.5,cex=1)}
        mtext(substitute(paste(hat(f)[n](x),",␣␣avec␣m=",mm,sep=""),list(mm=mvec[i]))
                ,side=2,line=3.5,cex=1)
        }
```

II.6 TD

1. Prouver le Théorème II.2 (pour établir l'expression de h_n^{opt}, on négligera les termes de reste en (II.6)).

2. Vérifier que la taille de fenêtre qui, si on néglige les termes de reste, minimise le risque ponctuel en (II.12) est donnée par (II.15) et que le risque ponctuel minimal qui en résulte satisfait (II.16).

3. Soit $\hat{f}_n(x)$ l'estimateur à noyau en (II.11) avec $h = h_n$. Montrer que

$$\int_{-\infty}^{\infty} x\hat{f}_n(x)\,dx = \bar{X}_n,$$

où $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est la moyenne empirique des observations. Ce résultat est-il surprenant ?

4. Soit f une fonction de densité telle $f(x) > 0$ pour tout $x \in (a, b)$ et $f(x) = 0$ pour tout $x \notin [a, b]$. Supposons que $f(a) > 0$.

   (a) Dans cette sous-question, on supposera en outre que f est continue à droite en a. Par un calcul direct, montrer que si $\hat{f}_n$ est l'estimateur par histogramme en (II.5) fondé sur une taille de fenêtre $h = h_n$ qui tend vers zéro quand n diverge vers l'infini, alors on a $\mathrm{E}[\hat{f}_n(a)] \to f(a)$ quand n diverge vers l'infini, de sorte que $\hat{f}_n(a)$ est un estimateur asymptotiquement sans biais de $f(a)$.

(b) Dans cette sous-question, on supposera plutôt que f est dérivable sur $[a, b]$ et que $\|f'\|_\infty < \infty$. Montrer que, si $\hat{f}_n(x)$ est l'estimateur à noyau en (II.11) fondé sur un noyau K symétrique par rapport à zéro (au sens où $K(-z) = K(z)$ pour tout z) vérifiant $e_K := \int_{-\infty}^{\infty} |z||K(z)|\,dz < \infty$ et sur une taille de fenêtre $h = h_n$ qui tend vers zéro quand n diverge vers l'infini, alors $\mathrm{E}[\hat{f}_n(a)] \to f(a)/2$, quand n diverge vers l'infini [1], de sorte que $\hat{f}_n(a)$ n'est pas un estimateur asymptotiquement sans biais de $f(a)$. On dit que l'estimateur à noyau *souffre d'un effet de bord*.

5. Cet exercice est relatif au jeu de données *quakes* de R, qui reprend certaines variables associées à des tremblements de terre (tapez `quakes` pour consulter le jeu de données et `?quakes` pour obtenir des informations) et se focalise plus particulièrement sur la variable *Depth* (accessible via la commande `quakes[,3]`).

(a) Faire le graphe de l'estimateur par histogramme associé à $a = \min(X_1, \ldots, X_n)$, $b = \max(X_1, \ldots, X_n)$ et $m = 30$ (ici, $X_1, \ldots, X_n$ représentent les $n = 1000$ valeurs de la variable *Depth* dans le jeu de données).

(b) Superposer à ce graphe l'estimateur à noyau associé au noyau gaussien et à la taille de la fenêtre $h_n = 20$ (observer l'effet de bord décrit à l'exercice précédent !)

(c) Faire de même avec l'estimateur à noyau fondé sur le noyau gaussien et (1) la taille de la fenêtre "rule-of-thumb"en (II.23), puis (2) la taille de la fenêtre de validation croisée en (II.25) [2].

6. Pour mieux apprécier l'effet de bord étudié à l'exercice 4, refaire l'exercice précédent en utilisant un échantillon aléatoire de taille $n = 1000$ engendré depuis la loi uniforme sur $[0, 1]$ (en (b), prendre $h_n = .05$).

7. Soit f de classe C_k et soit $\hat{f}_n(x)$ l'estimateur à noyau fondé sur la taille de fenêtre h_n et sur un noyau K d'ordre k (c'est-à-dire vérifiant $\int_{-\infty}^{\infty} z^r K(z)\,dz = 0$ pour $r = 1, 2, \ldots, k-1$ et $\int_{-\infty}^{\infty} |z|^k |K(z)|\,dz < \infty$). Dans les points ci-dessous, on procédera comme dans la preuve du Théorème II.3 (c'est-à-dire en faisant des développements de Taylor), mais sans contrôler proprement les termes de reste.

(a) Montrer que

$$b_x(\hat{f}_n, f) = \frac{h_n^k}{k!} f^{(k)}(x) d_{k,K} + o(h_n^k) \quad \text{et} \quad v_x(\hat{f}_n) = \frac{1}{nh_n} f(x) c_K + o\Big(\frac{1}{nh_n}\Big)$$

quand n diverge vers l'infini, où $f^{(k)}(x)$ est la $k$ème dérivée de f en x et où $d_{k,K}$ est une constante (à déterminer) ne dépendant que de K et de k.

1. Indice : s'inspirer du calcul de $\mathrm{E}[\hat{f}_n(x)]$ dans la preuve du Théorème II.3.

2. Vous serez peut-être déçu du résultat de l'estimateur par validation croisée, mais comme on peut le lire dans [6], *"Do not assume that, if the estimator $\hat{f}_n$ is wiggly, then cross-validation has let you down. The eye is not a good judge of risk" !*

(b) En déduire que

$$R_x(\hat{f}_n, f) = \frac{h_n^{2k}}{(k!)^2}(f^{(k)}(x))^2 d_{k,K}^2 + \frac{1}{nh_n} f(x) c_K + o(h_n^{2k}) + o\Big(\frac{1}{nh_n}\Big),$$

quand n diverge vers l'infini, et prouver que le risque ponctuel optimal qui en résulte vérifie (pour une certaine constante C ne dépendant pas de n)

$$R_x(\hat{f}_n, f) = Cn^{-2k/(2k+1)} + o\big(n^{-2k/(2k+1)}\big)$$

quand n diverge vers l'infini [3].

8. Soit f la densité gaussienne de moyenne μ et de variance σ^2.

(a) Vérifier que $\|f''\|_2$ ne dépend pas de μ.

(b) Prouver que $\|f''\|_2^2 = 3/(8\sqrt{\pi}\sigma^5)$.

9. Soit $X_1, \ldots, X_n$ un échantillon aléatoire associé à la densité f. Puisque la fonction de répartition correspondante vérifie $F(x) = \int_{-\infty}^{x} f(z)\, dz$, un estimateur naturel de $F(x)$ est

$$\hat{F}_n(x) = \int_{-\infty}^{x} \hat{f}_n(z)\, dz,$$

où $\hat{f}_n$ est un estimateur de f. Supposons qu'on travaille avec un estimateur à noyau associé à un noyau K qui est une fonction de densité.

(a) Identifier un avantage à utiliser cet estimateur de fonction de répartition plutôt que la fonction de répartition empirique.

(b) Montrer que

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \tilde{K}\Big(\frac{x - X_i}{h_n}\Big),$$

où $\tilde{K}$ est la fonction de répartition associée à la densité K.

(c) En utilisant un noyau gaussien et une taille de fenêtre que vous choisirez visuellement, superposer le graphe de ce nouvel estimateur à la Figure I.1.

10. Vérifier que la base d'histogrammes réguliers et la base trigonométrique décrites dans la Section II.4.1 sont bien constituées de fonctions orthonormées de $L_2([a, b])$ (par simplicité, on fera l'exercice seulement pour $a = 0$ et $b = 1$).

11. Cet exercice est à nouveau relatif au jeu de données *quakes* de R (voir l'exercice 5).

(a) Comme à l'exercice 5(a), faire le graphe de l'estimateur par histogramme associé à $a = \min(X_1, \ldots, X_n)$, $b = \max(X_1, \ldots, X_n)$ et $m = 30$ (ici, $X_1, \ldots, X_n$ représentent les $n = 1000$ valeurs de la variable *Depth* dans le jeu de données).

3. Le résultat est en ligne avec le Théorème II.6, qui concerne plutôt le risque intégré.

(b) Pour les mêmes a et b que ci-dessus, superposer les graphes des estimateurs par projection associés à la base trigonométrique et à trois valeurs de m que vous jugerez adaptées pour décrire la dépendance en m du biais et de la variance de ces estimateurs.

Chapitre III

Régression non paramétrique

Dans ce chapitre, nous considérons la problématique usuelle de la *régression*, où on veut étudier le lien existant entre une "variable à expliquer" Y et une "variable explicative" X. L'approche la plus classique consiste à postuler que le lien entre Y et X est linéaire, au sens où il existe des nombres réels β_0 et β_1 tels que

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où la variable d'erreur ε vérifie $\mathrm{E}[\varepsilon|X = x] = 0$ pour toute valeur x de X. Plus généralement, on peut adopter un modèle polynomial de la forme

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon,$$

où on fait la même hypothèse sur le terme d'erreur. Cette approche est *paramétrique* puisque la fonction de régression $m(x) = \mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p$ est connue dès qu'un nombre fini de paramètres—$\beta_0, \beta_1, \ldots, \beta_p$—le sont. Dans une optique de prédiction de Y sur la base de X, l'estimateur usuel de $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, à savoir l'estimateur des moindres carrés, permettra de construire le prédicteur

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \ldots + \hat{\beta}_p X^p,$$

lequel sera très performant si le modèle adopté coïncide avec le vrai modèle (ou en est proche), mais retournera des résultats qui peuvent être catastrophiques si ce n'est pas le cas. Ceci suggère d'adopter plutôt une approche de type *non paramétrique*, où on ne fait pas de restriction structurelle sur la fonction de régression $m(x)$. Le modèle associé est donc

$$Y = m(X) + \varepsilon,$$

où la variable d'erreur ε vérifie $\mathrm{E}[\varepsilon|X = x] = 0$ pour toute valeur x de X, de sorte que la fonction de régression $m(x)$ s'interprète encore de façon directe en termes d'espérance conditionnelle : $m(x) = \mathrm{E}[Y|X = x]$. Dans ce chapitre, nous décrivons des méthodes qui permettent d'estimer la fonction de régression m de façon non paramétrique sur la base de copies indépendantes $(X_1, Y_1), \ldots, (X_n, Y_n)$ du vecteur aléatoire (X, Y). Un exemple de telle fonction de régression m et d'échantillon aléatoire associé est donné à la Figure III.1.

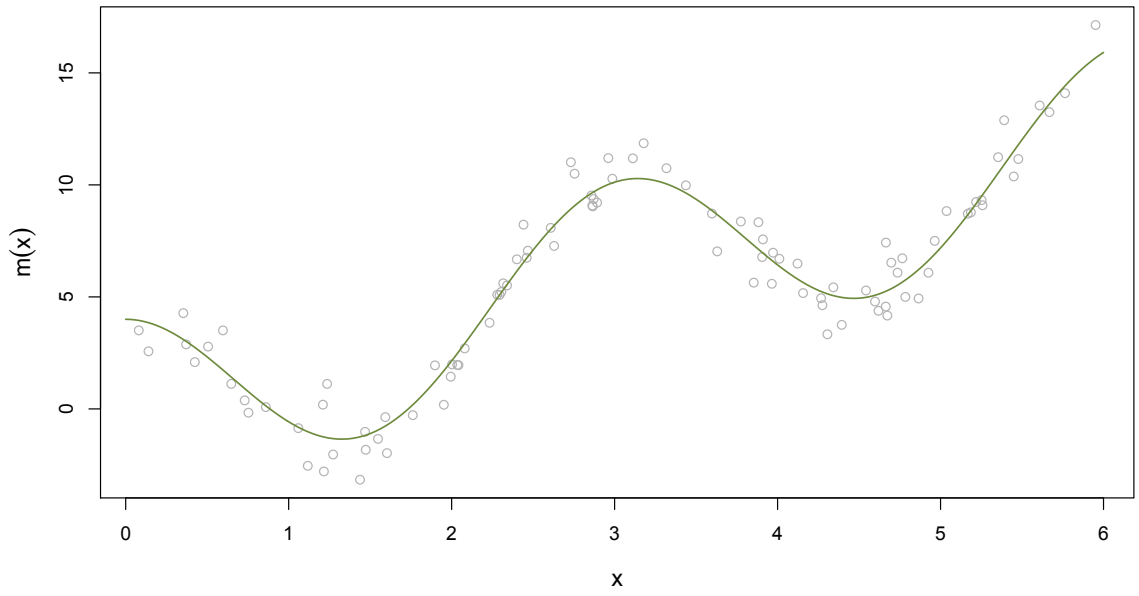

Figure III.1 – *Graphe de la fonction de régression* $m(x) = 2x - \sin(2x) + 4\cos(2x)$ *et un échantillon aléatoire* (X_i, Y_i), $i = 1, \ldots, n = 100$, *engendré depuis le modèle* $Y = m(X) + \varepsilon$, *avec* $\varepsilon \sim \mathcal{N}(0, 1)$ *indépendant de* $X \sim \mathrm{Unif}(0, 6)$.

III.1 Estimation par noyau

En régression non paramétrique, la méthode d'estimation par noyau est basée sur le fait que la fonction de régression m peut s'écrire

$$\begin{aligned} m(x) &= \mathrm{E}[Y|X=x] \\ &= \int_{-\infty}^{\infty} y\, f^{Y|X=x}(y)\, dy \\ &= \int_{-\infty}^{\infty} y\, \frac{f^{(X,Y)}(x,y)}{f^X(x)}\, dy \\ &= \frac{1}{f^X(x)} \int_{-\infty}^{\infty} y\, f^{(X,Y)}(x,y)\, dy, \end{aligned} \tag{III.1}$$

où $f^{Y|X=x}$ désigne la densité conditionnelle de Y sachant que $X = x$. Ceci montre qu'un estimateur convergent de $m(x)$ peut être obtenu en remplaçant, dans (III.1), les densités marginale f^X et jointe $f^{(X,Y)}$ par des estimateurs convergents. On a vu au chapitre précédent comment estimer f^X, et il nous faut donc considérer maintenant le problème de l'estimation d'une densité jointe.

