Cours modélisation de données complexes

Professeur Gilles Durrieu

Master 2 Mention Mathématiques Appliquées, Statistique
Parcours Data Science et Modélisation Statistique

Université Bretagne Sud

Table des matières

Introduction générale **7**

1 Estimation non paramétrique basée sur des estimateurs de type noyau **9**

1.1 Introduction 10
1.2 Modèle et estimateurs 11
1.3 Estimation non paramétrique de la densité de probabilité 11
1.3.1 Calcul du biais de $\hat{f}_{h_n}$ 13
1.3.2 Calcul de la variance de $\hat{f}_{h_n}$ 14
1.3.3 Propriétés asymptotiques de $\hat{f}_{h_n}$ 15
1.3.4 Choix de la taille de la fenêtre h_n 15
1.3.5 Par minimisation de l'Erreur Quadratique Moyenne (EQM) . 15
1.3.6 Par minimisation de l'Erreur Quadratique Moyenne Intégrée (EQMI) 16
1.3.7 Par la méthode de la validation croisée 17
1.4 Estimation non paramétrique de la fonction de régression 19
1.4.1 Propriétés asymptotiques 20
1.5 Régression polynomiale locale 22
1.6 Choix de la taille de la fenêtre h_n 24
1.7 Régression spline 24
1.8 Comparaison estimation Nadaraya-Watson et régression polynomiale locale 26

2 Applications en environnement **27**

2.1 Comparaison d'estimateurs de régression non paramétriques : application en valvométrie 31
2.1.1 Introduction 32
2.1.2 Modèle et estimateurs 33
2.1.3 Propriétés asymptotiques 34
2.1.4 Choix de la fenêtre 35
2.1.5 Application en valvométrie 36
2.2 Quantiles de régression : application à l'estimation de la croissance . . 38
2.2.1 Introduction 38
2.2.2 Modèle et estimateurs 39
2.2.3 Application 42

2.2.4 Article publié : Schwartzmann, C., Durrieu, G., Sow, M., Ciret, P., Lazareth, C. E. and Massabuau, J. C. (2011) In situ giant clam growth rate behavior in relation to temperature : A one-year coupled study of high-frequency noninvasive valvometry and sclerochronology, Limnology and oceanography, 56(5), 1940-1951. 45
2.3 Estimation par la théorie des shot noise 58
2.3.1 Introduction 58
2.3.2 Données 59
2.3.3 Méthodes 60
2.3.4 Densités de probabilité et extrêmes 60
2.3.5 Densité spectrale 61
2.3.6 Shot noise 62
2.3.7 Article publié : Schmitt, F. G., De Rosa, M., Durrieu, G., Sow, M., Ciret, P., Tran, D. and Massabuau, J. C. (2011). Statistical study of bivalve high frequency microclosing behavior : Scaling properties and shot noise analysis. International Journal of Bifurcation and Chaos, 21(12), 3565-3576. 64
2.4 Processus de Markov caché et surveillance de la qualité des eaux . . . 77
2.4.1 Article publié : Azais, R., Coudret, R. and Durrieu, G. (2014) A hidden renewal model for monitoring aquatic systems biosensors, Environmetrics, 25(3), 189-199. 77
2.5 Choix de la taille de la fenêtre quand le nombre de modes d'une distribution est connu 89
2.5.1 Article publié : Coudret, R., Durrieu, G. and Saracco, J. (2015) Comparison of kernel density estimators with assumption on number of modes, Communications in Statistics-Simulation and Computation, 44(1), 196-216. 89
2.6 Estimation de la dérivée de la fonction de régression en design aléatoire 113
2.6.1 Article publié : Bercu B., Capderou S. and Durrieu G (2019) A nonparametric statistical procedure for the detection of marine pollution, Journal of Applied Statistics, 46(1), 119-140. . 113
2.7 Estimation de la dérivée de la fonction de régression en design déterministe 137
2.7.1 Article publié : Bercu, B., Capderou S. and Durrieu G. (2019) Nonparametric recursive estimation of the derivative of the regression function with application to sea shores water quality, Statistical Inference for Stochastic Processes, 22, 17–40. 137
2.8 Théorie des valeurs extrêmes et environnement 162
2.8.1 Article publié : Durrieu, G., Grama, I., Pham, Q. K. and Tricot, J. M. (2015). Nonparametric adaptive estimation of conditional probabilities of rare events and extreme quantiles, Extremes, 18(3), 437-478. 162

