République Algérienne Démocratique et Populaire.
Ministère de lôEnseignement Supérieur et de la Recherche Scientifique.
Université Mentouri ï Constantine.
Faculté des Sciences Exactes.
Département de Mathématiques.

N° dôordre:
Série:

Mémoire

En vue de lôobtention du diplôme de

Magistère en Mathématiques

Option: Mathématiques Appliquées

Thème

Inférence Statistique dans les Modèles de Régression

Présenté par: Saadi Faiza

Soutenu le:

Devant le jury:

M. Denche	Prof	Université de Mentouri	Président.
Z. Mohdeb	Prof	Université de Mentouri	Rapporteur.
D. Boudaa	M.C	Université de Mentouri	Examinateur.
N. Nemouchi	M.C	Université de Mentouri	Examinatrice.

Soutenu le: ...../...../.....

Table des matières

Remerciements

Je remercie chaleureusement monsieur Mohdeb Zaher, Professeur à l'université Mentouri, Constantine, d'avoir accepté d'être l'encadreur de mon travail de mémoire.

Je remercie profondément monsieur M. Denche, professeur à l'université de Mentouri, Constantine, monsieur Z. Mohdeb, professeur à l'université de Mentouri, Constantine, monsieur D. Boudaa, maître de conférences à l'université de Mentouri, Constantine, madame N. Nemouchi maître de conférences à l'université de Mentouri, Constantine, d'avoir acceptés de faire partie des membres du jury de ce mémoire.

Je suis très reconnaissante à monsieur A. Chibat et monsieur A. Hemida de m'avoir aidé avec leurs conseils.

Je tiens à remercier particulièrement mon père, ma mère, mes frères, mes sœurs, mes amis, précisément I. Laroussi et Ch. Matmat pour leurs soutiens tout le long de mon travail au mémoire.

Résumé

La régression non paramétrique est un outil statistique permettant de décrire la relation entre une variable dépendante et une ou plusieurs variables explicatives, sans spécifier de forme stricte pour cette relation. Dans ce mémoire, nous *présentons une analyse d'une série de Fourier qui est employée dans de nombreuses sciences et techniques. Nous mettons donc l'accent sur un modèle de régression non paramétrique. Pour étudier ce type de modèle, on l'écrit sous forme d'un modèle paramétrique. Pour cette raison, on suppose que la fonction de régression est linéaire. L'estimation des paramètres est faite en appliquant la méthode de validation croisé qui est suivie par une étude asymptotique et un test d'hypothèse non paramétrique.*

Pour choisir un meilleur modèle d'ajustement, on applique la sélection avec le critère de l'information d'Akaike, où on décrit ces variantes corrigées, et on dérive leurs moments et probabilités et aussi on lie ces derniers à la performance via le concept du rapport signal-bruit.

A la fin, nous présentons un exemple de simulation pour étudier la consistance de nos estimateurs.

Mots Clés : *Régression non paramétrique, Régression linéaire multiple, La validation croisée, Le critère de l'information d'Akaike.*

Chapitre 1

Introduction

En statistique, on s'intéresse souvent à décrire et à comprendre les relations qui caractérisent certaines variables. Dans ce cas, la réalisation de l'étude fait généralement appel à des méthodes statistiques, puisqu'elles permettent d'obtenir des modèles qui tiennent compte d'une certaine partie de hasard dans les observations obtenues. Par ailleurs, les méthodes statistiques reposent habituellement sur des principes qui doivent être respectés pour que le modèle obtenu soit valide. L'avancement de la technologie a permis l'implantation de nouvelles méthodes de régression plus souples, qui laissent les données choisir la forme de la relation entre les variables. Ces méthodes sont regroupées sous l'appellation de *régression non paramétrique*. Plusieurs méthodes d'estimation non paramétriques ont été proposées au cours des années passées et elles possèdent toutes leurs avantages et leurs inconvénients. 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.

Il existe plusieurs méthodes de régression non paramétrique. Les plus connues

sont sûrement les fonctions de lissage, la méthode du noyau, ainsi que les fonctions splines qui, à elles seules, caractérisent plus d'un type d'estimateur (voir *Wegman* et *Wright* (1983)) dont les splines de régression et les splines de lissage. Ces méthodes permettent toutes de contrôler la flexibilité de l'estimateur.

Les estimateurs d'une fonction de régression μ obtenus de façon non paramétrique sont généralement appelés fonctions de lissage. Ces dernières lissent les données d'un échantillon pour obtenir des estimateurs qui se situent entre la régression paramétrique et l'interpolation des points. La flexibilité accordée à une fonction de lissage se contrôle habituellement par la valeur du paramètre de lissage λ, qui lui est associé. Les valeurs que peut prendre le paramètre de lissage dépendent du type de fonction de lissage.

L'objectif de la régression non paramétrique est d'estimer la relation de dépendance qui lie une variable d'intérêt Y, dite variable dépendante, a une variable explicative T a partir de couples d'observations $(t_j, y_j)_{j=1,...,n}$, tels que les t_j et les y_j sont reliés par le modèle de régression

$$y_j = \mu(t_j) + \varepsilon_j, j = 1, ..., n, \tag{1.1}$$

où μ est une fonction de régression inconnue que nous souhaitons estimer, et ε est le vecteur des erreurs ε_j, $j = 1, .., n$, centrées, de même variance σ^2 (homoscédasticité) et non corrélées entre elles ($\text{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, où $\delta_{ij} = 1$ lorsque $i = j$ et $\delta_{ij} = 0$ lorsque $i \neq j$, (δ_{ij} est le symbole de *Kronecker*)).

L'objectif de ce document est d'étudier l'inférence statistique d'une fonction de régression μ écrite sous la forme d'une série de *Fourier*. Historiquement, c'est les

astronomes qui ont été les premiers à utiliser l'analyse de *Fourier* pour des séries chronologiques. Leurs buts étaient de détecter des saisonnalités cachées au sein de leurs données. En 1772 et en 1778, *Lagrange* a utilisé ces méthodes pour détecter la périodicité cachée. Un demi-siècle plus tard, en 1847, *Buya* et *Ballot*, dans '*Les changements périodiques de températures*' ont proposés des méthodes pour étudier la périodicité de données astronomiques. On peut citer également d'autres travaux plus récents relevant de l'estimation dans les modèles de régression comme par exemple : *Berry*, *Carroll* et *Ruppert* (2002), dans '*Bayesian smoothing and regression splines for measurement error problems*' ; *Burnham* et *Anderson* (2002), dans '*Model Selection and multi-Model Inference'; Cai*, *Fan* et *Li* (2000), dans '*Efficient estimation and inferences for varying-coefficient models'; Delecroix* et *Thomas-Agnan* (2000) dans '*Spline and kernel regression under shape restriction'; ... ext.*

Ce document est composé de sept chapitres, dont le premier est consacré à une introduction. Le deuxième chapitre donne une description générale du modèle proposé, avec les suppositions introduites pour faciliter les calculs. Aussi, il donne une estimation de la fonction de régression basée sur une approximation de la décomposition en série de *Fourier,* et une estimation des coefficients de la fonction. Les techniques utilisées sont celles de la régression linéaire avec l'application de la méthode des moindres carrés.