III.1.1 Estimation d'une densité jointe

Soit un vecteur aléatoire bivarié (X, Y) admettant la densité $f = f^{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}$ (l'extension au cas d-varié ne pose que des difficultés de notation et est donc laissée comme exercice). Par analogie avec le cas univarié, on appellera *estimateur à noyau* un estimateur de la forme

$$\hat{f}_n(x,y) = \frac{1}{n h_{nx} h_{ny}} \sum_{i=1}^{n} K_2\Big(\frac{x - X_i}{h_{nx}}, \frac{y - Y_i}{h_{ny}}\Big),$$

où $h_{nx}, h_{ny} > 0$ sont des *tailles de fenêtre* et où K_2 est un noyau bivarié, c'est-à-dire une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ telle que (i) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K_2(u,v)\, dudv = 1$, (ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uK_2(u,v)\, dudv = 0$, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} vK_2(u,v)\, dudv = 0$, et (iii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K_2^2(u,v)\, dudv < \infty$ (typiquement, on utilisera pour K_2 une fonction de densité sur $\mathbb{R}^2$ de carré intégrable et d'espérance nulle).

Utiliser des tailles de fenêtre h_{nx} et h_{ny} différentes peut être indiqué en pratique quand les marginales X et Y ont des échelles hétérogènes. Par simplicité, nous nous restreindrons au cas $h_{nx} = h_{ny} (=: h_n)$ dans la suite. De même, nous ne considérerons que des *noyaux produits*, c'est-à-dire des noyaux de la forme $K_2(u,v) = K(u)K(v)$, où K est un noyau univarié (plus généralement, on pourrait en fait aussi considérer des noyaux produits de la forme $K_2(u,v) = K(u)\tilde{K}(v)$, où K et $\tilde{K}$ sont des noyaux univariés qui

peuvent être différents). L'estimateur à noyau ci-dessus prend donc la forme

$$\hat{f}_n(x,y) = \frac{1}{nh_n^2}\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h_n}\Big)K\Big(\frac{y-Y_i}{h_n}\Big).$$

La Figure III.2 illustre cet estimateur à noyau pour l'estimation de la loi normale standard en utilisant pour K le noyau gaussien. Un exemple plus sophistiqué, qui permettra d'appréhender l'impact du choix de h_n sur cet estimateur, sera considéré dans les exercices.

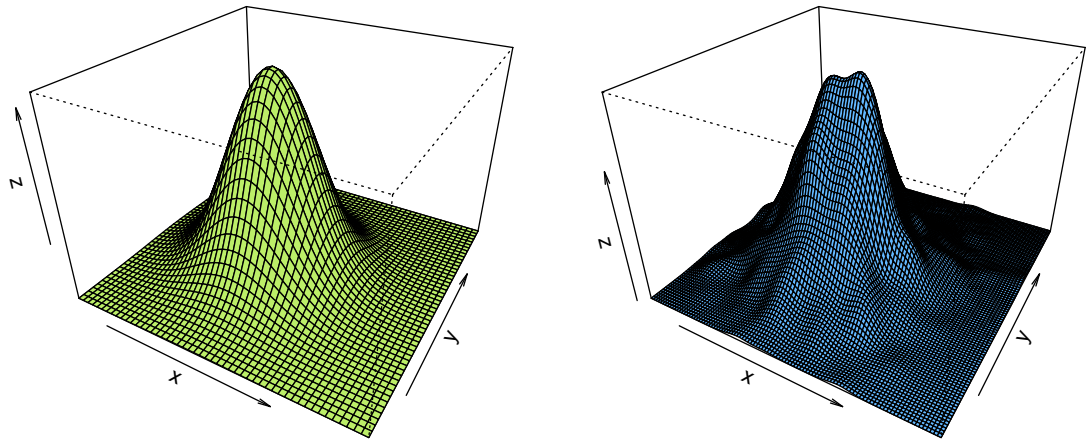

Figure III.2 – *(Gauche :) graphe de la densité de la loi normale standard bivariée. (Droite :) graphe de l'estimateur à noyau fondé sur le noyau gaussien et la taille de fenêtre $h_n = 1$, calculé sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la loi normale standard bivariée.*

Le biais $b_{x,y}(\hat{f}_n, f) := \mathrm{E}[\hat{f}_n(x,y)] - f(x,y)$, la variance $v_{x,y}(\hat{f}_n) := \mathrm{Var}[\hat{f}_n(x,y)]$ et le risque ponctuel $R_{x,y}(\hat{f}_n, f) = \mathrm{E}[\{\hat{f}_n(x,y) - f(x,y)\}^2]$ de cet estimateur de densité sont décrits dans le résultat suivant.

Théorème III.1 *Supposons que f soit de classe C^3 sur $\mathbb{R}^2$ et que toutes les dérivées partielles d'ordre 1 et d'ordre 3 de f soient bornées sur $\mathbb{R}^2$. Soit K un noyau univarié tel que $\int_{-\infty}^{\infty} |z|^3|K(z)|\,dz < \infty$ et $\int_{-\infty}^{\infty} |z|K^2(z)\,dz < \infty$. Soit $h = h_n$ une suite qui est $o(1)$ et telle que nh_n^2 diverge vers l'infini. Soit $\hat{f}_n$ l'estimateur à noyau de f associé à K et à h_n. Posons $d_K := \int_{-\infty}^{\infty} z^2K(z)\,dz$ et $c_K := \|K\|_2^2$. Alors, pour tout $(x,y) \in \mathbb{R}^2$, on a*

$$b_{x,y}(\hat{f}_n, f) = \frac{h_n^2}{2} d_K \triangle f(x,y) + o(h_n^2) \quad \textit{et} \quad v_{x,y}(\hat{f}_n) = \frac{1}{nh_n^2} c_K f(x,y) + o\Big(\frac{1}{nh_n^2}\Big),$$

quand n diverge vers l'infini (ici, $\triangle f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$ désigne le laplacien de f), de sorte que

$$R_{x,y}(\hat{f}_n, f) = \frac{h_n^4}{4} d_K^2 (\triangle f(x,y))^2 + \frac{1}{nh_n^2} c_K f(x,y) + o(h_n^4) + o\Big(\frac{1}{nh_n^2}\Big), \tag{III.2}$$

quand n diverge vers l'infini.

Preuve du Théorème II.3. Puisque les observations (X_i, Y_i) sont indépendantes et admettent la densité f, on obtient

$$\begin{aligned}
\mathrm{E}[\hat{f}_n(x,y)] &= \mathrm{E}\Big[\frac{1}{nh_n^2}\sum_{i=1}^n K\Big(\frac{x-X_i}{h_n}\Big) K\Big(\frac{y-Y_i}{h_n}\Big)\Big] = \frac{1}{h_n^2}\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big) K\Big(\frac{y-Y_1}{h_n}\Big)\Big] \\
&= \frac{1}{h_n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K\Big(\frac{x-u}{h_n}\Big) K\Big(\frac{y-v}{h_n}\Big) f(u,v)\, du dv \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(s)K(t) f(x-h_n s, y-h_n t)\, ds dt,
\end{aligned}$$

où on a effectué le changement de variables $(u,v) = (x - h_n s, y - h_n t)$. Par conséquent, le biais de $\hat{f}_n(x,y)$ est

$$b_{x,y}(\hat{f}_n, f) = \mathrm{E}[\hat{f}_n(x,y)] - f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(s)K(t)\{f(x-h_n s, y-h_n t) - f(x,y)\}\, ds dt,$$

où on a utilisé le fait que $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(s)K(t)\, ds dt = (\int_{-\infty}^{\infty} K(s)\, ds)(\int_{-\infty}^{\infty} K(t)\, dt) = 1$. Un développement de Taylor multivarié (Théorème A.3) fournit alors

$$\begin{aligned}
f(x-h_n s, y-h_n t) &= f(x,y) - h_n s \frac{\partial f}{\partial x}(x,y) - h_n t \frac{\partial f}{\partial y}(x,y) \\
&\quad + \frac{1}{2} h_n^2 s^2 \frac{\partial^2 f}{\partial x^2}(x,y) + \frac{1}{2} h_n^2 t^2 \frac{\partial^2 f}{\partial y^2}(x,y) + h_n^2 st \frac{\partial^2 f}{\partial x \partial y}(x,y) + R_n,
\end{aligned}$$

où, pour un certain $\theta_{n,x,y}$ entre (x,y) et $(x - h_n s, y - h_n t)$,

$$R_n = h_n^3 \sum_{\alpha_1+\alpha_2=3} \frac{1}{(\alpha_1!)(\alpha_2!)} \frac{\partial^3 f}{\partial x^{\alpha_1} \partial y^{\alpha_2}}(\theta_{n,x,y})(-s)^{\alpha_1}(-t)^{\alpha_2}.$$

En utilisant les identités $\int_{-\infty}^{\infty} K(s)\,ds = 1$ et $\int_{-\infty}^{\infty} sK(s)\,ds = 0$, ceci donne

$$\begin{aligned} b_{x,y}(\hat{f}_n, f) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(s)K(t)\Big\{\frac{1}{2}h_n^2 s^2 \frac{\partial^2 f}{\partial x^2}(x,y) + \frac{1}{2}h_n^2 t^2 \frac{\partial^2 f}{\partial y^2}(x,y) + R_n\Big\}\,dsdt \\ &= \frac{h_n^2}{2} d_K \Big(\frac{\partial^2 f}{\partial x^2}(x,y) + \frac{\partial^2 f}{\partial y^2}(x,y)\Big) + O(h_n^3) \\ &= \frac{h_n^2}{2} d_K \triangle f(x,y) + o(h_n^2) \end{aligned} \tag{III.3}$$

(on vérifiera aisément que, sous les hypothèses considérées, le terme de reste est effectivement $O(h_n^3)$).

Pour la variance de $\hat{f}_n(x,y)$, on a

$$\begin{aligned} v_{x,y}(\hat{f}_n) = \mathrm{Var}[\hat{f}_n(x,y)] &= \frac{1}{n^2 h_n^4}\mathrm{Var}\Big[\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h_n}\Big)K\Big(\frac{y-Y_i}{h_n}\Big)\Big] \\ &= \frac{1}{nh_n^4}\mathrm{Var}\Big[K\Big(\frac{x-X_1}{h_n}\Big)K\Big(\frac{y-Y_1}{h_n}\Big)\Big] \\ &= \frac{1}{nh_n^4}\Big\{\mathrm{E}\Big[\Big(K\Big(\frac{x-X_1}{h_n}\Big)K\Big(\frac{y-Y_1}{h_n}\Big)\Big)^2\Big] - \Big(\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)K\Big(\frac{y-Y_1}{h_n}\Big)\Big]\Big)^2\Big\} \\ &= \frac{1}{nh_n^4}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)K^2\Big(\frac{y-Y_1}{h_n}\Big)\Big] - \frac{1}{n}\Big(\frac{1}{h_n^2}\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)K\Big(\frac{y-Y_1}{h_n}\Big)\Big]\Big)^2 \\ &= \frac{1}{nh_n^4}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)K^2\Big(\frac{y-Y_1}{h_n}\Big)\Big] + O\Big(\frac{1}{n}\Big), \end{aligned}$$

où a utilisé le fait que

$$\begin{aligned} \frac{1}{n}\Big(\frac{1}{h_n^2}\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)K\Big(\frac{y-Y_1}{h_n}\Big)\Big]\Big)^2 &= \frac{1}{n}(\mathrm{E}[\hat{f}_n(x,y)])^2 \\ &= \frac{1}{n}(f(x,y)+o(1))^2 = \frac{1}{n}O(1) = O\Big(\frac{1}{n}\Big). \end{aligned}$$

En procédant comme pour le biais, on obtient

$$\begin{aligned}
\frac{1}{nh_n^4}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)K^2\Big(\frac{y-Y_1}{h_n}\Big)\Big] &= \frac{1}{nh_n^4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K^2\Big(\frac{x-u}{h_n}\Big)K^2\Big(\frac{y-v}{h_n}\Big)f(u,v)\,dudv \\
&= \frac{1}{nh_n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K^2(s)K^2(t)f(x-h_ns,y-h_nt)\,dsdt \\
&= \frac{1}{nh_n^2}c_Kf(x,y) + \frac{1}{nh_n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K^2(s)K^2(t)\{f(x-h_ns,y-h_nt)-f(x,y)\}\,dsdt \\
&= \frac{1}{nh_n^2}c_Kf(x,y) + o\Big(\frac{1}{nh_n^2}\Big).