2.8.2 Article publié : Durrieu, G., Pham, Q. K., Foltête, A. S., Maxime, V., Grama, I., Le Tilly, V., ... and Sire, O. (2016). Dynamic extreme values modeling and monitoring by means of sea shores water quality biomarkers and valvometry, Environmental monitoring and assessment, 188(7), 401. 206
2.8.3 Article publié : Durrieu, G., Grama I., Jaunatre K., Pham, Q. K. and Tricot J.M. (2016). Dynamic extreme values modeling and monitoring by means of sea shores water quality biomarkers and valvometry, Environmental monitoring and assessment, 188(7), 401. 215

Table des figures **237**

Bibliographie **239**

Introduction générale

Nous disposons aujourd'hui d'outils d'analyse de plus en plus performants, capables de générer des quantités de données de plus en plus grandes. Ces techniques ouvrent des champs d'analyse complètement nouveaux, souvent largement vierges. La protection de l'environnement, la santé, la découverte de gènes responsables de maladies, la finance, l'actuariat et l'écologie sont parmi les plus grands des enjeux scientifiques, thérapeutiques et économiques de ces dernières années.

Les récentes innovations réalisées sur les méthodes d'acquisitions et les appareils de mesures ainsi que l'utilisation intensive de moyens informatiques permettent souvent de récolter des volumes de données qui étaient inimaginables. De nouvelles questions, demandant de nouvelles solutions mathématiques, sont donc apparues. Par exemple, mais de façon très générale, les données récoltées présentent beaucoup de sources de problèmes (distributions non gaussiennes, gros volumes de données, valeurs manquantes, valeurs extrêmes, observations dépendantes, valeurs aberrantes, non linéarité, etc). Ainsi dès lors que l'on se confronte aux données réelles générées par ces technologies dans de nombreux domaines, la modélisation de systèmes complexes est indispensable et est devenue une méthodologie incontournable.

Il est merveilleux de constater que la théorie des probabilités et des statistiques fournit des modèles mathématiques efficaces et la statistique un cadre rigoureux d'analyse et de modélisation des résultats expérimentaux. Associée à la simulation numérique, elle montre une redoutable efficacité. L'évolution des moyens de calcul de ces dernières années a permis non seulement d'augmenter considérablement le volume des simulations mais aussi la résolution numérique de modèles complexes pouvant prendre en compte un grand nombre de variables.

Le cours se divise en deux chapitres. Le chapitre 1 présente la mise en place différentes approches de modélisation en régréssion non paramétrique basée sur des estimateurs de type noyau (Nadaraya-Watson, régression polynomiale locale, régression spline, régression quantile). Dans le chapitre 2, nous donnons des applications et des développements dans le domaine de l'écologie.

Chapitre 1

Estimation non paramétrique basée sur des estimateurs de type noyau

1.1 Introduction

Dans cette partie, nous développons des modèles statistiques pour l'étude de systèmes complexes. Nous appliquerons la méthodologie présentée afin de caractériser l'activité valvaire de bivalves (qui résume à elle seule l'ensemble du comportement locomoteur de ce type d'animal) en fonction des paramètres du milieu afin de décrire des comportements normaux en fonction des variations naturelles du milieu (éthologie marine) puis de discriminer ensuite des comportements atypiques liés à une perturbation du milieu.

Dans beaucoup d'applications, il est courant d'obtenir de grands ensembles de données.