Le troisième chapitre est consacré à l'estimation du risque de prédiction $P(\lambda)$. Cette estimation est faite par la minimisation de ce dernier où on trouve trois mesures pour sélectionner la valeur optimale λ.

Dans le quatrième chapitre nous étudions le comportement asymptotique du risque

($R(\lambda)$), du GCV, ainsi que la distribution asymptotique de l'estimateur obtenu.

Le chapitre cinq porte sur la construction de la statistique de test de signification des coefficients de la régression, complété par la mise en œuvre du test d'hypothèse. Le test est basé sur les deux tests de *Student* et de *Fisher-Snedecor.*

Le chapitre six présente la méthode de la sélection des modèles avec l'AIC et critères d'information dérivées.

Enfin dans le dernier chapitre, nous avons traité quelques exemples de simulation pour étudier la consistance de nos estimateurs.

Chapitre 2

Présentation du Modèle de Régression

2.1 La Description du Modèle

Soient $(t_j, y_j),\ j = 1, ..., n$ les n observations des deux variables indépendantes T et Y, tels que t_j et y_j sont liés par le modèle

$$y_j = \mu(t_j) + \varepsilon_j, j = 1, ..., n\ , \tag{2.1}$$

où μ est une fonction de régression inconnue que nous souhaitons estimer et (ε_j) est une suite de variable aléatoire, non covariées, de même loi de probabilité, telles que

$$\mathbb{E}\ (\varepsilon_j) = 0,\ var(\varepsilon_j) = \sigma^2,\ j = 1, .., n, \tag{2.2}$$

2.2 Cas d'un Modèle Linéaire

Supposons que les y_j, $j = 1, .., n$ sont obtenues comme des réponses de p variables d'entrées $X_1, ..., X_p$. Donc, y_j est la sortie obtenue par le vecteur d'entrées $\underline{X}_j =$

$(X_{j_1}, ..., X_{jp})'$, $j = 1, ..., n$, et $X' = [\ \underline{X}_1, ..., \underline{X}_n]$ à un rang égal à p.

Soit Λ l'ensemble consistant en tous les sous ensembles possibles d'indices des variables $X_1, ..., X_p$. Considérons

$$C(\Lambda) = \left\{ \underline{\mu}_\lambda / \ \lambda \in \Lambda \right\} \tag{2.3}$$

une classe d'estimateurs où $\underline{\mu} = (\mu_1, .., \mu_n)'$ est un vecteur moyen inconnu et Λ est un ensemble d'indices λ, (Le paramètre λ peut être un scalaire, ou un vecteur).

Les estimateurs de $\underline{\beta}$ et $\underline{\mu}$ sont donnés par la méthode des moindres carrés comme suit :

$$\underline{\beta}_\lambda = \left(X'_\lambda X_\lambda \right)^{-1} X'_\lambda \ \underline{y} \tag{2.4}$$

et

$$\underline{\mu}_\lambda = X_\lambda \ \underline{\beta}_\lambda = X_\lambda \ (\ X'_\lambda \ X_\lambda)^{-1} X'_\lambda \ \underline{y} \tag{2.5}$$

Pour chaque λ on associe une matrice $H(\lambda)$ d'ordre n, telle que

$$\underline{\mu}_\lambda = H \ (\lambda) \ \underline{y} \tag{2.6}$$

où

$$H \ (\lambda) = X_\lambda \ (X'_\lambda \ X_\lambda)^{-1} X'_\lambda. \tag{2.7}$$

Il est facile de vérifier que $H \ (H' = H)$ est symétrique et définie non-négative.

2.3 La Forme de L'Estimateur

L'un des plus importants espaces de fonctions est l'espace $L_2[a, b]$, l'ensemble de fonctions de carrés intégrable sur l'intervalle $[a, b]$.

$$L_2[a,b] = \left\{ g : [a,b] \to \mathbb{R} \text{ telle que } \int_a^b g^2(t)dt < \infty \right\}$$

Cet espace représente une riche collection de fonctions, parmi lesquelles, un système complet orthonormal ($CONS$) de $L_2[a,b]$ est fourni par l'exponentiel complexe (Une famille de fonctions $\{X_j\}$ est dite système orthonormé complet ($CONS$) si $\mu \perp X_j$ pour tout j que $\mu \equiv 0$).

Une des familles $CONS$ de $L_2[a,b]$ est donnée par la famille exponentielle suivante

$$X_j(t) = (b-a)^{-1/2} e^{\,[\,2\pi ijt/(b-a)]}, \;\; j = 0, \pm 1, ..., \tag{2.8}$$

où $e^{ix} = \cos x + i \, \sin x$ avec $i^2 = -1$.

Proposition 1 *Soit $\{X_j\}_{j=1}^{\infty}$ un $CONS$ de $L_2[a,b]$. Pour $\mu \in L_2[a,b]$, on définit*

$$\beta_j = \langle \, \mu, X_j \rangle \, , \;\; j = 1, 2, ... \tag{2.9}$$

Alors $\sum_{j=1}^{\lambda} \beta_j \, X_j$ est la meilleure approximation de μ dans le sens que

$$\left\| \mu - \sum_{j=1}^{\lambda} \beta_j X_j \right\| \leq \left\| \mu - \sum_{j=1}^{\lambda} \alpha_j X_j \right\|$$

pour tout $\underline{\alpha} = (\alpha_1, ..., \alpha_\lambda)' \in \mathbb{R}^\lambda$. En plus

$$\left\| \mu - \sum_{j=1}^{\lambda} \beta_j X_j \right\|^2 \;\; \to 0 \;\; \textit{quand } \lambda \to \infty.$$

Cette proposition déclare que la suite de fonctions $\sum_{j=1}^{\lambda} \beta_j X_j$ converges vers μ en norme de $L_2[a,b]$. On écrit ceci comme

$$\mu(t) \simeq \sum_{j=1}^{\infty} \beta_j \, X_j(t). \tag{2.10}$$

Comme conséquence de la proposition 1, on peut représenter μ (dans le sens de $L_2[a,b]$) comme une combinaison linéaire infinie des fonctions de base $\{X_j\}$. Les coefficients optimaux (2.9) utilisés dans l'expression (2.10) de μ sont appelés les coefficients de *Fourier* généralisés**,** tandis que l'expression même est appelé la série de *Fourier* généralisé de μ.

Les coefficients de *Fourier* généralisés satisfont,

$$\sum_{j=1}^{\infty} \mid \beta_j \mid^2 = \parallel \mu \parallel^2 \tag{2.11}$$

appelée la relation de *Parseval*. Par conséquent, les coefficients de *Fourier* généralisées des fonctions de $L_2[a,b]$ peuvent être vu comme des éléments de l'ensemble de la suite de carré sommable. L'inverse est aussi vraie ; si $\{\beta_j\}$ est de carré d'ordre sommable, alors $\mu = \sum_{j=1}^{\infty} \beta_j X_j$ est dans $L_2[a,b]$.