\end{aligned}$$

On peut donc conclure que

$$v_{x,y}(\hat{f}_n) = \frac{1}{nh_n^2}c_Kf(x,y) + o\Big(\frac{1}{nh_n^2}\Big) = \frac{1}{nh_n^2}c_Kf(x,y) + o\Big(\frac{1}{nh_n^2}\Big),$$

ce qui, en utilisant (III.3), fournit finalement

$$R_{x,y}(\hat{f}_n,f) = b_{x,y}^2(\hat{f}_n,f) + v_{x,y}(\hat{f}_n) = \frac{h_n^4}{4}d_K^2(\triangle f(x,y))^2 + o(h_n^4) + \frac{1}{nh_n^2}c_Kf(x,y) + o\Big(\frac{1}{nh_n^2}\Big).$$

Le résultat est démontré. □

Lorsqu'on compare ce résultat au résultat univarié correspondant (le Théorème II.3), on voit que le biais est du même ordre en univarié et en bivarié (le biais est d'ordre $O(h_n^2)$ dans les deux cas), mais que la variance est plus grande dans le cas bivarié que dans le cas univarié : comme h_n tend vers zéro, l'ordre de grandeur bivarié $O(\frac{1}{nh_n^2})$ est en effet supérieur à l'ordre de grandeur $O(\frac{1}{nh_n})$ obtenu en univarié. Ceci implique qu'il est plus difficile d'estimer une densité sur $\mathbb{R}^2$ qu'une densité sur $\mathbb{R}$. Pour quantifier cet effet, on peut comparer le risque minimal qui peut être atteint en bivarié à celui qu'on avait pu réaliser en univarié. Un calcul direct montre que le risque ponctuel en (III.2) est minimisé lorsque

$$h_n = \left(\frac{2f(x,y)c_K}{(\triangle f(x,y))^2d_K^2}\right)^{1/6} n^{-1/6}, \tag{III.4}$$

ce qui mène à un risque ponctuel minimal de la forme

$$R_{x,y}(\hat{f}_n,f) = C_{K,f,x}n^{-2/3} + o(n^{-2/3}), \tag{III.5}$$

pour une certaine constante $C_{K,f,x}$ ne dépendant pas de n. Par conséquent, le taux de convergence optimal bivarié est moins bon que le taux de convergence optimal $n^{-4/5}$ obtenu en univarié. Ceci est un phénomène général : le taux de convergence optimal en dimension d, qui est en fait $n^{-4/(d+4)}$, est d'autant moins satisfaisant que d est grand. Ceci

est connu sous le nom de *fléau de la dimensionalité* (en anglais, *curse of dimensionality*) : plus la dimension est grande, plus il faut d'observations pour estimer une densité avec une précision fixée. Ceci est si sévère (le nombre d'observations requises augmente de façon exponentielle) qu'il est déjà extrêmement périlleux d'estimer une densité sur $\mathbb{R}^4$ ou $\mathbb{R}^5$.

III.1.2 Estimateur de Nadaraya–Watson

Comme on l'a vu à la page 59, un estimateur de la fonction de régression m peut être obtenu en remplaçant les densités f^X et $f^{(X,Y)}$ par des estimateurs adéquats. En utilisant pour ce faire les estimateurs à noyau introduits aux Sections II.3.1 et III.1.1, ceci conduit à l'estimateur

$$\begin{aligned}
\hat{m}_n(x) &= \frac{1}{\hat{f}_n^X(x)} \int_{-\infty}^{\infty} y\, \hat{f}_n^{(X,Y)}(x,y)\, dy \\
&= \frac{\int_{-\infty}^{\infty} y \left\{ \frac{1}{nh_n^2} \sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big) K\big(\frac{y-Y_i}{h_n}\big) \right\} dy}{\frac{1}{nh_n} \sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big)} \\
&= \frac{\frac{1}{nh_n^2} \sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big) \big\{ \int_{-\infty}^{\infty} y\, K\big(\frac{y-Y_i}{h_n}\big) \big\}\, dy}{\frac{1}{nh_n} \sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big)} \\
&= \frac{\frac{1}{nh_n} \sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big) Y_i}{\frac{1}{nh_n} \sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big)},
\end{aligned}$$

puisqu'en posant $z = (y - Y_i)/h_n$, on obtient

$$\begin{aligned}
\frac{1}{h_n} \int_{-\infty}^{\infty} y\, K\Big(\frac{y-Y_i}{h_n}\Big)\, dy &= \int_{-\infty}^{\infty} (Y_i + h_n z)\, K(z)\, dz \\
&= Y_i \underbrace{\int_{-\infty}^{\infty} K(z)\, dz}_{=1} + h_n \underbrace{\int_{-\infty}^{\infty} z K(z)\, dz}_{=0} = Y_i.
\end{aligned}$$

L'estimateur $\hat{m}_n(x)$ est appelé *estimateur de Nadaraya–Watson*. Il se réécrit sous la forme

$$\hat{m}_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i, \tag{III.6}$$

où on a posé

$$W_{ni}(x) := \frac{K\big(\frac{x-X_i}{h_n}\big)}{\sum_{j=1}^n K\big(\frac{x-X_j}{h_n}\big)}, \quad i = 1, \ldots, n. \tag{III.7}$$

Clairement, les quantités $W_{ni}(x)$, $i = 1, \ldots, n$, se somment à 1. Si le noyau K est à valeurs dans $\mathbb{R}^+$, ils constituent donc des *poids*, ce qui permet d'interpréter l'estimateur

de Nadaraya–Watson en (III.6) comme une *moyenne pondérée des* Y_i. Pour les noyaux habituels (lesquels sont tels que $K(x)$ est d'autant plus grand que $|x|$ est petit), le poids de l'observation Y_i dans cette moyenne pondérée sera d'autant plus grand que $|x - X_i|$ est petit, c'est-à-dire d'autant plus grand que X_i est proche de x. Il s'agit donc d'une *moyenne pondérée locale*.

Comme pour les estimateurs de densité à noyau, le choix du noyau K n'est pas prépondérant pour l'estimateur de Nadaraya–Watson, ce qui est illustré à la Figure III.3. Par contre, le choix de la fenêtre h_n aura un impact très important, ce qui est en ligne avec la Figure III.4. Pour voir comment le biais, la variance et le risque ponctuel de l'estimateur de Nadaraya–Watson dépendent de h_n, nous présentons le résultat suivant (dont la démonstration ne sera pas donnée avec autant de détails que les autres démonstrations de ce cours, notamment en ce qui concerne le contrôle des termes de reste).

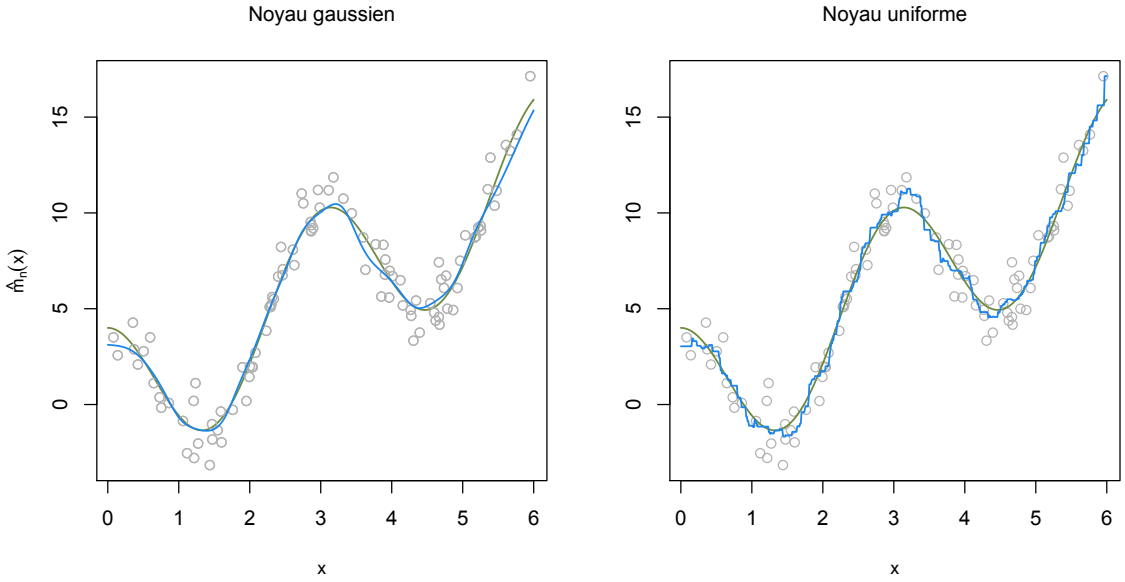

Figure III.3 – *Graphes de l'estimateur de Nadaraya–Watson* $\hat{m}_n(x)$ *calculé sur l'échantillon aléatoire de la Figure III.1, en utilisant le noyau gaussien (gauche) ou le noyau uniforme (droite), avec* $h_n = .2$ *dans les deux cas.*

Théorème III.2 *Soit* $h = h_n$ *une suite qui est* $o(1)$ *et telle que* nh_n *diverge vers l'infini. Posons* $\sigma^2(x) := \mathrm{Var}[Y|X = x] = \int_{-\infty}^{\infty}\{y - m(x)\}^2 f^{Y|X=x}(y)\,dy$. *Alors, sous des hypothèses de régularité adéquates, on a, pour tout* $x \in \mathbb{R}$,

$$\mathrm{E}[\hat{m}_n(x)] - m(x) = \frac{h_n^2}{2} d_K \Big(m''(x) + 2m'(x)\frac{(f^X)'(x)}{f^X(x)} \Big) + o(h_n^2)$$

et

$$\mathrm{Var}[\hat{m}_n(x)] = \frac{\sigma^2(x)}{nh_n f^X(x)} c_K + o\Big(\frac{1}{nh_n}\Big),$$

quand n diverge vers l'infini. Le risque ponctuel de l'estimateur de Nadaraya–Watson vérifie donc, pour tout $x \in \mathbb{R}$,

$$\mathrm{E}[\{\hat{m}_n(x)-m(x)\}^2] = \frac{h_n^4}{4} d_K^2 \Big(m''(x)+2m'(x)\frac{(f^X)'(x)}{f^X(x)} \Big)^2 + \frac{\sigma^2(x)}{nh_n f^X(x)} c_K + o(h_n^4) + o\Big(\frac{1}{nh_n}\Big),$$

quand n diverge vers l'infini.

Preuve du Théorème III.2 (esquisse). Définissons $\hat{r}_n(x)$ par la relation

$$\hat{m}_n(x) = \frac{\frac{1}{nh_n}\sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big) Y_i}{\frac{1}{nh_n}\sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big)} = \frac{\hat{r}_n(x)}{\hat{f}_n^X(x)}.$$

On a

$$\mathrm{E}[\hat{r}_n(x)] = \frac{1}{h_n}\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)Y_1\Big] = \frac{1}{h_n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K\Big(\frac{x-u}{h_n}\Big) y \underbrace{f^{(X,Y)}(u,y)}_{=f^{Y|X=u}(y)f^X(u)} \, du dy$$

$$= \frac{1}{h_n}\int_{-\infty}^{\infty} \Big(\underbrace{\int_{-\infty}^{\infty} y f^{Y|X=u}(y)\, dy}_{=m(u)} \Big) K\Big(\frac{x-u}{h_n}\Big) f^X(u)\, du = \frac{1}{h_n}\int_{-\infty}^{\infty} K\Big(\frac{x-u}{h_n}\Big) r(u)\, du,$$

où on a défini $r(x) := m(x)f^X(x)$ pour tout x. En posant $x-u = h_n z$ et en effectuant un développement de Taylor comme dans la preuve du Théorème II.3, on obtient alors, sous des hypothèses de régularité adéquates,

$$\begin{aligned}
\mathrm{E}[\hat{r}_n(x)] &= \int_{-\infty}^{\infty} K(z) r(x-h_n z)\, dz \\
&= \int_{-\infty}^{\infty} K(z)\Big\{ r(x) - h_n z r'(x) + \frac{1}{2}h_n^2 z^2 r''(x)\Big\}\, dz + o(h_n^2) \\
&= r(x) + \frac{h_n^2}{2} d_K r''(x) + o(h_n^2), \qquad \text{(III.8)}
\end{aligned}$$

où, comme d'habitude, le terme en z tombe car $\int_{-\infty}^{\infty} zK(z)\, dz = 0$. Pour ce qui est de la variance, on a

$$\mathrm{Var}[\hat{r}_n(x)] = \frac{1}{nh_n^2}\mathrm{Var}\Big[K\Big(\frac{x-X_1}{h_n}\Big)Y_1\Big] = \frac{1}{nh_n^2}\Big\{ \mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)Y_1^2\Big] - \Big(\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)Y_1\Big]\Big)^2 \Big\}$$

$$= \frac{1}{nh_n^2}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)Y_1^2\Big] - \frac{1}{n}\Big(\frac{1}{h_n}\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)Y_1\Big]\Big)^2$$

$$= \frac{1}{nh_n^2}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)Y_1^2\Big] + o\Big(\frac{1}{nh_n}\Big),$$

où on a utilisé le fait que

$$\frac{1}{n}\Big(\frac{1}{h_n}\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)Y_1\Big]\Big)^2 = \frac{1}{n}(\mathrm{E}[\hat{r}_n(x)])^2 = \frac{1}{n}(r(x)+o(1))^2 = O\Big(\frac{1}{n}\Big) = o\Big(\frac{1}{nh_n}\Big).$$

En posant

$$m_2(x) := \mathrm{E}[Y^2|X=x] = \int_{-\infty}^{\infty} y^2 f^{Y|X=x}(y)\,dy,$$

on obtient donc

$$\begin{aligned}
\mathrm{Var}[\hat{r}_n(x)] &= \frac{1}{nh_n^2}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)Y_1^2\Big] + o\Big(\frac{1}{nh_n}\Big) \\
&= \frac{1}{nh_n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K^2\Big(\frac{x-u}{h_n}\Big)y^2 \underbrace{f^{(X,Y)}(u,y)}_{=f^{Y|X=u}(y)f^X(u)}\,dudy + o\Big(\frac{1}{nh_n}\Big) \\
&= \frac{1}{nh_n^2}\int_{-\infty}^{\infty} K^2\Big(\frac{x-u}{h_n}\Big)\Big(\underbrace{\int_{-\infty}^{\infty} y^2 f^{Y|X=u}(y)\,dy}_{=m_2(u)}\Big)f^X(u)\,du + o\Big(\frac{1}{nh_n}\Big),
\end{aligned}$$

ce qui en effectuant le même changement de variable que ci-dessus et en faisant un développement de Taylor, livre

$$\begin{aligned}
\mathrm{Var}[\hat{r}_n(x)] &= \frac{1}{nh_n}\int_{-\infty}^{\infty} K^2(z)m_2(x-h_nz)f^X(x-h_nz)\,dz + o\Big(\frac{1}{nh_n}\Big) \\
&= \frac{1}{nh_n}\int_{-\infty}^{\infty} K^2(z)\big\{m_2(x)+o(1)\big\}\big\{f^X(x)+o(1)\big\}\,dz + o\Big(\frac{1}{nh_n}\Big) \\
&= \frac{1}{nh_n}c_K m_2(x)f^X(x) + o\Big(\frac{1}{nh_n}\Big).