Les modèles de régression servent à représenter mathématiquement la relation entre une variable aléatoire Y et un ensemble de covariables X pouvant être aléatoires ou/et déterministes. On les utilise souvent pour prédire Y en fonction des valeurs de X. Le modèle de régression le plus élémentaire est le modèle de régression linéaire simple contenant une seule covariable X. Pour les données $(X_1, Y_1) \ldots, (X_n, Y_n)$, ce modèle énonce que pour $i = 1, \ldots, n$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1.1}$$

où l'on suppose généralement que les variables aléatoires ε_i sont indépendantes et identiquement distribuées (i.i.d) de moyenne nulle. À partir des estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$, on obtient l'équation de prédiction de Y en $X = x$:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Lorsque la relation entre Y et X est non linéaire, le modèle (1.1) est insatisfaisant. On s'intéresse alors pour $i = 1, \ldots, n$ au modèle plus général :

$$Y_i = m(X_i) + \varepsilon_i, \tag{1.2}$$

où m est la fonction de régression/la fonction de lien qui est inconnue. Lorsque $E(\varepsilon/X = x) = 0$, nous avons $m(x) = E(Y/X = x)$. Si on ne suppose pas une forme paramétrique pour m, on parlera d'une fonction de régression non paramétrique.

La théorie de l'estimation non paramétrique s'est développée considérablement ces deux dernières decennies. Un des problèmes souvent rencontrés en statistique est celui de l'estimation fonctionnelle associée à la loi des observations telles que par exemple la fonction de densité ou la fonction de régression.
Dans ce chapitre, nous nous intéressons à la mise en place d'un modèle de régréssion non paramétrique basée sur des estimateurs de type noyau pour l'analyse du comportement de bivalves. Nous présentons d'abord l'estimateur à noyau classique d'une densité de probabilité (Rosenblatt, 1956 ; Parzen, 1962 ; Silverman, 1986), l'estimateur de la fonction de régression (Nadaraya, 1964 et Watson, 1964 ; Härdle, 1990), puis nous rappelons les propriétés asymptotiques de ces estimateurs. Ensuite nous établissons les rythmes biologiques liés aux marées chez l'huître *Crassostrea gigas*, puis nous proposons une approche statistique pour prouver cette corrélation.

1.2 Modèle et estimateurs

Un des modèles le plus fréquemment rencontré en statistique paramétrique ou non paramétrique est le modèle de régression dont nous donnons ci-dessous une description sommaire.

Nous disposons d'un échantillon composé de n couples indépendants de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ et nous considérons le modèle de régression non paramétrique donné, pour $i = 1, \ldots, n$, par

$$Y_i = m(X_i) + \varepsilon_i. \tag{1.3}$$

Dans ce modèle intervient une fonction m inconnue à estimer qui exprime la valeur moyenne de Y en fonction de X et un terme aléatoire d'erreur ε de loi inconnue et indépendant de X. Dans le cas classique, l'erreur commise est modélisé par une variable aléatoire gaussienne et la relation entre la variable dépendante Y et la ou les variables explicatives X est linéaire.

Il arrive fréquemment que ces postulats ne soient pas respectés, souvent de façon plus évidente lorsque que l'on possède un nombre important de données. Dans ce cas, le chercheur désire habituellement obtenir un modèle plus complexe, qui reflète mieux la relation entre Y et X. Il y a différentes façons d'y arriver, nous considérons ici des estimateurs non paramétriques de type noyau dans le cas du dispositif expérimental à effets fixes (la distance entre les électrodes est mesurée de manière équidistante).
Le principal avantage de la régression non paramétrique est qu'elle ne suppose aucune forme spécifique pour l'estimateur, ce qui lui donne beaucoup plus de flexibilité.

1.3 Estimation non paramétrique de la densité de probabilité

L'estimateur de type histogramme est classiquement utilisé en estimation de loi mais il reste très sensible au choix du nombre de classes. Nous avons donc choisi d'utiliser l'estimateur à noyau de la fonction de densité de probabilité (Rosenblatt, 1956 ; Parzen, 1962 ; Silverman, 1986 ; Wand et Jones 1995). Plus précisement, si on considère un échantillon de n variables aléatoires indépendantes et de même loi (f inconnue), $X_1, X_2, \ldots, X_n$, l'estimateur de type noyau de f s'ecrit :

$$\hat{f}_n(x) = \frac{1}{n\,h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{1.4}$$

ou dans sa forme récursive :

$$\tilde{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right). \tag{1.5}$$

La fenêtre h_n désigne une suite de nombres réels strictement positifs vérifiant (*C1*) $h_n \to 0$ et $n\,h_n \to \infty$ lorsque $n \to \infty$. Le noyau est une fonction mesurable, positive

et bornée satisfaisant (*C2*) $\int_{\mathbb{R}} K(x)\,dx = 1$, $\int x\,K(x)\,dx = 0$, $\int_{\mathbb{R}} |x|\,K(x)\,dx < +\infty$ et $\int_{\mathbb{R}} K^2(x)\,dx = \tau^2$. La condition (*C3*) précise que f est de classe $\mathcal{C}^2$.