Remarque 2 *D'après la proposition* 1*, si* μ *est une fonction de* $L_2[a,b]$*, alors*

$$\mu(t) \sim (b-a)^{-1} \sum_{j=-\infty}^{\infty} \beta_j e^{[\,2\pi i j t / (b-a)]} \tag{2.12a}$$

où

$$\beta_j = \int_a^b \mu(s) e^{[\,-2\pi i j s / (b-a)]} ds. \tag{2.12b}$$

Ceci est l'expression de la série de Fourier classique de μ*.*

Dans ce cas (2.12*a*) *peut être exprimé sous une forme alternative pour* μ *un réel*

$$\begin{aligned} \beta_j^* &= \int_a^b \mu^*(t) e^{[\,2\pi i j t / (b-a)]} dt \\ &= \int_a^b \mu(t) e^{[\,2\pi i j t / (b-a)]} dt \\ &= \beta_{-j} \end{aligned}$$

donc

$$\begin{aligned}(b-a)\mu(t) &\sim \beta_0 + \sum_{j=1}^{\infty}\left\{\beta_j e^{[\,2\pi ijt/(b-a)]} + \beta_{-j}e^{[\,-2\pi ijt/(b-a)]}\right\} \\ &= \beta_0 + \sum_{j=1}^{\infty}\left\{\beta_j e^{[\,2\pi ijt/(b-a)]} + \beta_j^* e^{[\,-2\pi ijt/(b-a)]}\right\}.\end{aligned}$$

L'utilisation algébrique de la définition de l'exponentiel complexe indique que la série de Fourier de μ peut être exprimée en fonction de sinus et cosinus comme suit,

$$\frac{1}{b-a}\left[\beta_0 + \sum_{j=1}^{\infty}\left\{c_j \cos\left[\frac{2\pi jt}{b-a}\right] + s_j \sin\left[\frac{2\pi jt}{b-a}\right]\right\}\right] \tag{2.13a}$$

où

$$c_j = 2\int_a^b \cos\left[\frac{2\pi jx}{b-a}\right]\mu(x)dx \tag{2.13b}$$

et

$$s_j = 2\int_a^b \sin\left[\frac{2\pi jx}{b-a}\right]\mu(x)dx. \tag{2.13c}$$

Les deux formes (2.12) *et* (2.13) *sont équivalentes.*

Pour simplifier ce qui suit, on suppose que les points $t_1, ..., t_n$ de réalisations sont uniformément espacés dans $[a, b] = [0, 1]$ (pour obtenir l'orthogonalité de la fonction exponentiel), c'est-à-dire

$$t_j = \frac{j-1}{n}, j = 1, ..., n. \tag{2.14}$$

La matrice des réalisations est

$$X_\lambda = \left(e^{(2\pi ijr/n)}\right) \begin{array}{l} r = 0, n-1 \\ j = -\lambda, \lambda \end{array}. \tag{2.15}$$

En utilisant l'identité

$$\frac{1}{n}\sum_{r=0}^{n-1} e^{(\,2\pi ijr/n)}e^{(\,-2\pi ikr/n)} = \begin{cases} 1 \text{ si } j = k \\ 0 \text{ si } j \neq k \end{cases}, \tag{2.16}$$

on a

$$X_\lambda^* X_\lambda = nI$$

(où la notation $(^*)$ indique le complexe transposé du conjugué). Alors le $j^{\text{ème}}$ élément du vecteur

$$\underline{\beta}_\lambda = (\ X_\lambda^* X_\lambda)^{-1}\, X_\lambda^* \underline{y}$$

a pour expression,

$$\beta_{\lambda j} = \frac{1}{n}\sum_{r=1}^{n} y_r\ e^{[\ -2\pi i j(r-1)/n]},\ \ j = -\lambda, ..., \lambda \tag{2.17}$$

et l'estimation de la fonction de régression est alors donné par

$$\mu_\lambda(t) = \sum_{j=-\lambda}^{\lambda} \beta_{\lambda j}\ e^{\,2\pi i j t}. \tag{2.18}$$

Notons que, $\beta_{\lambda j}$ ne dépend pas de λ. $\beta_{\lambda j}$ est la $j^{\text{ème}}$ composante de β_λ dans l'expression de *Fourier* de μ_λ.

L'expression de $\mu_\lambda(t)$ peut être exprimé en terme de sinus et cosinus comme suit

$$\mu_\lambda(t) = \beta_{\lambda 0} + \sum_{j=1}^{\lambda} [\ c_{\lambda j}\cos(2\pi j t) + s_{\lambda j}\sin(2\pi j t)] \tag{2.19}$$

où

$$c_{\lambda j} = \frac{2}{n}\sum_{r=1}^{n} y_r \cos\left[\frac{2\pi\ j\ (r-1)}{n}\right] \tag{2.20a}$$

et

$$s_{\lambda j} = \frac{2}{n}\sum_{r=1}^{n} y_r \sin\left[\frac{2\pi\ j\ (r-1)}{n}\right]. \tag{2.20b}$$

Il y a une autre expression de μ_λ, qui peut être obtenue par la première écriture, (2.18) donnée par,

$$\mu_\lambda(t) = \frac{1}{n}\sum_{r=1}^{n} y_r \sum_{j=-\lambda}^{\lambda} e^{[\ 2\pi i j(t-t_r)]}. \tag{2.21}$$

Remarque 3 *En utilisant l'identité*

$$\sum_{j=-\lambda}^{\lambda} e^{2\pi ijx} = \frac{\sin\ [\ \pi\ (2\lambda+1)x]}{\sin\ (\pi x)} = K_\lambda(x), \tag{2.22}$$

on obtient

$$\mu_\lambda(t) = \frac{1}{n}\sum_{r=1}^{n} y_r K_\lambda(t-t_r). \tag{2.23}$$

La fonction $K_\lambda(x)$ est appelée le noyau de Dirichlet (Kernel of Dirichlet). C'est une fonction périodique de période égale à 1. Remarquons aussi que

$$\sum_{r=1}^{n} K_\lambda(t-t_r) = n.$$

Donc, l'expression de $\mu_\lambda(t)$ est une moyenne pondéré de y_r avec les poids obtenu à partir du noyau de Dirichlet.

Chapitre 3

Estimation de la Fonction de Régression

Considérons le modèle de régression

$$y_j = \mu(t_j) + \varepsilon_j, j = 1, ..., n,$$

où μ est une fonction réelle définie sur l'intervalle $[a, b]$, t_j est un échantillonnage fixé sur $[a, b]$; (ε_j) est une suite de variables aléatoires indépendantes, centrées de même variance σ^2.

On pose $\underline{\mu} = (\mu_1, ..., \mu_n)'$ où $\mu_j = \mu(t_j)$, $j = 1, ..., n$; $\underline{y} = (y_1, ..., y_n)'$, $\underline{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$.

Considérons une classe d'estimateurs pour $\underline{\mu}$ donnée par

$$C(\Lambda) = \left\{ \underline{\mu}_\lambda / \ \lambda \in \Lambda \right\}$$

avec Λ représentant un ensemble d'indices. Le paramètre λ peut être un scalaire, un vecteur ou même un ensemble.