\end{aligned}$$

Conjointement avec (III.8), ceci donne

$$\begin{aligned}
\mathrm{E}[\{\hat{r}_n(x)-r(x)\}^2] &= \big\{\mathrm{E}[\hat{r}_n(x)]-r(x)\big\}^2 + \mathrm{Var}[\hat{r}_n(x)] \\
&= \frac{h_n^4}{4}d_K^2(r''(x))^2 + \frac{1}{nh_n}c_K m_2(x)f^X(x) + o(h_n^4) + o\Big(\frac{1}{nh_n}\Big),
\end{aligned}$$

ce qui montre que $\hat{r}_n(x)$ converge vers $r(x)$ en probabilité. Puisque, par ailleurs, $\hat{f}^X(x)$ converge vers $f^X(x)$ en probabilité, on en déduit que $\hat{m}_n(x)$ converge en probabilité vers $m(x)$ (autrement dit, l'estimateur de Nadaraya–Watson est faiblement convergent pour la fonction de régression $m(x)$ inconnue).

En adoptant la notation classique des "o_P" (la suite de variables aléatoires (Z_n) est $o_P(s_n)$ si la suite (Z_n/s_n) converge vers zéro en probabilité ; voir la Section 2.2 de [5]

pour une discussion), la décomposition

$$\begin{aligned}\hat{m}_n(x) - m(x) &= (\hat{m}_n(x) - m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\Big(1 + \frac{f^X(x) - \hat{f}_n^X(x)}{\hat{f}_n^X(x)}\Big) \\ &= (\hat{m}_n(x) - m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\big(1 + o_P(1)\big) \qquad \text{(III.9)}\end{aligned}$$

suggère que

$$\mathrm{E}[\hat{m}_n(x) - m(x)] = \mathrm{E}\Big[(\hat{m}_n(x) - m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\Big] + R_n$$

et que

$$\mathrm{Var}[\hat{m}_n(x)] = \mathrm{Var}\Big[(\hat{m}_n(x) - m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\Big] + S_n,$$

où R_n et S_n sont des termes de reste (d'ordre inférieur). Nous allons utiliser ceci pour calculer le biais et la variance de $\hat{m}_n(x)$. Commençons par le biais. Puisque

$$\begin{aligned}(\hat{m}_n(x) - m(x))\frac{\hat{f}_n^X(x)}{f^X(x)} &= \frac{1}{f^X(x)}\big(\hat{r}_n(x) - m(x)\hat{f}_n^X(x)\big) \\ &= \frac{1}{nh_n f^X(x)}\sum_{i=1}^n K\Big(\frac{x - X_i}{h_n}\Big)(Y_i - m(x)),\end{aligned}$$

on a

$$\begin{aligned}\mathrm{E}\Big[(\hat{m}_n(x) - m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\Big] &= \frac{1}{f^X(x)}\big(\mathrm{E}\big[\hat{r}_n(x)\big] - m(x)\mathrm{E}\big[\hat{f}_n^X(x)\big]\big) \\ &= \frac{1}{f^X(x)}\Big\{\Big(r(x) + \frac{h_n^2}{2}d_K r''(x) + o(h_n^2)\Big) - m(x)\Big(f^X(x) + \frac{h_n^2}{2}(f^X)''(x)d_K + o(h_n^2)\Big)\Big\} \\ &= \frac{h_n^2}{2}d_K\Big(\frac{r''(x)}{f^X(x)} - m(x)\frac{(f^X)''(x)}{f^X(x)}\Big) + o(h_n^2) \\ &= \frac{h_n^2}{2}d_K\Big(m''(x) + 2m'(x)\frac{(f^X)'(x)}{f^X(x)}\Big) + o(h_n^2),\end{aligned}$$

ce qui fournit

$$\mathrm{E}[\hat{m}_n(x)] - m(x) = \frac{h_n^2}{2}d_K\Big(m''(x) + 2m'(x)\frac{(f^X)'(x)}{f^X(x)}\Big) + o(h_n^2). \qquad \text{(III.10)}$$

Pour ce qui est de la variance,

$$\begin{aligned}\mathrm{Var}\Big[(\hat{m}_n(x)-m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\Big] &= \frac{1}{nh_n^2(f^X(x))^2}\mathrm{Var}\Big[K\Big(\frac{x-X_1}{h_n}\Big)(Y_1-m(x))\Big]\\ &= \frac{1}{nh_n^2(f^X(x))^2}\Big\{\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)(Y_1-m(x))^2\Big]-\Big(\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)(Y_1-m(x))\Big]\Big)^2\Big\}\\ &= \frac{1}{nh_n^2(f^X(x))^2}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)(Y_1-m(x))^2\Big]+o\Big(\frac{1}{nh_n}\Big),\end{aligned}$$

où on a utilisé le fait que

$$\frac{1}{nh_n^2}\Big(\mathrm{E}\Big[K\Big(\frac{x-X_1}{h_n}\Big)(Y_1-m(x))\Big]\Big)^2=\frac{1}{n}\big(\mathrm{E}\big[\hat{r}_n(x)\big]-m(x)\mathrm{E}\big[\hat{f}_n^X(x)\big]\big)^2=O\Big(\frac{1}{n}\Big)=o\Big(\frac{1}{nh_n}\Big).$$

Donc

$$\begin{aligned}\mathrm{Var}\Big[(\hat{m}_n(x)-m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\Big] &= \frac{1}{nh_n^2(f^X(x))^2}\mathrm{E}\Big[K^2\Big(\frac{x-X_1}{h_n}\Big)(Y_1-m(x))^2\Big]+o\Big(\frac{1}{nh_n}\Big)\\ &= \frac{1}{nh_n^2(f^X(x))^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K^2\Big(\frac{x-u}{h_n}\Big)(y-m(x))^2f^{Y|X=u}(y)f^X(u)\,dydu+o\Big(\frac{1}{nh_n}\Big)\\ &= \frac{1}{nh_n^2(f^X(x))^2}\int_{-\infty}^{\infty}K^2\Big(\frac{x-u}{h_n}\Big)\underbrace{\Big(\int_{-\infty}^{\infty}\{y-m(x)\}^2f^{Y|X=u}(y)\,dy\Big)}_{=m_2(u)-2m(x)m(u)+m^2(x)}f^X(u)\,du+o\Big(\frac{1}{nh_n}\Big).\end{aligned}$$

En posant $x-u=h_nz$, il vient

$$\begin{aligned}&\mathrm{Var}\Big[(\hat{m}_n(x)-m(x))\frac{\hat{f}_n^X(x)}{f^X(x)}\Big]\\ &= \frac{1}{nh_n(f^X(x))^2}\int_{-\infty}^{\infty}K^2(z)\{m_2(x-h_nz)-2m(x)m(x-h_nz)+m^2(x)\}f^X(x-h_nz)\,dz+o\Big(\frac{1}{nh_n}\Big)\\ &= \frac{1}{nh_n(f^X(x))^2}\int_{-\infty}^{\infty}K^2(z)\{m_2(x)-m^2(x)\}f^X(x)\,dz+o\Big(\frac{1}{nh_n}\Big)\\ &= \frac{\sigma^2(x)}{nh_nf^X(x)}c_K+o\Big(\frac{1}{nh_n}\Big),\end{aligned}$$

puisque $\sigma^2(x)=m_2(x)-m^2(x)$, ce qui fournit

$$\mathrm{Var}[\hat{m}_n(x)]=\frac{\sigma^2(x)}{nh_nf^X(x)}c_K+o\Big(\frac{1}{nh_n}\Big). \tag{III.11}$$

L'expression du risque ponctuel (qui, comme d'habitude, est la somme du biais au carré et de la variance) découle directement de (III.10) et de (III.11). □

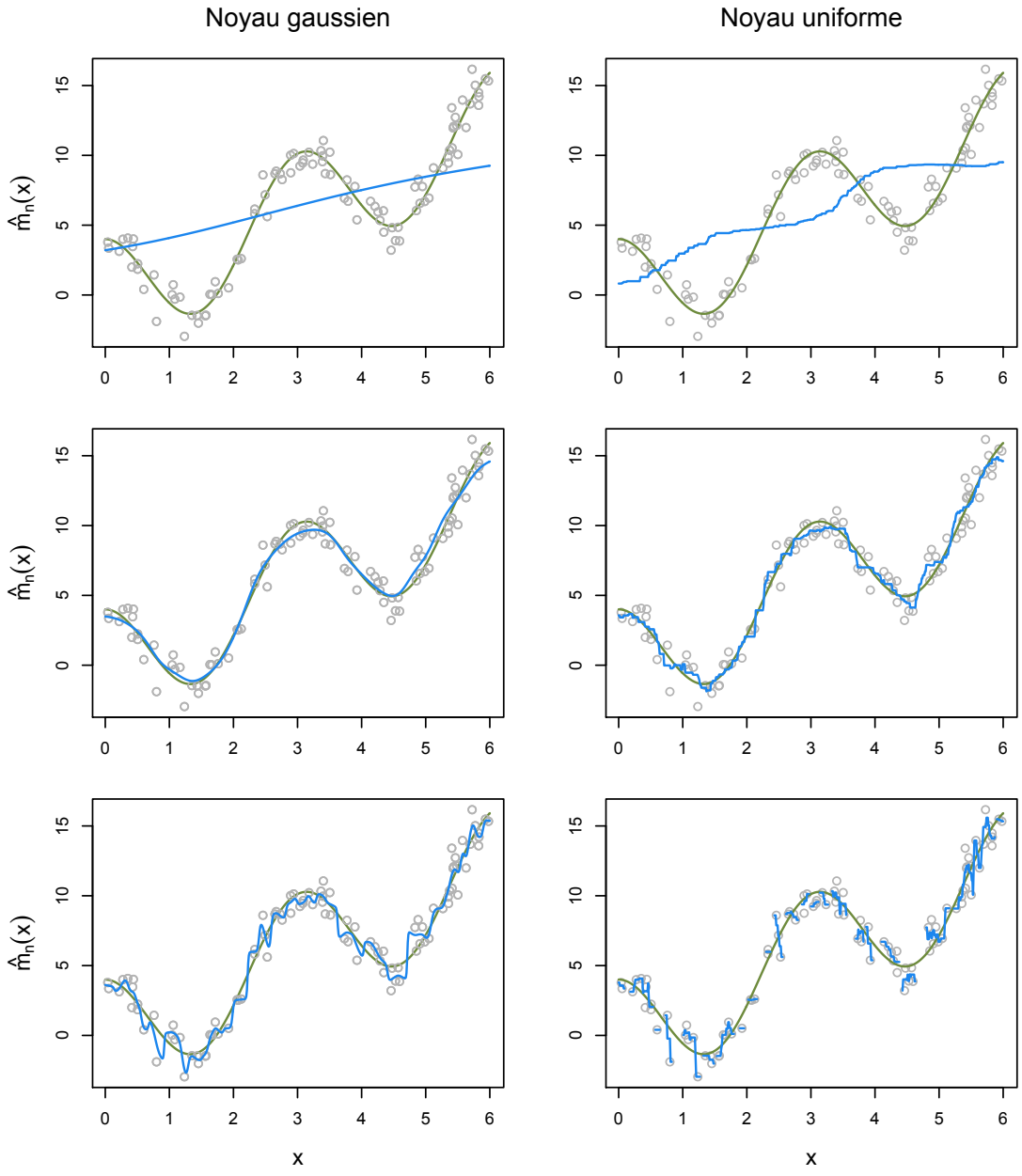

Figure III.4 – *(Gauche :) Graphes des estimateurs de Nadaraya–Watson $\hat{m}_n(x)$ calculés sur l'échantillon de la Figure III.1, en utilisant le noyau gaussien et, avec $h_0 = .2$, une taille de fenêtre $h = 10h_0$ (haut), $h = h_0$ (milieu) et $h = h_0/5$ (bas). (Droite :) Les graphes correspondants obtenus en utilisant le noyau uniforme et les mêmes valeurs de h_0.*

En procédant comme pour l'estimation de densité par noyau, on peut déduire de ce théorème que la taille de fenêtre optimale est de la forme $h_n = cn^{-1/5}$, ce qui mène à un risque ponctuel optimal qui tend vers zéro au taux $n^{-4/5}$ (ceci peut paraître, étonnant à

première vue, puisque l'estimateur de Nadaraya–Watson a été fondé notamment sur un estimateur de $f^{(X,Y)}$ pour lequel le taux de convergence est moins rapide). Sans rentrer dans les détails, nous mentionnons ici que des méthodes existent pour choisir de façon adéquate la taille de la fenêtre h_n en pratique (en particulier, il existe des méthodes de type validation croisée, qui sont l'équivalent des méthodes que nous avons étudiées au chapitre précédent pour l'estimation de densité).