Les avantages de de cet estimateur sont :
- une écriture mathématique simple,
- une bonne prise en compte de la contribution des observations situées dans le voisinage d'un point x.

Nous donnons maintenant différents noyaux vérifiant (*C1*) :
- **Noyau uniforme** :
$$K(x) = \frac{1}{2}\,\mathbf{1}(|x| \leq 1).$$
- **Noyau triangulaire** :
$$K(x) = (1 - |x|)\,\mathbf{1}(|x| \leq 1).$$
- **Noyau Epanechnikov** :
$$K(x) = \frac{3}{4}(1 - x^2)\,\mathbf{1}(|x| \leq 1).$$
- **Noyau quadratique** :
$$K(x) = \frac{15}{16}(1 - x^2)^2\,\mathbf{1}(|x| \leq 1).$$
- **Noyau Gaussien** :
$$K(x) = \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{1}{2}x^2\right).$$
- **Noyau cosinus** :
$$K(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}x\right)\,\mathbf{1}(|x| \leq 1)$$
- **Noyau cubique** :
$$K(x) = \frac{35}{32}(1 - x^2)^3\,\mathbf{1}(|x| \leq 1)$$

Exercice. *Vérifier pour chaque noyau en utilisant le logiciel R les conditions (C1).*
Remarques :
- L'estimateur à noyau est une fonction de densité de probabilité.
- L'estimateur à noyau hérite des propriétés de la fonction noyau.
- Si h_n et K sont connus, l'estimateur appliqué à un jeu de données est unique.

Nous remarquons aussi que la taille de fenêtre (le paramètre de lissage) h_n contrôle le degré de lissage :
- h_n petit : estimateur oscillant.
- h_n grand : estimateur lisse.
- $h_n \to \infty$: estimateur prenant progressivement la forme de la fonction noyau.

1.3.1 Calcul du biais de $\hat{f}_{h_n}$

Nous considérons une suite de variables aléatoires $X_1, \ldots, X_n$ i.i.d dont la densité de probabilité f est inconnue.

On a en utilisant la linéarité de l'espérance mathématique :

$$E(\hat{f}_{h_n}(x)) = E\left(\frac{1}{n\,h_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\right) = \frac{1}{n\,h_n}\sum_{i=1}^{n} E\left(K\left(\frac{x-X_i}{h_n}\right)\right).$$

La fonction noyau K est symétrique et les variables aléatoires X_i pour $i = 1, \ldots, n$ sont de même loi. On en déduit :

$$E(\hat{f}_{h_n}(x)) = \frac{1}{n\,h_n} \times n \times E\left(K\left(\frac{x-X}{h_n}\right)\right) = \frac{1}{h_n} E\left(K\left(\frac{X-x}{h_n}\right)\right),$$

et donc par définition de l'espérance mathématique

$$E(\hat{f}_{h_n}(x)) = \frac{1}{h_n}\int_{-\infty}^{\infty} K\left(\frac{u-x}{h_n}\right)\, f(u)\, du.$$

En effectuant le changement de variable

$$s = \frac{u-x}{h_n}$$

on obtient :

$$E(\hat{f}_{h_n}(x)) = \int_{-\infty}^{\infty} K(s)\, f(x+sh_n)\, ds.$$

Un développement limité de f à l'ordre 2 au voisinage du point x donne :

$$f(x+sh) = f(x) + shf'(x) + \frac{s^2h^2}{2} f''(x) + o(h^2). \tag{1.6}$$

En utilisant (1.6), on obtient :

$$E(\hat{f}_{h_n}(x)) = \int_{-\infty}^{\infty} K(s)\, \left(f(x) + sh_n f'(x) + \frac{s^2h_n^2}{2} f''(x)\right) + o(h_n^2),$$

et donc

$$E(\hat{f}_{h_n}(x)) = f(x)\int_{-\infty}^{\infty} K(s)ds + h_n f'(x)\int_{-\infty}^{\infty} sK(s)\, ds + \frac{h_n^2 f''(x)}{2}\int_{-\infty}^{\infty} s^2K(s)\, ds + o(h_n^2).$$