Pour simplifier l'étude, nous supposerons aussi que les éléments de $C(\Lambda)$ sont des estimateurs linéaires. Ce qui signifie que pour tout $\lambda \in \Lambda$, il existe une $(n \times n)$ matrice

$H(\lambda)$ tel que

$$\underline{\mu}_\lambda = H(\lambda)\ \underline{y}.$$

$H(\lambda)$ est supposé symétrique et définie positive.

3.1 L'Estimation

Il existe des critères qui sont largement utilisés pour choisir un meilleur estimateur. Voici quelques exemples.

Définissons la fonction perte en estimant $\underline{\mu}$ par

$$L(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\ \left[\ (\mu_j - \mu_{\lambda_j})^2\right] \tag{3.1}$$

où μ_{λ_j} est le $j^{\text{ème}}$ élément de $\underline{\mu}_\lambda$. $L(\lambda)$ est le carré de la distance euclidienne entre $\underline{\mu}$ et $\underline{\mu}_\lambda$ multiplié par le facteur n^{-1}.

On définit le risque comme étant la moyenne de la fonction perte donné par

$$\begin{aligned} R(\lambda) &= \mathbb{E}\ [L(\lambda)\ ] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\ (\mu_j - \mu_{\lambda j})^2\right] \\ &= \frac{1}{n}\mathbb{E}\ \left[\ (\underline{\mu} - \underline{\mu}_\lambda)'(\underline{\mu} - \underline{\mu}_\lambda)\right]. \end{aligned} \tag{3.2}$$

Les valeurs de ces critères sont petites indiquant une bonne estimation.

Un autre critère de mesure est le risque de prédiction. Le principe de mesure de l'erreur de prédiction est le suivant.

Supposons que nous envisageons d'observer n nouvelles observations $\underline{y}^* = \underline{\mu} + \underline{\varepsilon}^*$, avec $\underline{y}^* = (y_1^*, ..., y_n^*)'$, $\underline{\mu} = (\mu(t_1), ..., \mu(t_n))'$, $\underline{\varepsilon}^* = (\varepsilon_1^*, ..., \varepsilon_n^*)'$, où $\underline{\varepsilon}^*$ un vecteur de variables aléatoires centrées, indépendantes, de variance commune σ^2.

pour évaluer la performance de $\underline{\mu}_\lambda$ comme un prédicteur des futures observations, on peut utiliser

$$\begin{aligned} P(\lambda) &= \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\left[\ (y_j^* - \mu_{\lambda_j})^2\right]\right] \\ &= \sigma^2 + R(\lambda). \end{aligned} \tag{3.3}$$

Un estimateur qui minimise le risque $R(\lambda)$, minimise également $P(\lambda)$ et réciproquement.

L'idéal est de sélectionner un certain λ qui minimise le risque (qui est équivalent à $P(\lambda)$) ou la perte. L'estimateur qui en découle doit fournir un bon estimateur de $\underline{\mu}$. Malheureusement, ni $L(\lambda)$ ni $R(\lambda)$ ne peuvent actuellement être calculés sans connaître $\underline{\mu}$ qui est inconnue. Donc en pratique, le critère ($R(\lambda)$ ou $L(\lambda)$) doit être estimé par les données observées et ensuite le minimiser pour trouver un λ qui donne une meilleure estimation de $\underline{\mu}$.

La fonction perte, le risque et le risque de prédiction fournissent trois mesures de performance possibles d'un estimateur de $\underline{\mu}$. Part ailleurs, ils en existent d'autres, par exemple, au lieu d'utiliser le carré de l'erreur, on peut utiliser la valeur absolue. Cependant, d'un point de vue mathématique le carré des erreurs est généralement plus facile à calculer.

3.2 L'Estimation du Risque de Prédiction

Dans cette partie, nous allons construire un estimateur de $P(\lambda)$. L'estimateur naïf de $\mathbb{E}\left[(y_j^* - \mu_{\lambda j})^2\right]$ est dans ce cas le carré des résidus $\left(y_j - \mu_{\lambda j}\right)$. D'où le choix

d'estimer $P(\lambda)$ par l'erreur moyenne quadratique MSE définie par

$$\begin{aligned} MSE(\lambda) &= \frac{1}{n}\sum_{j=1}^{n}\left[\ (\ y_j - \mu_{\lambda j})^2\right] \\ &= \frac{1}{n}(\ \underline{y} - \underline{\mu}_\lambda)'(\underline{y} - \underline{\mu}_\lambda) \\ &= \frac{1}{n}\underline{y}'(\ I - H(\lambda))^2\underline{y}. \end{aligned}$$

D'après cette égalité, on a $\mathbb{E}\left[MSE(\lambda)\right]$ est égal à $R(\lambda)$. Pour évaluer cette moyenne, nous avons besoin du lemme suivant :

Lemme 4 *Soit $\underline{z}$ un vecteur aléatoire à n composantes, de moyenne $\underline{\theta}$ et de matrice de variance-covariance Σ. Si A est une matrice symétrique $(n \times n)$, alors*

$$\mathbb{E}\ [\underline{z}'A\ \underline{z}] = \underline{\theta}'A\ \underline{\theta} + \mathrm{tr}(\ A\Sigma)$$

où $\mathrm{tr}(A\Sigma)$ est la trace de $A\Sigma$. Si en plus, $\underline{z}$ est un vecteur aléatoire de loi normale alors

$$\mathrm{Var}\left[\ \underline{z}'A\underline{z}\right] = 2\,\mathrm{tr}[\ (A\Sigma)^2] + 4\ \underline{\theta}'A\ \Sigma\ A\ \underline{\theta}.$$

Preuve. Voir *Searle (*1971*, Chapitre* 2*).* ■

Une application directe du lemme 1 pour $\underline{z} = \underline{y}$, $\underline{\theta} = \underline{\mu}$ et $\Sigma = \sigma^2 I$ nous donne,

$$\begin{aligned} \mathbb{E}\left[\ MSE(\lambda)\right] &= \frac{1}{n}\underline{\mu}'(\ I - H(\lambda))^2\underline{\mu} + \frac{1}{n}\ \sigma^2\,\mathrm{tr}\ \left[(I - H(\lambda))^2\right] \\ &= \frac{1}{n}\underline{\mu}'(\ I - H(\lambda))^2\underline{\mu} + \sigma^2 + \frac{1}{n}\ \sigma^2\,\mathrm{tr}[H(\lambda)^2] \\ &\quad -2\ \frac{1}{n}\ \sigma^2\,\mathrm{tr}[H(\lambda)]. \end{aligned} \tag{3.4}$$

Et nous avons

$$\begin{aligned} P(\lambda) &= \sigma^2 + R(\lambda) \\ &= \sigma^2 + \frac{1}{n}\mathbb{E}\ \left[(\underline{\mu} - \underline{\mu}_\lambda)'(\underline{\mu} - \underline{\mu}_\lambda)\right] \\ &= \sigma^2 + \frac{1}{n}\underline{\mu}'(\ I - H(\lambda))^2\underline{\mu} + \frac{1}{n}\sigma^2\ \text{tr}\ [H(\lambda)^2]. \end{aligned} \tag{3.5}$$

Cette dernière égalité est obtenu en utilisant le lemme précédent pour $\underline{z} = \underline{\varepsilon}$, $\underline{\theta} = \underline{0}$ et $\Sigma = \sigma^2 I$, et le fait que $\underline{\mu}_\lambda = H(\lambda)\ \underline{y} = H(\lambda)\ \ \underline{\mu} + H(\lambda)\ \underline{\varepsilon}$.