III.2 Estimation par polynômes locaux

Une autre méthode d'estimation de $m(x)$, qui, comme on le verra, est en lien avec la méthode d'estimation par noyau, consiste à effectuer—de façon locale à x—une régression constante, linéaire, ou plus généralement polynomiale. Pour rappel, la régression (globale) polynomiale d'ordre p est une méthode de nature paramétrique, qui estime $m(x)$ par

$$\hat{m}_n(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \ldots + \hat{\beta}_p x^p,$$

où $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ est solution du problème des moindres carrés

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i + \ldots + \beta_p X_i^p)\}^2.$$

Pour $p = 0$, ceci estime $m(x)$ par $\hat{m}(x) = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ (indépendamment de x), tandis que, pour $p = 1$, ceci est l'estimateur linéaire de $m(x)$ associé à la droite des moindres carrées habituelle. Pour un p fixé, cet estimateur global ne sera convergent que si la fonction m inconnue est elle-même un polynôme de degré p (ou moins).

La régression polynomiale *locale* d'ordre p consiste plutôt à estimer $m(x)$ par

$$\hat{m}_n^{(p)}(x) = \hat{\beta}_{x0} + \hat{\beta}_{x1} x + \ldots + \hat{\beta}_{xp} x^p,$$

où $\hat{\beta}_x = (\hat{\beta}_{x0}, \hat{\beta}_{x1}, \ldots, \hat{\beta}_{xp})$ est solution du problème *local* des moindres carrés

$$\hat{\beta}_x = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} W_{ni}(x) \{Y_i - (\beta_0 + \beta_1 X_i + \ldots + \beta_p X_i^p)\}^2,$$

qui est basé sur les poids $W_{ni}(x)$ en (III.7). Ce sont ces poids qui localisent la méthode au point x : ils font en sorte qu'une observation (X_i, Y_i) ne jouera un rôle important dans la minimisation de la somme de carrés que si X_i est proche de x. La régression polynomiale locale est une méthode non paramétrique, qui fournit un estimateur convergent de m dès que certaines conditions de régularité sont satisfaites. La Figure III.5 illustre la régression polynomiale locale pour $p = 0$, $p = 1$ et $p = 2$.

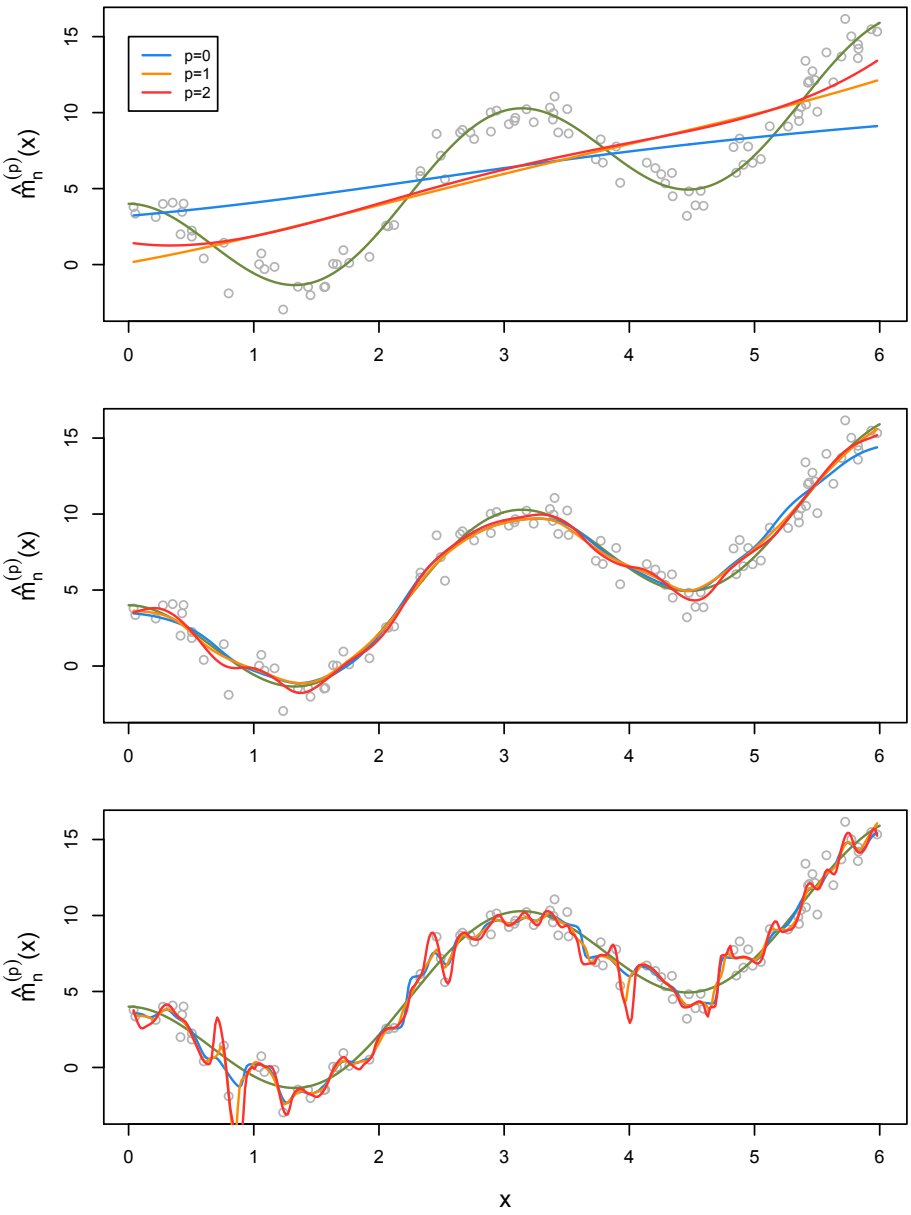

Figure III.5 – *Graphes des estimateurs par régression polynomiale locale* $\hat{m}_n^{(p)}(x)$*, pour* $p = 0, 1, 2$*, calculés sur l'échantillon de la Figure III.1, en utilisant le noyau gaussien et, avec* $h_0 = .2$*, une taille de fenêtre* $h = 10h_0$ *(haut),* $h = h_0$ *(milieu) et* $h = h_0/4$ *(bas).*

Comme on propose de le vérifier dans les exercices de ce chapitre, l'estimateur par polynômes locaux obtenu pour $p = 0$ (régression localement constante) coïncide avec l'estimateur de Nadaraya–Watson. Donc la méthode de régression non paramétrique par polynômes locaux généralise en fait celle de Nadaraya–Watson. On peut se demander quel est l'avantage à utiliser $p > 0$ par rapport à $p = 0$ (Nadaraya–Watson). Le résultat suivant, qui se focalise sur le cas $p = 1$ (régression localement linéaire) permet de répondre

à cette question (nous renvoyons à [1] pour une preuve).

Théorème III.3 *Soit $h = h_n$ une suite qui est $o(1)$ et telle que nh_n diverge vers l'infini. Alors, sous des hypothèses de régularité adéquates, on a, pour tout $x \in \mathbb{R}$,*

$$\mathrm{E}[\hat{m}_n^{(1)}(x)] - m(x) = \frac{h_n^2}{2} d_K m''(x) + o(h_n^2)$$

et

$$\mathrm{Var}[\hat{m}_n^{(1)}(x)] = \frac{\sigma^2(x)}{nh_n f^X(x)} c_K + o\Big(\frac{1}{nh_n}\Big) \quad \Big(= \mathrm{Var}[\hat{m}_n^{(0)}(x)]\Big),$$

quand n diverge vers l'infini. Le risque ponctuel de l'estimateur $\hat{m}_n^{(1)}(x)$ vérifie donc, pour tout $x \in \mathbb{R}$,

$$\mathrm{E}[\{\hat{m}_n^{(1)}(x) - m(x)\}^2] = \frac{h_n^4}{4} d_K^2 \big(m''(x)\big)^2 + \frac{\sigma^2(x)}{nh_n f^X(x)} c_K + o(h_n^4) + o\Big(\frac{1}{nh_n}\Big),$$

quand n diverge vers l'infini.

Ce résultat montre que, contrairement à l'estimateur de Nadaraya–Watson $\hat{m}_n^{(0)}(x)$, l'estimateur $\hat{m}_n^{(1)}(x)$ présente un biais qui ne dépend pas de la densité f^X de la variable explicative, donc du "design" (le design désigne la configuration-type prise par les variables X_i). On peut aussi vérifier que si f^X est à support compact (de support $[a, b]$, disons), alors le biais de l'estimateur au bord du domaine (c'est-à-dire le biais de $\hat{m}_n^{(1)}(a)$ et $\hat{m}_n^{(1)}(b)$) est un $O(h_n^2)$ comme à l'intérieur du support, alors que le bais correspondant pour l'estimateur de Nadaraya–Watson est un $O(h_n)$ (on dit que l'estimateur $\hat{m}_n^{(0)}(x)$ souffre d'un *effet de bord*).

Parfois, l'estimateur résultant de la régression polynomiale locale d'ordre p est plutôt défini sous la forme

$$\hat{m}_n^{(p)}(x) = \hat{\gamma}_{x0},$$

où $\hat{\gamma}_x = (\hat{\gamma}_{x0}, \hat{\gamma}_{x1}, \ldots, \hat{\gamma}_{xp})$ est solution du problème local des moindres carrés

$$\hat{\gamma}_x = \arg \min_{\gamma \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} W_{ni}(x) \{Y_i - (\gamma_0 + \gamma_1 (X_i - x) + \ldots + \gamma_p (X_i - x)^p)\}^2,$$

qui est encore basé sur les poids $W_{ni}(x)$ en (III.7). On vérifiera aisément que les deux définitions de régression polynomiale locale sont équivalentes, dans le sens où elles mènent au même estimateur $\hat{m}_n^{(p)}(x)$. Cette nouvelle définition est motivée par le fait qu'on peut approximer la fonction de régression inconnue m au voisinage de x par le développement

limité

$$m(t) \approx m(x) + m'(x)(t-x) + \frac{1}{2}m''(x)(t-x)^2 + \ldots + \frac{1}{p!}m^{\underline{p}}(x)(t-x)^p,$$

où $m^{\underline{k}}(x)$ représente la $k$ème dérivée de m au point x. Ceci rend clair non seulement que $\hat{\gamma}_{0x}$ est un estimateur $m(x)$, mais aussi que $\hat{\gamma}_{1x}$ est un estimateur de $m'(x)$, que $\hat{\gamma}_{2x}$ est un estimateur de $m''(x)/2$, ..., et que $\hat{\gamma}_{px}$ est un estimateur de $m^{\underline{p}}(x)/(p!)$. L'avantage principal de cette nouvelle définition est donc qu'elle fournit aussi des estimateurs des dérivées de m.

III.3 Quelques autres méthodes d'estimation

Dans cette section, nous décrivons quelques autres méthodes classiques d'estimation en régression non paramétrique, sans entrer dans autant de détails que pour les méthodes étudiées dans les Sections III.1 et III.2.

III.3.1 Estimation par plus proches voisins

Comme on l'a vu en (III.6), l'estimateur de Nadaraya–Watson $\hat{m}_n(x)$ peut s'interpréter comme une moyenne locale (pondérées) des observations. En particulier, quand on utilise le noyau uniforme $K(z) = \frac{1}{2}\mathbb{I}[z \in (-1,1)]$, $\hat{m}_n(x)$ est simplement la moyenne arithmétique des Y_i pour lesquels le X_i associé appartient à l'intervalle $(x-h_n, x+h_n)$, ce qui justifie une nouvelle fois qu'on appelle h_n la taille de la fenêtre. L'estimation par plus proches voisins est également une moyenne locale, mais qui est cette fois fondée sur un nombre fixe, k disons, de Y_i. Plus précisément, l'estimateur des plus proches voisins associé à l'entier positif k est défini par

$$\hat{m}_{n,k}(x) := \frac{1}{k}\sum_{i \in \mathcal{I}_k(x)}^{n} Y_i,$$

où

$$\mathcal{I}_k(x) := \{i = 1, \ldots, n : X_i \text{ est un des } k \ X_j \text{ les plus proches de } x\}.$$

La différence principale avec l'estimateur de Nadaraya–Watson est que la "fenêtre effective" associée à l'estimateur des plus proches voisins dépend de x : si $f(x)$ est grand, alors il y aura beaucoup de X_i proches de x et les k plus proches voisins sur lesquels sera fondée l'estimation $\hat{m}_{n,k}(x)$ de $m(x)$ appartiendront à une petite fenêtre comprenant x ; au contraire, si $f(x)$ est petit, alors il y aura peu de X_i proches de x et les k plus proches voisins sur lesquels sera fondée l'estimation $\hat{m}_{n,k}(x)$ de $m(x)$ appartiendront à une fenêtre plus étendue.

Bien entendu, la convergence requiert que $k = k_n$ diverge vers l'infini, pour que le mécanisme de la loi des grands nombres puisse opérer. Mais il est aussi clair que k_n ne peut pas diverger vers l'infini trop vite, au risque que la fenêtre effective soit trop grande et que les plus proches "voisins" ne soient plus des voisins du point x, ce qui augmentera le biais de l'estimateur (en particulier, si on prend $k_n = n$, on aura $\hat{m}_{n,k}(x) = \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ indépendamment de x, de sorte que la fonction $\hat{m}_{n,k}$ estimée sera constante quelles que soient les observations (X_i, Y_i) !) Cette intuition est confirmée par le résultat suivant, qui décrit le biais, la variance et le risque ponctuel de l'estimateur des plus proches voisins.