On en déduit en utilisant (*C1*) :

$$E(\hat{f}_{h_n}(x)) = f(x) + \frac{h_n^2 f''(x)}{2}\int_{-\infty}^{\infty} s^2K(s)\, ds + o(h_n^2),$$

le biais de l'estimateur $\hat{f}_{h_n}$ de f est :

$$\text{biais}(\hat{f}_{h_n}(x)) = E(\hat{f}_{h_n}(x)) - f(x) = \frac{h_n^2 f''(x)}{2}\int_{-\infty}^{\infty} s^2K(s)\, ds + o(h_n^2)$$

que l'on peut écrire :

$$\text{biais}(\hat{f}_{h_n}(x)) = \frac{h_n^2 f''(x)}{2}\mu_2(K) + o(h_n^2), \tag{1.7}$$

avec la notation

$$\mu_2(K) = \int_{-\infty}^{\infty} s^2 K(s)\, ds.$$

Par conséquent :

— le biais de $\hat{f}_{h_n}(x)$ est quadratique en h_n.
— si $h_n \to 0$ alors $\text{biais}(\hat{f}_{h_n}(x)) \to 0$ quand $n \to \infty$.
— dépendance biais et de la courbure.

1.3.2 Calcul de la variance de $\hat{f}_{h_n}$

En utilisant les propriétés de la variance et puisque les variables aléatoires X_i pour $i = 1, \ldots, n$ sont indépendantes, on a :

$$\text{Var}(\hat{f}_{h_n}(x)) = \text{Var}\left(\frac{1}{n\,h_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right) = \frac{1}{n^2\,h_n^2}\sum_{i=1}^{n}\text{Var}\left(K\left(\frac{x - X_i}{h_n}\right)\right).$$

Puisque les variables aléatoires X_i pour $i = 1, \ldots, n$ sont de même loi, on a aussi :

$$\text{Var}(\hat{f}_{h_n}(x)) = \frac{1}{n\,h_n^2}\left(E\left(K^2\left(\frac{x - X}{h_n}\right)\right) - \left(E\left(K\left(\frac{x - X}{h_n}\right)\right)\right)^2\right). \tag{1.8}$$

Par définition de l'espérance mathématique, nous avons par symétrie de la fonction K :

$$\frac{1}{h_n^2}\,E\left(K^2\left(\frac{x - X}{h_n}\right)\right) = \frac{1}{h_n^2}\int_{-\infty}^{\infty} K^2\left(\frac{x - u}{h_n}\right) f(u)\, du = \frac{1}{h_n^2}\int_{-\infty}^{\infty} K^2\left(\frac{u - x}{h_n}\right) f(u)\, du.$$

En effectuant le changement de variable

$$s = \frac{u - x}{h_n},$$

on obtient :

$$\frac{1}{h_n^2}\,E\left(K^2\left(\frac{x - X}{h_n}\right)\right) = \frac{1}{h_n}\int_{-\infty}^{\infty} K^2(s)\, f(x + sh_n)\, ds,$$

et

$$\frac{1}{h_n^2}\left(E\left(K\left(\frac{x - X}{h_n}\right)\right)\right)^2 = \left(\int_{-\infty}^{\infty} K(s)\, f(x + sh_n)\, ds\right)^2.$$

On en déduit en remplaçant les deux précédentes expressions dans (1.8) :

$$\text{Var}(\hat{f}_{h_n}(x)) = \frac{1}{n}\left(\frac{1}{h_n}\int_{-\infty}^{\infty} K^2(s)\, f(x + sh)\, ds - \left(\int_{-\infty}^{\infty} K(s)\, f(x + sh)\, ds\right)^2\right).$$