En comparant les deux équations (3.4), (3.5), nous obtenons que $MSE(\lambda)$ est un estimateur biaisé de $P(\lambda)$ avec un biais égale à $-2n^{-1}\sigma^2\,\text{tr}[H(\lambda)]$.

Si σ^2 est connu, l'estimateur sans biais de $P(\lambda)$ est par conséquent donné par

$$\hat{P}(\lambda) = MSE(\lambda) + 2\ \frac{1}{n}\sigma^2\ \ \text{tr}[H(\lambda)] \tag{3.6}$$

et l'estimateur sans biais du risque est

$$\hat{R}(\lambda) = \hat{P}(\lambda) - \sigma^2. \tag{3.7}$$

En général, la dépendance de (3.6) et (3.7) de σ^2 ne constitue pas un sérieux problème.

Pour notre modèle, il existe dans la littérature de nombreux estimateurs de σ^2. Par exemple, si on pose $\#(\lambda)$ le nombre d'estimateurs dans λ, alors

$$\frac{n\ MSE(\lambda)}{n - \#(\lambda)}$$

fournit un estimateur de σ^2 pour tout $\lambda \in \Lambda$.

Une des approches est d'utiliser l'estimateur de σ^2 correspondant à $\lambda = \{1, ..., p\}$, c'est-à-dire l'estimateur obtenu lorsque toutes les variables sont utilisées en estimant $\underline{\mu}$.

Il existe un cas pour lequel l'estimation de $P(\lambda)$ n'exige pas l'estimation de σ^2, même lorsque σ^2 est inconnu. Cette situation existe, lorsque $\text{tr}[H(\lambda)] = 0$ pour tout $\lambda \in \Lambda$. Les estimateurs de ce type sont en général appelé "*estimateurs de trace-zéro*" (*nil-trace estimators*).

Beaucoup d'estimateurs n'ont pas la propriété de trace-zéro ; cependant, il y a plusieurs situations pour lesquelles, il n'existe pas d'estimateurs évidents de σ^2, en particulier dans le cas de l'étude d'un modèle non paramétrique. Donc $\hat{P}(\lambda)$ et $\hat{R}(\lambda)$ ne sont pas entièrement des estimateurs satisfaisants. Par conséquent, nous considérons maintenant deux estimateurs du risque de prédiction qui n'exige pas d'estimation de σ^2.

Un estimateur possible de $P(\lambda)$ qui n'exige pas d'estimation de σ^2 peut être motivé par une autre étude de $MSE(\lambda)$. Dans l'utilisation du $MSE(\lambda)$ pour estimer $P(\lambda)$, nous avons utilisé $y_j - \mu_{\lambda j}$ comme estimateurs des futures erreurs de prédiction. Puisque $\mu_{\lambda j}$ est construit en utilisant les données $y_1, ..., y_n$, nous devons anticiper que $\mu_{\lambda j}$ doit être meilleur dans la prédiction y_j qu'une future réponse y_j^*. Comme un résultat, on doit examiner que $MSE(\lambda)$ doit être plus petit en moyenne que $P(\lambda)$. Ce que est, en fait le cas. D'après (3.4) et (3.5), $MSE(\lambda)$ tend à sous-estimer $P(\lambda)$ par un facteur de $2\ n^{-1}\sigma^2\ \text{tr}[H(\lambda)]$.

Une approche intuitive pour corriger $MSE(\lambda)$ concernant la sous-estimation de $P(\lambda)$ peut être obtenue en remplaçant $(y_j - \mu_{\lambda j})^2$ par une quantité qui entrainerait une meilleur estimation de $\mathbb{E}\left[(y_j^* - \mu_{\lambda j})^2\right]$. Pour donner quelques indications, essayons de voir comment la situation se découle, lorsque nous avons des observations multiples

pour chaque moyenne. Considérons par exemple

$$y_{kj} = \mu_j + \varepsilon_{kj},\ k = 1, 2,\ j = 1, ..., n.$$

En d'autres termes, au lieu de considérer une observation pour chaque moyenne, nous en considérons deux. Alors, un estimateur de $\underline{\mu}$ peut être obtenu en utilisant $\underline{y} = (y_{11}, ..., y_{1n})'$ et $\mathbb{E}\left[(y_j^* - \mu_{\lambda j})^2\right]$ peut être estimé en utilisant $(y_{2j} - \mu_{\lambda j})^2$. Puisque y_{2j} n'est pas utilisé dans le calcul de $\mu_{\lambda j}$, nous n'allons pas établir les propriétés de prédiction de l'estimateur par les données utilisées pour sa construction. Par conséquent, $n^{-1}\sum_{j=1}^{n}\left[\ (\ y_{2j} - \mu_{\lambda j})^2\right]$, avec $\mu_{\lambda j}$ basé seulement sur y_{1r}, $r = 1, ..., n$, entraine un meilleur estimateur de $P(\lambda)$ que $MSE(\lambda)$. Il peut être montré que cet estimateur est sans biais de $P(\lambda)$, c'est-à-dire que

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}(\ y_{2j} - \mu_{\lambda j})^2\right] = P(\lambda).$$

Bien sûr, nous n'avons pas en général des observations multiples pour chaque μ_j. Aussi, si c'est le cas, il est préférable de les inclure dans le processus de l'estimation plutôt que les considérer en dehors de la validation de l'estimateur. Cependant, l'idée de l'utilisation d'un sous échantillon de données pour l'estimation de l'erreur de prédiction suggère d'autre approche appliquée dans certaines situations.

Dans quelques cas, il est possible de calculer un analogue ou une approximation de l'estimateur $\mu_{\lambda j}$ qui n'intègre pas la $j^{\text{ème}}$ observation. Dans de telles situations, les données peuvent être regroupées en n sous échantillons de taille $(n-1)$ et l'observation qui n'est pas dans chaque sous échantillon est utilisée pour l'estimation de l'erreur de prédiction. Avant de donner une formulation générale, il peut être utile de considérer un exemple où ce type de développement est possible.