Théorème III.4 *Soit $k = k_n$ une suite qui diverge vers l'infini de telle sorte que $k_n = o(n)$. Alors, sous des hypothèses de régularité adéquates, on a, pour tout $x \in \mathbb{R}$,*

$$\mathrm{E}[\hat{m}_{n,k}(x)] - m(x) = \frac{d_K}{8(f^X(x))^2}\Big(m''(x) + 2m'(x)\frac{(f^X)'(x)}{f^X(x)}\Big)\frac{k_n^2}{n^2} + o\Big(\frac{k_n^2}{n^2}\Big)$$

et

$$\mathrm{Var}[\hat{m}_{n,k}(x)] = \frac{2\sigma^2(x)}{k_n}c_K + o\Big(\frac{1}{k_n}\Big),$$

quand n diverge vers l'infini. Le risque ponctuel associé est donc

$$\begin{aligned}&\mathrm{E}[\{\hat{m}_{n,k}(x) - m(x)\}^2] \\ &\quad = \frac{d_K^2}{64(f^X(x))^4}\Big(m''(x) + 2m'(x)\frac{(f^X)'(x)}{f^X(x)}\Big)^2\frac{k_n^4}{n^4} + \frac{2\sigma^2(x)}{k_n}c_K + o\Big(\frac{k_n^4}{n^4}\Big) + o\Big(\frac{1}{k_n}\Big),\end{aligned}$$

quand n diverge vers l'infini.

Ce résultat confirme l'intuition qu'une valeur trop grande de k_n mènera à un grand biais, mais à une petite variance, alors qu'une valeur trop petite de k_n mènera au contraire à un biais réduit, mais à une grande variance. L'entier k_n joue donc pour cette méthode d'estimation le rôle de paramètre de lissage, dont il faudra choisir la valeur de sorte à réaliser un équilibre entre biais et variance. La Figure III.6 illustre ceci sur l'exemple considéré dans les sections précédentes.

III.3.2 Estimation par splines

Si on laisse de côté l'interprétation de la fonction de régression $m(x)$ en termes de moyenne conditionnelle de Y sachant X, nous sommes typiquement à la recherche d'une fonction $\hat{m}$ qui vise à interpoler les observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, dans le sens où la somme des carrés résiduelle

$$S_n(g) := \sum_{i=1}^n \{Y_i - g(X_i)\}^2$$

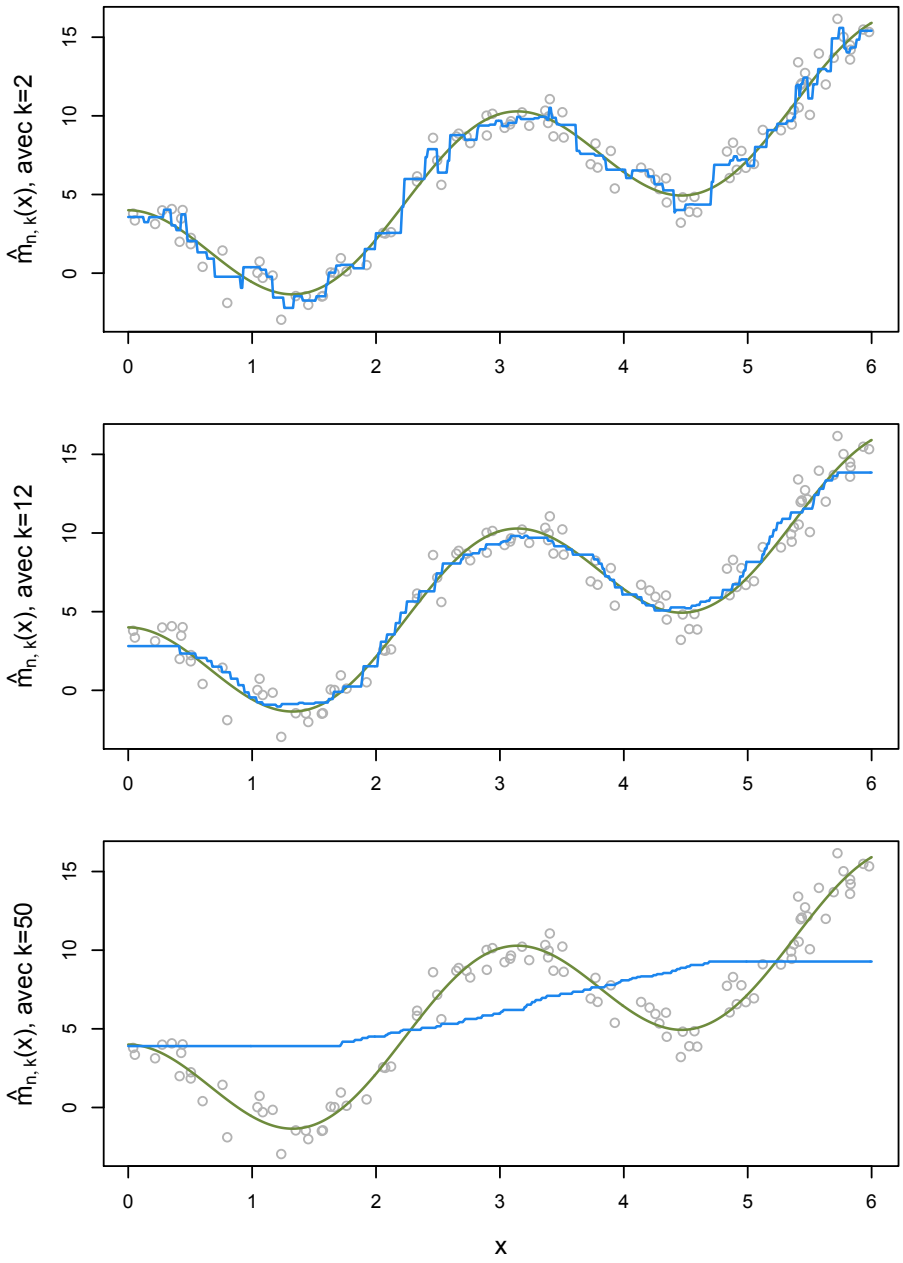

Figure III.6 – *Graphes des estimateurs par plus proches voisins $\hat{m}_{n,k}(x)$ calculés sur l'échantillon de la Figure III.1, en utilisant $k = 2$ voisins (haut), $k = 12$ voisins (milieu), et $k = 50$ voisins (bas).*

sera petite pour $g = \hat{m}$. Ceci suggère de choisir pour $\hat{m}$ la fonction g qui minimise $S_n(g)$. Sous l'hypothèse que les X_i sont indépendants et admettent la même fonction de densité f^X, ces X_i seront presque sûrement distincts deux à deux, de sorte qu'il existe des fonctions g (en fait, une infinité de fonctions g) qui fournissent la valeur minimale (nulle) de $S_n(g)$. Non seulement procéder de la sorte n'assure donc pas une solution unique

pour $\hat{m}$, mais les fonctions auxquelles ce principe conduit sont peu satisfaisantes dans le sens où elles se contentent d'interpoler de façon parfaite toutes les observations (X_i, Y_i), ce qui résultera en des fonctions très peu régulières menant à des prédictions médiocres (*overfitting*).

Dans ce cadre, il est classique de modifier la fonction objectif $S_n(g)$ de façon à pénaliser l'éventuel manque de régularité de g. La solution la plus classique consiste à considérer la fonction objectif

$$S_{n,\lambda}(g) := \sum_{i=1}^{n} \{Y_i - g(X_i)\}^2 + \lambda \|g''\|_2^2, \qquad \text{avec } \|g''\|_2^2 := \int_a^b (g''(x))^2 \, dx,$$

où λ est un paramètre réel strictement positif fixé et où g'' désigne la seconde dérivée de g. Bien entendu, ceci suppose tacitement qu'on se restreigne à l'ensemble $W_2([a,b])$ des fonctions $g : [a,b] \to \mathbb{R}$ qui sont deux fois dérivables sur $[a,b]$ et pour lesquelles on a $\|g''\|_2 < \infty$ (en pratique, on prend pour a le minimium des X_i et pour b le maximum des X_i). Pour λ petit, le terme de pénalisation va jouer un rôle négligeable, de sorte que la fonction minimisant $S_{n,\lambda}(g)$—notons la $\hat{m}_{n,\lambda}$—sera proche d'une fonction interpolant toutes les observations (X_i, Y_i), ce qui mènera typiquement à un petit biais et à une grande variance. Au contraire, pour λ grand, le terme de pénalisation va dominer, ce qui va conduire à prendre $\hat{m}_{n,\lambda}$ linéaire (puisque ce n'est que pour $g(x) = \beta_0 + \beta_1 x$ qu'on aura $\|g''\|_2 = 0$) ; la fonction $\hat{m}_{n,\lambda}$ qui en résultera sera très régulière mais sera loin de bien interpoler les observations, ce qui mènera à une variance faible mais un biais important. Le paramètre λ joue donc le rôle de paramètre de lissage pour cette méthode. Ceci est illustré à la Figure III.7.

On peut montrer que la solution $\hat{m}_{n,\lambda}$ au problème d'optimisation ci-dessus est ce qu'on appelle une *spline cubique*, c'est-à-dire une fonction de classe C^2 sur $[a,b]$ qui est C^3 par morceaux (les morceaux étant en fait les intervalles joignant deux X_i consécutifs). C'est bien entendu ce point crucial qui donne le nom à cette méthode d'estimation non paramétrique.

III.3.3 Estimation par projection

Finalement, nous considérons très brièvement une méthode d'estimation qui est le pendant, pour la régression non paramétrique, de la méthode de projection présentée dans la Section II.4 pour l'estimation de densité. On supposera ici que la fonction de régression m peut être approximée de façon satisfaisante par une fonction de la forme

$$\tilde{m}_{c_1,\ldots,c_m}(x) = \sum_{r=1}^{m} c_r g_r(x),$$

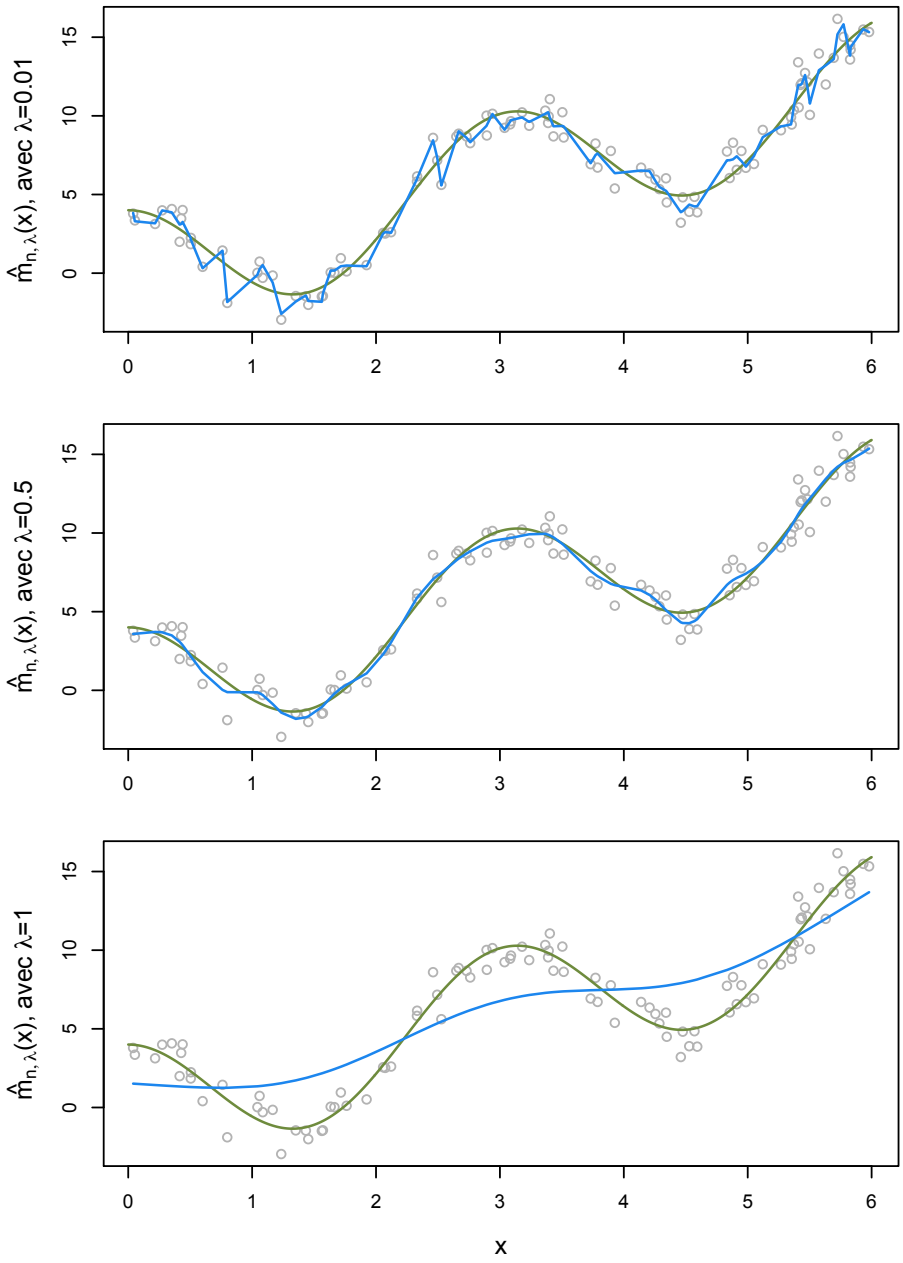

Figure III.7 – *Graphes des estimateurs par splines cubiques $\hat{m}_{n,\lambda}(x)$ calculés sur l'échantillon de la Figure III.1, en utilisant $\lambda = .01$ (haut), $\lambda = .5$ (milieu), et $\lambda = 1$ (bas).*

où $g_1, \ldots, g_m$ sont des fonctions fixées et où $c_1, \ldots, c_m$ sont des paramètres réels (on ne confondra pas, évidemment, la fonction m à estimer avec le nombre m de fonctions g_r considérées). Il est parfois avantageux de prendre des fonctions g_r qui, comme dans la Section II.4, forment une base orthonormale de l'espace des fonctions qu'elles engendrent (même si c'est moins crucial que pour l'estimation de densité). Ainsi, les g_r peuvent être, par exemple, des fonctions choisies dans une base de Fourier ou dans une base

d'ondelettes. La méthode la plus classique pour estimer les paramètres c_r consiste à minimiser la fonction objectif

$$\sum_{i=1}^{n} \left\{ Y_i - \tilde{m}_{c_1,\ldots,c_m}(X_i) \right\}^2 \tag{III.12}$$

en $c_1, \ldots, c_m$. D'autres méthodes sont disponibles, mais elles sont particulières à la base de fonctions—dans ce cas, *orthonormale*—choisie, et ce sont celles-là qui donnent son sens à la terminologie d'*estimation par projection*. Bien entendu, c'est l'entier m qui joue le rôle de paramètre de lissage pour cette méthode : pour m petit, on aura peu de paramètres à estimer, ce qui assurera une faible variance, mais la fonction qu'on estime sera (même dans une situation où les c_r seraient estimés sans erreur) une très pauvre approximation de la fonction m à estimer, ce qui résultera en un grand biais. Au contraire, pour m grand, cette approximation sera précise, donc le biais faible, mais le nombre important de coefficients c_r à estimer conduira à une variance importante. C'est donc à nouveau un équilibre biais-variance qui guidera le choix du nombre m de fonctions à considérer.