Par un développement limité de f à l'ordre 1 au voisinage de x, on obtient :

$$\mathrm{Var}(\hat{f}_{h_n}(x)) = \frac{1}{nh_n} f(x) \int_{-\infty}^{\infty} K^2(s)\, ds + o\left(\frac{1}{nh_n}\right). \tag{1.9}$$

Nous remarquons alors que la variance $\mathrm{Var}(\hat{f}_{h_n}(x))$ diminue quand h_n augmente tandis que le biais $\mathrm{biais}(\hat{f}_{h_n}(x))$ diminue quand h_n diminue. Comme nous souhaitons réduire simultanément le biais et la variance (compromis biais-variance), nous proposerons dans la suite du cours différentes méthodes pour choisir la taille de fenêtre h_n.

1.3.3 Propriétés asymptotiques de $\hat{f}_{h_n}$

Théorème 1. Sous les conditions (*C1*), (*C2*) et (*C3*), nous avons quand $n \to \infty$ pour toutes valeurs de x

$$\sqrt{nh_n}\left(\hat{f}_{h_n}(x) - f(x)\right) \xrightarrow{L} N\left(\mathrm{biais}(\hat{f}_{h_n}(x)),\, f(x) \int_{-\infty}^{\infty} K^2(s)\, ds\right), \tag{1.10}$$

et

$$\hat{f}_{h_n}(x) \xrightarrow{p.s.} f(x),$$

où

$$\mathrm{biais}(\hat{f}_{h_n}(x)) = \frac{h_n^2 f''(x)}{2} \mu_2(K) + o(h_n^2).$$

Quand h_n tend vers 0, le biais de $\hat{f}_{h_n}$ tend aussi vers 0. La preuve du Théorème (1) n'est pas donnée dans ce cours.

Nous déduisons du Théorème (1) sous les conditions (*C1*), (*C2*) et (*C3*) un intervalle de confiance asymptotique de la densité de probabilité $f(x)$ de niveau de confiance $(1-\alpha)\%$ en un point x fixé donné par :

$$\mathrm{IC}_f = \left[\hat{f}_{h_n}(x) - z_{1-\alpha/2}\sqrt{\frac{\hat{f}_{h_n}(x) \int_{-\infty}^{\infty} K^2(s)\, ds}{nh_n}},\, \hat{f}_{h_n}(x) + z_{1-\alpha/2}\sqrt{\frac{\hat{f}_{h_n}(x) \int_{-\infty}^{\infty} K^2(s)\, ds}{nh_n}}\right]$$

en remplaçant la densité $f(x)$ par son estimateur $\hat{f}_{h_n}(x)$ dans la variance où $z_{1-\alpha/2}$ désigne la valeur critique/valeur théorique d'une loi $N(0,1)$.

1.3.4 Choix de la taille de la fenêtre h_n

1.3.5 Par minimisation de l'Erreur Quadratique Moyenne (EQM)

L'erreur quadratique moyenne donne une expression combinant la variance et le biais de l'estimateur $\hat{f}_{h_n}$ de f. Nous définissons au sens de l'erreur quadratique minimum la taille de fenêtre h_h tel que :

$$h_{EQM} = \arg\min_{h_n} \mathrm{EQM}(\hat{f}_{h_n}(x)).$$

Par définition, l'Erreur Quadratique Moyenne (Mean Squared Error en anglais) est donnée par :

$$\text{EQM}\left(\hat{f}_{h_n}(x)\right) = \frac{1}{nh_n} f(x) \int_{-\infty}^{\infty} K^2(s)\, ds + \frac{h_n^4}{4} \left(f''(x)\mu_2(K)\right)^2, \qquad h_n \to 0 \text{ et } nh_n \to \infty. \tag{1.11}$$

On observe que lorsque $h_n \to 0$ et $nh_n \to \infty$

$$\text{EQM}\left(\hat{f}_{h_n}(x)\right) \to 0,$$

En notant :

$$A_1 = f(x) \int_{-\infty}^{\infty} K^2(s)\, ds \qquad \text{et} \qquad A_2 = \left(f''(x)\right)^2 \left(\mu_2(K)\right)^2,$$

on a :

$$\text{EQM}\left(\hat{f}_{h_n}(x)\right) = \frac{1}{nh_n} A_1 + \frac{h_n^4}{4} A_2.$$