Pour les problèmes de sélection de la variable, il est possible d'éliminer le point $(\underline{X}_j^{'}, y_j)$ de notre approche, où $\underline{X}_j = (X_{j_1}, ..., X_{jp})'$, et pour tous les sous ensembles donnés, on recalcule les coefficients estimés. Plus précisément, pour le sous-ensemble de variables avec les indices dans λ, soit la matrice d'éléments X_{kr},

$$X_{\lambda(j)} = \{X_{k\,r}\} \quad \begin{matrix} k = 1, ..., j-1, j+1, ..., n \\ r \in \lambda \end{matrix}$$

et soit

$$\underline{y}_{(j)} = (\ y_1, ..., y_{j-1}, y_{j+1}, ..., y_n)'$$

Alors

$$\underline{\beta}_{\lambda(j)} = (\ X'_{\lambda(j)} X_{\lambda(j)})^{-1} X'_{\lambda(j)} \underline{y}_{(j)}$$

est l'estimateur des moindres carrés des coefficients pour les variables X_r, $r \in \lambda$, calculés sans l'utilisation de $(\underline{X}_j^{'}, y_j)$.

Une estimation de μ_j est fournit par

$$\mu_{\lambda(j)} = \underline{X}'_{\lambda j}\ \underline{\beta}_{\lambda(j)}$$

avec $\underline{X}_{\lambda j}$ le vecteur des valeurs de X avec les indices dans λ, correspondant à la $j^{\text{ème}}$ réponse.

L'estimation $\mu_{\lambda(j)}$ est directement relié à $\mu_{\lambda j}$ dans le sens que tous les deux sont des estimateurs des moindres carrés de μ_j impliquant les prédicteurs indexés par λ.

En fait, on peut montrer que

$$\mu_{\lambda(j)} = \mu_{\lambda j} - \frac{h_{jj}(\lambda)}{(\ 1 - h_{jj}(\lambda))}(\ y_j - \mu_{\lambda j}) \tag{3.8}$$

où $h_{jj}(\lambda)$ est le $j^{\text{ème}}$ élément diagonal de $H(\lambda)$ (voir,e.g., *Gunst* et *Mason* 1980, chapitre 7).

En revenant au cas général, supposons que $\mu_{\lambda(j)}$ représente un parallèle de $\mu_{\lambda j}$ calculé sans la $j^{\text{ème}}$ observation ; l'estimateur (3.8) en est un exemple.

Alors, l'estimateur $\mu_{\lambda(j)}$ est construit à partir du sous-échantillon de taille $(n-1)$ pris de l'échantillon d'origine et la $j^{\text{ème}}$ observation est traitée comme une observation additionnelle. Notre discussion précédente suggère l'utilisation de $(y_j - \mu_{\lambda(j)})^2$ pour estimer $\mathbb{E}(y_j^* - \mu_{\lambda j})^2$ et l'estimation de $P(\lambda)$ par

$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n} \left[\left(y_j - \mu_{\lambda(j)} \right)^2 \right]. \tag{3.9}$$

$CV(\lambda)$ est appelé l'erreur moyenne quadratique de la validation croisée (*cross-validation mean squared error*)**.** La sélection de λ à travers la minimisation de $CV(\lambda)$ est appelée la validation croisée (CV). La quantité $nCV(\lambda)$ est quelques fois appelée prédiction de la somme des carrées (*prediction sum of squares*) (*Allen* 1974).

Cependant, puisque il existe qu'une valeur pour chaque $\mu_j = \mu(t_j)$, on montre que le $CV(\lambda)$ est généralement biaisé pour $P(\lambda)$ (il est cependant sans biais pour le risque de prédiction qui est reliée à $P(\lambda)$ dans le cas de l'estimateur trace-zéro). *Stone (*1974*)* et *Geisser (*1975*)* ont considéré des schémas de validation croisée plus élaborés que celui traité ici. Divers applications du CV sont fournies par *Titterington (*1985*)*. *Efron (*1982*)* a donné la comparaison des CV pour la technique de réutilisation d'échantillons.

Une autre méthode utile pour la sélection de λ est la validation croisée généralisée (*Generalized Cross Validation*) (GCV). En supposant que tr $[H(\lambda)] \leq n$, le critère

de GCV est défini par

$$\begin{aligned} GCV(\lambda) &= \frac{MSE(\lambda)}{(\ \frac{1}{n}\ \text{tr}\ [\ I - H(\lambda)])^2} \\ &= \frac{\frac{1}{n}\sum_{j=1}^{n}(\ y_j - \mu_{\lambda(j)})^2}{\left(\ \frac{1}{n}\,\text{tr}[\ I - H(\lambda)]\right)^2}. \end{aligned} \tag{3.10}$$

Bien que l'exemple précédent montre que GCV et CV peuvent être reliés, il montre aussi que la validation croisée n'est pas un cas spécial de GCV. Donc le titre "*validation croisée généralisée*" est quelque part une fausse appellation.

Cependant, la meilleur motivation pour le critère (3.10) est probablement fournie par le théorème appelé "*Théorème GCV* " établi par *Craven* et *Wahba (*1979*) (*voir également *Wahba* (1977*, c)* et *Golub, Heath et Wahba (*1979*)* pour d'autres variantes du théorème).

Une forme générale du théorème GCV peut être formulée comme suit.

Théorème 5 ***(Le Théorème*** GCV***)***

Soit $\tau_j(\lambda) = n^{-1}\,\text{tr}[\ H(\lambda)^j]\ ,\ j = 1, 2,$ *et supposons que* $\tau_1(\lambda) \leq 1,$ *alors*

$$\frac{|\ \mathbb{E}GCV(\lambda) - P(\lambda)\ |}{R(\lambda)} \leq g(\lambda)$$

où

$$g(\lambda) = \frac{[\ 2\tau_1(\lambda) + \tau_1(\lambda)^2 \diagup \tau_2(\lambda)]}{(\ 1 - \tau_1(\lambda))^2}. \tag{3.11}$$

On peut interpréter le contenu de ce théorème en disant que, si $g(\lambda)$ est petit, alors une conséquence de ce théorème est que la distance entre $P(\lambda)$ et $\mathbb{E}\,[GCV(\lambda)]$ est petite relativement à $R(\lambda)$. Dans ce cas, $GCV(\lambda)$ est presque un estimateur sans biais de $P(\lambda)$.

Remarque 6 *Il-y-a en fait beaucoup de cas où $g(\lambda)$ est petit, particulièrement pour les plus grandes tailles de l'échantillon. Par exemple, dans le problème de la sélection de variable*

$$\tau_1(\lambda) = \frac{1}{n}\,\mathrm{tr}\ [\ H(\lambda)] = \frac{\#(\lambda)}{n} \leq \frac{p}{n}$$

et puisque $H(\lambda)$ est idempotente, $\tau_2(\lambda) = \tau_1(\lambda)$. Donc,

$$g(\lambda) = \frac{3\tau_1(\lambda)}{(1-\tau_1(\lambda))^2} \leq \frac{3\frac{p}{n}}{\left(1-\frac{p}{n}\right)^2} = 3\frac{p}{n} + o\left((\frac{p}{n})^2\right).$$

comme p est fixé, alors $g(\lambda)$ est petit pour n assez grand.