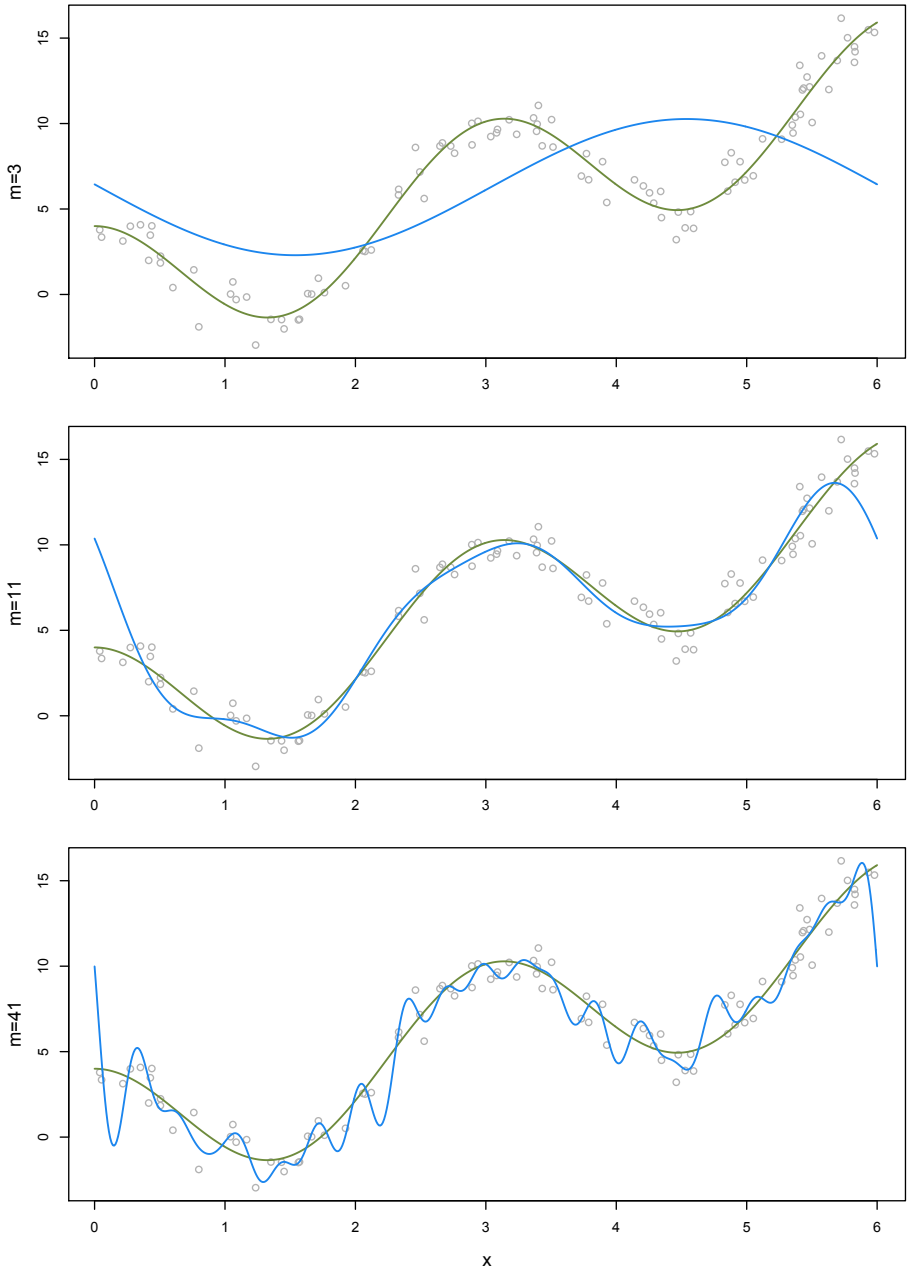

Figure III.8 – *Graphes des estimateurs par projection minimisant (III.12), calculés sur l'échantillon de la Figure III.1, en utilisant, avec* $a = 0$ *et* $b = 6$*, la base trigonométrique associée à* $m = 3$ *(haut),* $m = 11$ *(centre), et* $m = 41$ *(bas).*

III.4 Codes R

Figure III.1

```
rm(list = ls())
#########
mexemple<-function(z){return(2*z+4*cos(2*z)-sin(2*z))}
#########
a=0
b=6
n=100
sigma=1
set.seed(1)
x=runif(n,a,b)
y=mexemple(x)+sigma*rnorm(n)
z=seq(a,b,by=.01)
par(mfrow=c(1,1))
par(oma=c(2,2,2,1))
par(mar=c(3,3,1,1))
plot(x,y,main="",col=grey(.7),xlab="",ylab="")
points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
mtext("x",side=1,line=3,cex=1.2)
mtext(expression(m(x)),side=2,line=2.6,cex=1.2)
```

Figure III.2

```
rm(list = ls())
#######
library(MASS)
n=1000
hn=1
f<-function(x,y) {return(dnorm(x,0,1)*dnorm(y,0,1))}
######
set.seed(1)
X=array(rnorm(2*n),c(n,2))
######
x=seq(min(X[,1]),max(X[,1]),length=51)
y=seq(min(X[,2]),max(X[,2]),length=51)
z=outer(x,y,f)
par(mfrow=c(1,2))
par(oma=c(1,1,1,1))
par(mar=c(1,1,1,1))
persp(x,y,z,theta=30,phi=30,expand=0.7,xlab="x",ylab="y",zlab="z",col="darkolivegreen2")
est=kde2d(X[,1],X[,2],n=100,h=hn)
persp(est$x,est$y,est$z,theta=30,phi=30,expand=0.7,xlab="x",ylab="y",zlab="z",col="steelblue1")
```

Figure III.3

```
rm(list = ls())
#########
mexemple<-function(z){return(2*z+4*cos(2*z)-sin(2*z))}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
KUnif<-function(z){return( (1/2)*(z<=1)*(z>=-1) )}
```

```
#########
a=0
b=6
n=100
sigma=1
hGauss=.2
hUnif=.2
set.seed(1)
x=runif(n,a,b)
y=mexemple(x)+sigma*rnorm(n)
z=seq(a,b,by=.01)
mhatGauss=rep(0,length=length(z))
mhatUnif=rep(0,length=length(z))
for (r in (1:length(z)))
        {
        mhatGauss[r]=sum( KGauss((z[r]-x)/hGauss)*y )/sum( KGauss((z[r]-x)/hGauss) )
        mhatUnif[r]=sum( KUnif((z[r]-x)/hUnif)*y )/sum( KUnif((z[r]-x)/hUnif) )
        }
par(mfrow=c(1,2))
par(oma=c(2,1,2,1))
par(mar=c(3,4,1,1))
plot(x,y,main="",col=grey(.7),xlab="",ylab="")
points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
points(x,y,main="",col=grey(.7),xlab="",ylab="")
points(z,mhatGauss,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
        mtext("Noyau␣gaussien",side=3,line=1.4,cex=1)
        mtext("x",side=1,line=3,cex=1.2)
        mtext(expression(hat(m)[n](x)),side=2,line=2.6,cex=1.2)
plot(x,y,main="",col=grey(.7),xlab="",ylab="")
points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
points(z,mhatUnif,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
        mtext("Noyau␣uniforme",side=3,line=1.4,cex=1)
        mtext("x",side=1,line=3,cex=1.2)
```

Figure III.4

```
rm(list = ls())
#########
mexemple<-function(z){return(2*z+4*cos(2*z)-sin(2*z))}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
KUnif<-function(z){return( (1/2)*(z<=1)*(z>=-1) )}
#########
a=0
b=6
n=100
sigma=1
hGauss=.2
hUnif=.2
set.seed(4)
hGaussvec=.2*c(10,1,1/5)
hUnifvec=.2*c(10,1,1/5)
x=runif(n,a,b)
y=mexemple(x)+sigma*rnorm(n)
z=seq(a,b,by=.01)
mhatGauss=rep(0,length=length(z))
mhatUnif=rep(0,length=length(z))
par(mfrow=c(length(hGaussvec),2))
```

```
par(oma=c(2,1,2,1))
par(mar=c(3,4,1,1))
for (i in (1:length(hGaussvec)))
        {
        hGauss=hGaussvec[i]
        hUnif=hUnifvec[i]
        for (r in (1:length(z)))
                {
                mhatGauss[r]=sum( KGauss((z[r]-x)/hGauss)*y )/sum( KGauss((z[r]-x)/hGauss) )
                mhatUnif[r]=sum( KUnif((z[r]-x)/hUnif)*y )/sum( KUnif((z[r]-x)/hUnif) )
                }
        plot(x,y,main="",col=grey(.7),xlab="",ylab="")
        points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        points(x,y,main="",col=grey(.7),xlab="",ylab="")
        points(z,mhatGauss,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
                if (i==1) {mtext("Noyau␣gaussien",side=3,line=1.4,cex=1)}
                if (i==length(hGaussvec)) {mtext("x",side=1,line=3,cex=.85)}
                mtext(expression(hat(m)[n](x)),side=2,line=2.6,cex=.85)
        plot(x,y,main="",col=grey(.7),xlab="",ylab="")
        points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        points(z,mhatUnif,typ="l",lwd=1.4,main="",col="dodgerblue2",xlab="",ylab="")
                if (i==1) {mtext("Noyau␣uniforme",side=3,line=1.4,cex=1)}
                if (i==length(hGaussvec)) {mtext("x",side=1,line=3,cex=.85)}
        }
```

Figure III.5

```
rm(list = ls())
#########
library(KernSmooth)
mexemple<-function(z){return(2*z+4*cos(2*z)-sin(2*z))}
KGauss<-function(z){return( (1/sqrt(2*pi))*exp(-z^2/2) )}
#########
a=0
b=6
n=100
sigma=1
hGaussvec=.2*c(10,1,1/4)
colestim=c("dodgerblue2","darkorange","firebrick1")
set.seed(4)
x=runif(n,a,b)
y=mexemple(x)+sigma*rnorm(n)
z=seq(a,b,by=.01)
mhatGauss0=rep(0,length=length(z))
par(mfrow=c(length(hGaussvec),1))
par(oma=c(2,1,2,1))
par(mar=c(3,4,1,1))
for (i in (1:length(hGaussvec)))
        {
        hGauss=hGaussvec[i]
        for (r in (1:length(z)))
                {
                mhatGauss0[r]=sum( KGauss((z[r]-x)/hGauss)*y )/sum( KGauss((z[r]-x)/hGauss) )
                }
        plot(x,y,main="",col=grey(.7),xlab="",ylab="")
        points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        #points(z,mhatGauss0,typ="l",lwd=1.4,main="",col=colestim[1],xlab="",ylab="")
```

```
        lines(locpoly(x,y,degree=0,kernel="normal",bandwidth=hGauss),col=colestim[1],lwd=1.4)
        lines(locpoly(x,y,degree=1,kernel="normal",bandwidth=hGauss),col=colestim[2],lwd=1.4)
        lines(locpoly(x,y,degree=2,kernel="normal",bandwidth=hGauss),col=colestim[3],lwd=1.4)
                if (i==length(hGaussvec)) {mtext("x",side=1,line=3,cex=.85)}
                mtext(expression(paste(hat(m)[n]^{(p)},"(x)",sep="")),side=2,line=2.6,cex=.85)
                if (i==1) {legend(0,15,legend=c("p=0","p=1","p=2"),
                        col=colestim,lwd=1.4,lty=1,cex=.85)}
        }
```

Figure III.6

```
rm(list = ls())
#########
mexemple<-function(z){return(2*z+4*cos(2*z)-sin(2*z))}
#########
a=0
b=6
n=100
sigma=1
kvec=c(2,12,50)
colestim=c("dodgerblue2","darkorange","firebrick1")
set.seed(4)
x=runif(n,a,b)
y=mexemple(x)+sigma*rnorm(n)
z=seq(a,b,by=.01)
mhatkNN=rep(0,length=length(z))
par(mfrow=c(length(kvec),1))
par(oma=c(2,1,2,1))
par(mar=c(3,4,1,1))
for (i in (1:length(kvec)))
        {
        k=kvec[i]
        for (r in (1:length(z)))
                {
            indexneighbourshat=which(rank(abs(z[r]-x))<=k)
                mhatkNN[r]=mean(y[indexneighbourshat])
                }
        plot(x,y,main="",col=grey(.7),xlab="",ylab="")
        points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        points(z,mhatkNN,typ="l",lwd=1.4,main="",col=colestim[1],xlab="",ylab="")
        if (i==length(kvec)) {mtext("x",side=1,line=3,cex=.85)}
        mtext(substitute(paste(hat(m)[list(n,k)],"(x),␣avec␣k=",kk,sep=""),list(kk=kvec[i])),
                side=2,line=2.6,cex=.85)
        }
```

Figure III.7

```
rm(list = ls())
#########
mexemple<-function(z){return(2*z+4*cos(2*z)-sin(2*z))}
#########
a=0
b=6
n=100
sigma=1
```

```
lambdavec=c(.01,.5,1)
colestim=c("dodgerblue2","darkorange","firebrick1")
set.seed(4)
x=runif(n,a,b)
y=mexemple(x)+sigma*rnorm(n)
z=seq(a,b,by=.01)
mhatkNN=rep(0,length=length(z))
par(mfrow=c(length(lambdavec),1))
par(oma=c(2,1,2,1))
par(mar=c(3,4,1,1))
for (i in (1:length(lambdavec)))
        {
        lambda=lambdavec[i]
        mhatspline=smooth.spline(x,y,spar=lambda)
        plot(x,y,main="",col=grey(.7),xlab="",ylab="")
        points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
        lines(mhatspline,lwd=1.4,col=colestim[1],xlab="")
        if (i==length(lambdavec)) {mtext("x",side=1,line=3,cex=.85)}
        mtext(substitute(paste(hat(m)[list(n,lambda)],"(x),␣avec␣",lambda,"=",ll,sep=""),
                list(ll=lambdavec[i])),side=2,line=2.6,cex=.85)
        }
```

Figure III.8

```
rm(list = ls())
#########
mexemple<-function(z){return(2*z+4*cos(2*z)-sin(2*z))}
#########
a=0
b=6
n=100
sigma=1
kvec=c(1,5,20) # Donc m=3,11,41
colestim=c("dodgerblue2","darkorange","firebrick1")
set.seed(4)
x=runif(n,a,b)
y=mexemple(x)+sigma*rnorm(n)
z=seq(a,b,by=.01)
par(mfrow=c(length(kvec),1))
par(oma=c(2,1,2,1))
par(mar=c(3,4,1,1))
for (i in (1:length(kvec)))
   {
   k=kvec[i]
   xgrid=z
   Xcovariate=c()
   for (j in (1:k))
     {
          Xcovariate=cbind(Xcovariate,sqrt(2/(b-a))*cos(2*pi*j*(x-a)/(b-a)))
          Xcovariate=cbind(Xcovariate,sqrt(2/(b-a))*sin(2*pi*j*(x-a)/(b-a)))
          }
   tempdata<-data.frame(y,Xcovariate)

   Xgridcovariate=c()