Par dérivation de la fonction $\text{EQM}\left(\hat{f}_{h_n}(x)\right)$ par rapport à h_n, on a

$$\frac{\partial \text{EQM}\left(\hat{f}_{h_n}(x)\right)}{\partial h_n} = -\frac{1}{nh_n^2} A_1 + \frac{4h_n^3}{4} A_2,$$

et en résolvant l'équation

$$\frac{\partial \text{EQM}\left(\hat{f}_{h_n}(x)\right)}{\partial h_n} = 0,$$

on obtient :

$$-\frac{1}{nh_n^2} A_1 + \frac{4h_n^3}{4} A_2 = 0$$

et donc

$$h_{EQM} = \left(\frac{A_1}{nA_2}\right)^{1/5} = \left(\frac{f(x) \int_{-\infty}^{\infty} K^2(s)\, ds}{n \left(f''(x)\right)^2 \left(\mu_2(K)\right)^2}\right)^{1/5}. \tag{1.12}$$

Le problème de choix du paramètre h_n par cette méthode est que h_{EQM} dépend de f et de f'' qui sont inconnues.

1.3.6 Par minimisation de l'Erreur Quadratique Moyenne Intégrée (EQMI)

Par définition, l'Erreur Quadratique Moyenne Intégrée (EQMI) s'écrit :

$$\begin{aligned} \text{EQMI}\left(\hat{f}_{h_n}(x)\right) &= \int \text{EQM}\left(\hat{f}_{h_n}(x)\right) dx \\ &= \tfrac{1}{nh_n} \int K^2(s)\, ds \int f(x)\, dx + \frac{h_n^4}{4} \mu_2(K)^2 \int \left(f''(x)\right)^2 dx. \end{aligned}$$

On en déduit par dérivation au sens de $\text{EQMI}\left(\hat{f}_{h_n}(x)\right)$ minimum :

$$h_{EQMI} = \left(\frac{\int K^2(s)\, ds}{n\, \mu_2^2(K) \int \left(f''(x)\right)^2 dx}\right)^{1/5}, \tag{1.13}$$

qui dépend aussi de $f''(x)$.

Si on considère pour densité de probabilité f la densité de probabilité d'une loi $N(\mu, \sigma^2)$, on obtient :

$$\int \left(f''(x)\right)^2 \, dx = \frac{3}{8\sqrt{\pi}\,\sigma^5},$$

et donc

$$h_{EQMI} = \left(\frac{8\sqrt{\pi} \int K^2(x)\,dx}{3\,n\,\mu_2^2(K)} \right)^{1/5} \times S$$

où S est l'estimateur empirique usuel de σ donné par :

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}.$$

De plus :

$$\left(\frac{8\sqrt{\pi} \int K^2(x)\,dx}{3\,\mu_2^2(K)} \right)^{1/5} = \begin{cases} 1.06 & \text{pour un noyau Gaussien,} \\ 2.34 & \text{pour un noyau Epanechnikov,} \\ 2.78 & \text{pour un noyau quadratique.} \end{cases}$$

Théorème 2. Soit f une fonction continue définie sur $[a, b]$ et à valeurs dans $\mathbb{R}$. Alors

$$\lim_{n\to+\infty} \frac{b-a}{n} \sum_{k=1}^{n} f\left(a + \frac{k(b-a)}{n}\right) = \int_a^b f(t)\,dt.$$

On en déduit que

$$\mathrm{EQMI}\left(\hat{f}_{h_n}(x)\right) = \int_{-5}^{5} \mathrm{EQM}(\hat{f}_{h_n}(x))\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} \left(\frac{5-(-5)}{n}\right) \mathrm{EQM}(\hat{f}_{h_n}(x_i)$$

où $x_i = -5 + 0.02 \times i$.