Preuve. En utilisant le *Lemme* 4, on peut montrer que

$$\begin{aligned}
\mathbb{E}\,[\ GCV(\lambda)] &= \frac{\frac{1}{n}\mathbb{E}\ \left[\ \underline{y}'(I-H(\lambda))^2\ \underline{y}\right]}{(1-\tau_1(\lambda))^2} \\
&= \frac{\frac{1}{n}\left\{\ \underline{\mu}'(I-H(\lambda))^2\underline{\mu}\ +\ \sigma^2\ \mathrm{tr}\ [(I-H(\lambda))^2]\right\}}{(1-\tau_1(\lambda))^2}.
\end{aligned}$$

Donc, en vue de (3.3) et (3.5)

$$\begin{aligned}
\mathbb{E}\,[\ GCV(\lambda)] - P(\lambda) &= \frac{R(\lambda)}{(1-\tau_1(\lambda))^2} + \sigma^2\left(\frac{1-2\tau_1(\lambda)}{(1-\tau_1(\lambda))^2}\right) - (R(\lambda)+\sigma^2) \\
&= \left(\frac{\tau_1(\lambda)\ (2-\tau_1(\lambda))}{(1-\tau_1(\lambda))^2}\right) R(\lambda) - \sigma^2\left(\frac{\tau_1(\lambda)^2}{(1-\tau_1(\lambda))^2}\right).
\end{aligned}$$

Ce qui entraine que

$$\begin{aligned}
\frac{|\ \mathbb{E}\,[\ GCV] - P(\lambda)|}{R(\lambda)} &= \left|\frac{\tau_1(\lambda)\ (2-\tau_1(\lambda))}{(1-\tau_1(\lambda))^2} - \sigma^2\left(\frac{\tau_1(\lambda)^2}{R(\lambda)(1-\tau_1(\lambda))^2}\right)\right| \\
&\leq \frac{\tau_1(\lambda)\ (2-\tau_1(\lambda))}{(1-\tau_1(\lambda))^2} + \frac{\tau_1(\lambda)^2/\tau_2(\lambda)}{(1-\tau_1(\lambda))^2} \\
&\leq g(\lambda).
\end{aligned}$$

Pour obtenir les deux dernières inégalités, nous avons utiliser le fait que $R(\lambda) \geq \sigma^2\tau_2(\lambda)$ et que $\tau_1(\lambda) \leq 1$ qui nécessite que $1 \leq 2-\tau_1(\lambda) \leq 2$. ■

Ainsi, nous avons maintenant trois critères disponibles $\hat{P}(\lambda)$ (et $\hat{R}(\lambda)$), $CV(\lambda)$ et $GCV(\lambda)$ pour sélectionner une bonne valeur de λ.

Comme $\#(\lambda)$ dénote le nombre d'éléments dans λ, et du fait que $\mathrm{tr}[\, H(\lambda)] = \#(\lambda)$ on trouve que notre estimateur du risque de prédiction s'écrit sous la forme suivante

$$\begin{aligned} \hat{P}(\lambda) &= MSE(\lambda) + 2\ \hat{\sigma}^2 \left(\frac{\#(\lambda)}{n} \right) \\ \text{et} \qquad GCV(\lambda) &= \frac{MSE(\lambda)}{\left(1 - \frac{\#(\lambda)}{n}\right)}\ . \end{aligned} \tag{3.12}$$

Chapitre 4

Etude Asymptotique

4.1 Introduction

Supposons que nous observons les valeurs d'une variable réponse Y d'un modèle de régression de la forme

$$y_j = \mu(t_j) + \varepsilon_j, j = 1, ..., n \ , \tag{4.1}$$

où μ est une fonction réelle inconnue à estimer définie sur l'intervalle $[0, 1]$, $\underline{\varepsilon} = (\varepsilon_1, .., \varepsilon_n)'$ est un vecteur aléatoire centré tel que les ε_j soient non corrélées de même variance σ^2.

Pour motiver le type d'estimateurs que nous allons étudier, supposons que le modèle est linéaire et qu'il existe des fonctions $x_1, ..., x_\lambda$ connues telles que $\mu(t) = \sum_{j=1}^{\lambda} \beta_j x_j(t)$, avec $\beta_1, ..., \beta_\lambda$ des coefficients réels appelés coefficients de régression.

Estimer μ, revient donc à estimer $\beta_1, ..., \beta_\lambda$. Ceci peut être accompli par plusieurs méthodes dont une des méthodes classique est la méthode des moindres carrés.

L'utilisation d'un modèle linéaire pour μ suppose que la forme de μ est connue,

ce qui n'est pas souvent le cas. Lorsque la connaissance de μ est restreinte, on préfère élargir la classe de fonctions à laquelle μ est supposée appartenir. Par exemple, on peut supposer que μ est lisse dans le sens qu'elle soit continue et différentiable, ou d'autres hypothèses de cette nature.

Lorsque μ est supposée appartenir à une classe générale de fonctions (dimension infinie), le modèle (4.1) est appelé "modèle de régression non paramétrique". Un modèle de ce type peut parfois être approché par un modèle linéaire. Mais, parfois cette approximation ne peut pas se faire sous de nouvelles hypothèses, telle que la continuité, la possibilité de développer la fonction de régression en une somme infinie, etc,.... Cette approximation nous permet d'adapter les techniques classiques de l'inférence des modèles linéaires.

Plusieurs classes de fonctions dont les éléments admettent un développement de la forme de modèles linéaires infinis, c'est-à-dire de la forme

$$\mu(t) = \sum_{j=1}^{\infty} \beta_j x_j(t)$$

pour une famille de fonctions $\{x_j\}$.

Dans ce cas, une approche possible est d'approximer μ par $\sum_{j=1}^{\lambda} \beta_j x_j(t)$, pour un entier λ et ensuite estimer les β_j pour obtenir un estimateur de μ, appelé "estimateur série".

Ce chapitre est consacré estimateurs série dont le développement suivant une base de fonctions orthogonales.

On notera $L_2[a, b]$ l'espace des fonctions de carré intégrables sur $[a, b]$, muni du produit scalaire usuel. Pour toutes fonctions μ_1, $\mu_2 \in L_2[a, b]$, on définit le produit

scalaire par

$$\langle\ \mu_1, \mu_2\rangle = \int_a^b \mu_1(t)\mu_2(t)dt$$

et la norme par

$$\|\ \mu_1\| = \left(\int_a^b \mu_1^2(t)dt\right)^{1/2}.$$

Puisque il est souvent raisonnable de supposer que μ dans le modèle (4.1) est une fonction lisse, une direction pour satisfaire cette condition est de s'intéresser à une classe de fonctions lisses au sens que ces fonctions satisfassent les conditions de continuité et de différentiabilité.

Une de ces collections est

$$C^m[a,b] = \left\{\ \mu / \mu^{(j)} \text{ est continue, } j = 0, 1, ..., m\right\}, \tag{4.2}$$

avec la convention $C^0[a,b] = C[a,b]$.

En d'autres termes, $C^m[a,b]$ est l'ensemble de toutes les fonctions sur $[a,b]$ avec les m dérivées continues. Il est clair que $L_2[a,b] \supset C[a,b] \supset C^1[a,b] \supset ....$

Dans notre travail, il n'est pas nécessaire pour μ d'avoir les m dérivées continues. Il suffit pour $\mu^{(m)}$ d'être de carré intégrable.