   Xgridcovariate=cbind(Xgridcovariate,rep(1,length(xgrid)))
   for (j in (1:k))
          {
```

```
        Xgridcovariate=cbind(Xgridcovariate,sqrt(2/(b-a))*cos(2*pi*j*(xgrid-a)/(b-a)))
        Xgridcovariate=cbind(Xgridcovariate,sqrt(2/(b-a))*sin(2*pi*j*(xgrid-a)/(b-a)))
        }
ygridhat=Xgridcovariate%*%(lm(tempdata)$coefficients)
plot(x,y,main="",col=grey(.7),xlab="",ylab="")
points(z,mexemple(z),typ="l",lwd=1.4,main="",col="darkolivegreen4",xlab="",ylab="")
points(z,ygridhat,typ="l",lwd=1.4,main="",col=colestim[1],xlab="",ylab="")
if (i==length(kvec)) {mtext("x",side=1,line=3,cex=.85)}
mtext(substitute(paste("m=",kk,sep="")),list(kk=2*kvec[i]+1)),side=2,line=2.6,cex=.85)
}
```

III.5 TD

1. Soit (X, Y) un vecteur aléatoire bivarié tels que X et Y sont de variance finie. Toute fonction $g : \mathbb{R} \to \mathbb{R}$ peut être utilisée pour construire une prédiction $g(X)$ de Y sur la base de X. Bien entendu, on cherche une fonction g qui rendra l'erreur de prédiction aussi petite que possible, par exemple au sens où $\mathrm{E}[\{Y - g(X)\}^2]$ est minimal quand on considère toutes les fonctions $g \in \mathcal{G} = \{g : \mathbb{R} \to \mathbb{R} : \mathrm{E}[g^2(X)] < \infty\}$ (une condition requise pour que l'espérance ci-dessus soit bien définie). Le but de cet exercice est de montrer qu'on peut déterminer explicitement la fonction g_{opt} qui réalise ce minimum.

   (a) Posons $g_{\mathrm{opt}}(x) = \mathrm{E}[Y|X = x]$ pour tout x. En utilisant l'identité $\mathrm{E}[\mathrm{E}[Z|X]] = \mathrm{E}[Z]$, montrer que $\mathrm{E}[\{Y - g_{\mathrm{opt}}(X)\}\{g_{\mathrm{opt}}(X) - g(X)\}] = 0$.
   (b) En déduire que $\mathrm{E}[\{Y - g(X)\}^2] = \mathrm{E}[\{Y - g_{\mathrm{opt}}(X)\}^2] + \mathrm{E}[\{g_{\mathrm{opt}}(X) - g(X)\}^2]$ (pour ce faire, décomposer $Y - g(X)$ en $Y - g_{\mathrm{opt}}(X) + g_{\mathrm{opt}}(X) - g(X)$).
   (c) Conclure que la fonction $g \in \mathcal{G}$ qui minimise $\mathrm{E}[\{Y - g(X)\}^2]$ est effectivement la fonction $g = g_{\mathrm{opt}}$. Ceci justifie le modèle de régression non paramétrique adopté à la Section III.1.

2. Considérons un vecteur aléatoire (X, Y) dont la loi est décrite de la façon suivante. Soit Z une variable qui prend la valeur 1 avec probabilité $p = .7$ et 0 avec la probabilité $1-p$. Conditionnellement à l'événement $[Z = 1]$, le vecteur (X, Y) suit une loi normale standard bivariée. Conditionnellement à l'événement $[Z = 0]$, les variables X et Y sont indépendantes, X est de loi normale de moyenne 3 et de variance $1/4$, et Y est de loi normale standard. On vérifie facilement que la densité de (X, Y) est

$$(x, y) \mapsto f^{(X,Y)}(x, y) = p\phi_{0,1}(x)\phi_{0,1}(y) + (1 - p)\phi_{3,1/4}(x)\phi_{0,1}(y),$$

où ϕ_{μ,σ^2} est la densité de la loi normale de moyenne μ et de variance σ^2. Engendrer un échantillon aléatoire de taille $n = 1000$ depuis cette loi bivariée et faire les graphes des estimateurs à noyau fondés sur un noyau gaussien et les tailles de fenêtre $h_n = .5$, $h_n = 1.5$ et $h_n = 4$. Les comparer à la vraie densité et apprécier la dépendance en h_n du biais et de la variance des estimateurs à noyau.

3. Montrer que lorsqu'on travaille avec un noyau partout non négatif, l'estimateur par polynômes locaux obtenu pour $p = 0$ (régression localement constante) coïncide avec l'estimateur de Nadaraya–Watson.

4. Ecrire un code R qui engendre un échantillon aléatoire (X_i, Y_i), $i = 1, \ldots, n = 100$, suivant le modèle de régression $Y_i = (3X_i + 1) + \varepsilon_i$, où X_i suit une loi Beta de paramètre $\alpha = 2$ et $\beta = 5$ et où ε_i (qui est indépendante de X_i) suit une loi normale de moyenne zéro et de variance $1/100$. Pour le noyau gaussien et les tailles de fenêtre $h_n = 1$, $h_n = .1$ et $h_n = .025$, faire le graphe des estimateurs par régression polynomiale locale d'ordre $p = 0$ (Nadaraya–Watson) et $p = 1$. Lire les résultats en termes de biais à la lumière des Théorèmes III.2 et III.3.

5. Cet exercice est relatif au jeu de données `prestige`, qui est disponible dans le package `car` de R. On considèrera la régression de la variable `income` par rapport à la variable `prestige`. Tracer, pour des valeurs des paramètres de lissage que vous jugerez visuellement adéquates dans chaque cas, les estimateurs par régression polynomiale associé à $p = 0$ et $p = 1$ (et un noyau gaussien), un estimateur par plus proches voisins, un estimateur par spline et un estimateur par projection utilisant la base trigonométrique.

Annexe A

Quelques résultats d'analyse

Théorème A.1 *Soient $f : [a, b] \to \mathbb{R}$ une fonction continue et $g : [a, b] \to \mathbb{R}$ une fonction telle que $g(x) \geq 0$ pour tout $x \in [a, b]$. Alors,*

$$\int_a^b f(x)g(x)\,dx = f(\xi)\int_a^b g(x)\,dx \tag{A.1}$$

pour un certain $\xi \in [a, b]$. En particulier, pour $g \equiv 1$, on obtient que

$$\int_a^b f(x)\,dx = f(\xi)(b-a)$$

pour un certain $\xi \in [a, b]$.

Preuve du Théorème A.1. Puisque f est continue sur $[a, b]$, il existe $c \leq d$ tels que $f([a, b]) = [c, d]$. Comme $g(x) \geq 0$ pour tout $x \in [a, b]$, on a donc

$$c\int_a^b g(x)\,dx \leq \int_a^b f(x)g(x)\,dx \leq d\int_a^b g(x)\,dx. \tag{A.2}$$

Posons $s = \int_a^b g(x)\,dx$. Si $s = 0$, alors (A.2) implique que

$$\int_a^b f(x)g(x)\,dx = 0,$$

et (A.1) tient pour tout $\xi \in [a, b]$. On peut donc supposer que $s \neq 0$. Dans ce cas, on peut réécrire (A.2) sous la forme

$$c \leq \frac{1}{s}\int_a^b f(x)g(x)\,dx \leq d.$$

Par le théorème de la valeur intermédiaire, la fonction continue f atteint (sur $[a,b]$) toute valeur entre c et d, donc en particulier la valeur

$$\frac{1}{s}\int_a^b f(x)g(x)\,dx.$$

Ceci prouve l'existence d'un $\xi \in [a,b]$ tel

$$f(\xi) = \frac{1}{s}\int_a^b f(x)g(x)\,dx,$$

ce qui établit le résultat. □

Théorème A.2 *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction de classe C^2. Alors, pour tout $a, x \in \mathbb{R}$, on a*

$$(i) \qquad f(x) = f(a) + (x-a)f'(a) + \int_a^x (x-t)f''(t)\,dt,$$

de sorte que

$$(ii) \qquad f(x+h) = f(x) + hf'(x) + h^2\int_0^1 (1-s)f''(x+sh)\,ds$$

pour tout $x, h \in \mathbb{R}$.

Preuve du Théorème A.2. En appliquant le théorème fondamental du calcul différentiel et intégral, puis la formule d'intégration par parties, on obtient

$$f(x) - f(a) = \int_a^x f'(t)\,dt = \Big[tf'(t)\Big]_a^x - \int_a^x tf''(t)\,dt = xf'(x) - af'(a) - \int_a^x tf''(t)\,dt.$$

En utilisant de nouveau le théorème fondamental, on a alors

$$\begin{aligned} f(x) - f(a) &= (x-a)f'(a) + x(f'(x) - f'(a)) - \int_a^x tf''(t)\,dt \\ &= (x-a)f'(a) + x\int_a^x f''(t)\,dt - \int_a^x tf''(t)\,dt \\ &= (x-a)f'(a) + \int_a^x (x-t)f''(t)\,dt, \end{aligned}$$

ce qui établit (i). Le résultat (ii) s'obtient alors en replaçant x et a par $x+h$ et x, respectivement, et en faisant le changement de variables $t = x + sh$. □

Pour énoncer le résultat suivant, nous avons besoin d'introduire quelques notations. La quantité $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ est appelée *multi-indice* lorsqu'elle est utilisée pour

définir des dérivées partielles multiples suivant la convention

$$\partial^\alpha f(x) = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_1}}(x),$$

où $x = (x_1, \dots, x_d) \mapsto f(x)$ est une fonction définie sur (un sous-ensemble de) $\mathbb{R}^d$; ci-dessus, nous avons posé $|\alpha| = \alpha_1 + \ldots + \alpha_d$. On notera aussi $\alpha! = (\alpha_1!) \dots (\alpha_d!)$ et $h^\alpha = h_1^{\alpha_1} \dots h_d^{\alpha_d}$ pour $h \in \mathbb{R}^d$.

Théorème A.3 *Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^{k+1}. Alors, pour tout $a, h \in \mathbb{R}^d$, on a*

$$f(a+h) = \sum_{|\alpha| \le k} \frac{1}{\alpha!} \partial^\alpha f(a) h^\alpha + R_{a,h},$$

où

$$R_{a,h} = \sum_{|\alpha| = k+1} \frac{1}{\alpha!} \partial^\alpha f(a+ch) h^\alpha,$$

pour un certain $c \in (0,1)$. En particulier, pour $d = 1$, on a

$$f(a+h) = \sum_{\ell=1}^{k} \frac{1}{\ell!} f^{\underline{\ell}}(a) h^\ell + \frac{1}{(k+1)!} f^{\underline{k+1}}(a+ch) h^{k+1},$$

pour un certain $c \in (0,1)$, où $f^{\underline{\ell}}$ désigne la ℓème dérivée de f.

Bibliographie

[1] Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall. New York.

[2] Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag, Berlin Heidelberg.

[3] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics* **12**, 1285–1297.

[4] Tsybakov, A.B. (2004). *Introduction to Nonparametric Estimation*. Springer, New York.

[5] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

[6] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer-Verlag, New York.