1.3.7 Par la méthode de la validation croisée

Le critère de la validation croisée consiste à minimiser par rapport à h_n la fonction

$$\mathrm{CV}(h_n) = \mathrm{E}\left(\int_{-\infty}^{\infty} \left(\hat{f}_{h_n}(x) - f(x)\right)^2 \, dx \right).$$

Du fait de la linéarité de l'espérance mathématique, on obtient :

$$\mathrm{CV}(h_n) = \mathrm{E}\left(\int_{-\infty}^{\infty} \hat{f}^2_{h_n}(x)\,dx \right) - 2\,\mathrm{E}\left(\int_{-\infty}^{\infty} \hat{f}_{h_n}(x) f(x)\,dx \right) + \int_{-\infty}^{\infty} f^2(x)\,dx.$$

Comme $\int_{-\infty}^{\infty} f^2(x)\,dx$ ne dépend pas de h_n, cela revient à minimiser :

$$\mathrm{CV}(h_n) - \int_{-\infty}^{\infty} f^2(x)\,dx = \mathrm{E}\left(\int_{-\infty}^{\infty} \hat{f}^2_{h_n}(x)\,dx \right) - 2\,\mathrm{E}\left(\int_{-\infty}^{\infty} \hat{f}_{h_n}(x) f(x)\,dx \right).$$

On a par définition de l'espérance mathématique

$$E\left(\hat{f}_{h_n}(x)\right) = \int_{-\infty}^{\infty} \hat{f}_{h_n}(x) f(x)\, dx.$$

Un estimateur de $\int_{-\infty}^{\infty} \hat{f}_{h_n}(x) f(x)\, dx$ est alors donné par

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$$

où

$$\hat{f}_{(-i)}(x) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} K\left(\frac{x - X_j}{h_n}\right).$$

La fonction de validation croisée est donc :

$$\mathrm{CV}(h_n) = \int_{-\infty}^{\infty} \hat{f}_{h_n}^2(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i). \tag{1.14}$$

Ainsi, le choix de h_n par la méthode de la validation croisée consiste à minimiser en h_n la fonction

$$\mathrm{CV}(h_n) = \int_{-\infty}^{\infty} \hat{f}_{h_n}^2(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i).$$

Voici dans la Figure (1.1) une représentation classique de la fonction de validation croisée CV en fonction de h :

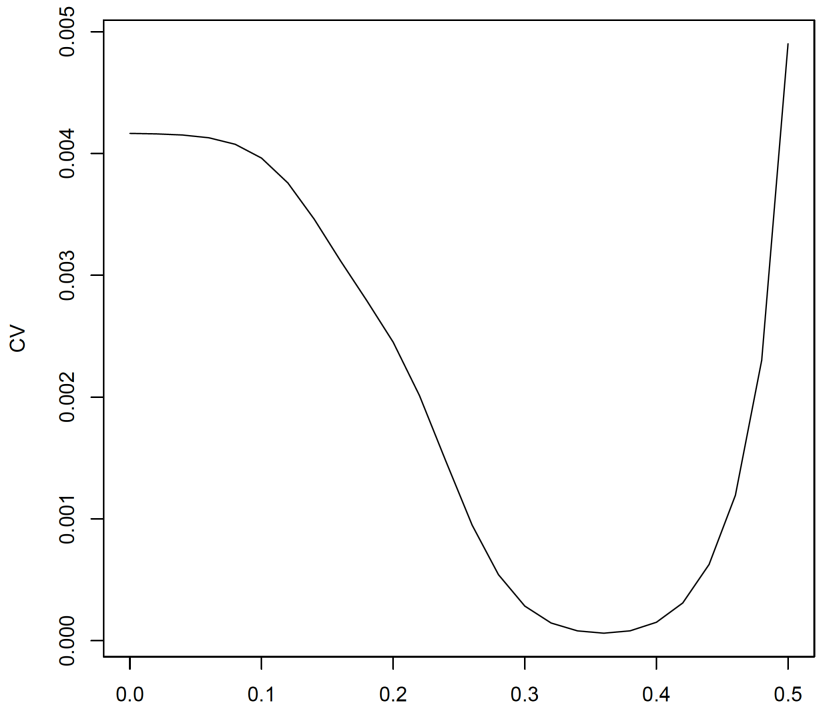

Figure 1.1 – Représentation classique de la fonction de validation croisée.