On définit, alors l'espace de *Sobolev* d'ordre m par

$$W_2^m[a,b] = \left\{\ \mu\ /\ \mu^{(j)} \text{ absolument continue, } j = 0, 1, ..., m-1 \text{ et } \mu^{(m)} \in L_2[a,b]\right\}. \tag{4.3}$$

Notons que $W_2^m[a,b] \supset C^m[a,b]$.

Lorsque $m = 0$, nous avons $W_2^0[a,b] = L_2[a,b]$.

4.2 Comportement Asymptotique du Risque

Rappelons la forme de l'estimateur μ_λ de la fonction de régression μ introduite dans le deuxième chapitre

$$\mu_\lambda(t) = \sum_{j=-\lambda}^{\lambda} \beta_{\lambda j}\ e^{2\pi i j t} \tag{4.4}$$

où

$$\beta_{\lambda j} = \frac{1}{n}\sum_{r=1}^{n} y_r e^{[\ -2\pi i j(r-1)/n]},\ j = -\lambda, ..., \lambda. \tag{4.5}$$

L'expression simple précédente pour μ_λ permet d'obtenir une forme simple pour le risque $R_n(\lambda)$.

Dans cette section, nous allons exploiter cet avantage pour étudier le comportement asymptotique du risque lorsque n et λ augmentent.

On définit le risque, lorsqu'on estime la fonction de régression μ par μ_λ, par

$$R_n(\lambda) = \frac{1}{n}\sum_{r=0}^{n-1} \mathbb{E}\left|\ \mu\left(\frac{r}{n}\right) - \mu_\lambda\left(\frac{r}{n}\right)\right|^2.$$

L'objet de cette partie d'étudier le comportement de $R_n(\lambda)$ comme fonction de n et λ sous certaines restrictions sur la fonction μ.

Nous allons d'abord supposer que $\mu \in W^2_{2,per}[0,1]$, où $W^2_{2,per}[0,1] = \{\ \mu \diagup \mu \in W^2_2[0,1]$ et $\mu(0) = \mu(1), \mu'($

Notons que $W^2_{2,per}[0,1]$ représente une classe de fonctions lisses et périodiques. Par ailleurs, il est bien connu que la série de *Fourier* pour μ converge uniformément en t vers μ sous cette hypothèse (voir par exemple *Davis* (1975), chapitre XII).

Puisque

$$\mathrm{tr}\, X_\lambda(\ X^*_\lambda X_\lambda)^{-1} X^*_\lambda = \frac{1}{n}\,\mathrm{tr}\,[\ X^*_\lambda X_\lambda] = 2\lambda + 1,$$

où

$$X_\lambda = \left(e^{(2\pi ijr/n)}\right) \begin{array}{l} r = 0, ..., n-1 \\ \\ j = -\lambda, ..., \lambda \end{array},$$

d'après l'étude faite dans le chapitre 3, on peut vérifier que

$$R_n(\lambda) = \frac{1}{n}\sum_{r=0}^{n-1} \left| \ \mu\left(\frac{r}{n}\right) - \mathbb{E}\ \left[\ \mu_\lambda\left(\frac{r}{n}\right)\right]\right|^2 + \frac{2\lambda+1}{n}\ \sigma^2. \tag{4.6}$$

Puisque la somme dans l'expression de (4.6) est non négative, il s'ensuit que $R_n(\lambda)$ est minorée par zéro pourvu que $\frac{\lambda}{n} \to 0$ quand $n \to \infty$.

Donc une condition nécessaire pour que le risque tende vers zéro est que le nombre de termes dans l'estimateur (4.4) de la série de *Fourier* soit plus petit que la taille de l'échantillon n. Il s'agit donc d'étudier la quantité $\frac{1}{n}\sum_{r=0}^{n-1} |\ \mu(r/n) - \mathbb{E}\,[\ \mu_\lambda(r/n)]|^2$.

Commençons par développer l'expression de $\mu(t) - \mathbb{E}\,[\ \mu_\lambda(t)]$. Puisque μ est représenté uniformément par sa série de *Fourier*, on peut donc écrire en particulier

$$\mu\left(\frac{k}{n}\right) = \sum_{j=-\infty}^{\infty} \beta_j\, e^{(\ 2\pi ijk\ /n)},\ k = 0, ..., n-1, \tag{4.7}$$

où

$$\beta_j = \int_0^1 \mu(t)\ e^{-2\pi ijt}\ dt.$$

On a

$$\mathbb{E}\ \ [\ \mu_\lambda(t)] = \sum_{j=-\lambda}^{\lambda} \mathbb{E}\left(\ \beta_{\lambda j}\right)\ e^{2\pi ijt}.$$

Par ailleurs, en utilisant les relations (4.4) et (4.7), on a

$$\begin{aligned} \mathbb{E}\ \ [\ \beta_{\lambda j}] &= \frac{1}{n}\sum_{k=0}^{n-1} \mu\left(\frac{k}{n}\right) e^{(\ -2\pi ijk/n)} \\ &= \sum_{r=-\infty}^{\infty} \beta_r \left\{\frac{1}{n}\sum_{k=0}^{n-1} e^{[\ -2\pi ik(j-r)/n]}\right\} \\ &= \beta_j + \sum_{r\neq 0} \beta_{j+nr}. \end{aligned} \tag{4.8}$$

(L'expression (4.8) est obtenue, en utilisant le fait que $\cos(2k\pi) + i\sin(2k\pi) = 1$).

En combinant les résultats (4.7) et (4.8), on obtient

$$\mu(t) - \mathbb{E}\left[\ \mu_\lambda(t)\right] = \sum_{|j|>\lambda} \beta_j e^{2\pi i j t} - \sum_{|j|\leq\lambda} \left\{ \sum_{r\neq 0} \beta_{j+nr} \right\} e^{2\pi i j t}. \tag{4.9}$$

Notons que la relation (4.8) nous fournit l'expression pour le biais de $\beta_{\lambda j}$ estimateur de β_j, c'est-à-dire

$$\begin{aligned} \text{Biais}\left(\ \beta_{\lambda j}\right) &= \mathbb{E}\left(\ \beta_{\lambda j}\right) - \beta_j \\ &= \sum_{r\neq 0} \beta_{j+nr}. \end{aligned}$$

Lorsque $n \to +\infty$, cette expression contient seulement les coefficients de *Fourier* d'indices très grands. Par conséquent, nous examinerons le biais de $\beta_{\lambda j}$ pour qu'il soit négligeable, quand n devient grand.

En substituant (4.9) dans (4.6) et en utilisant la relation (2.16) du chapitre 2 ainsi que les mêmes techniques pour obtenir (4.8), on obtient

$$R_n(\lambda) = \sum_{|j|>\lambda} \left|\ \beta_j\right|^2 + \frac{2\lambda+1}{n}\ \sigma^2 + \sum_{|j|>\lambda} \beta_j \sum \theta_{j+nr} \tag{4.10}$$

où

$$\theta_j = \begin{cases} \beta_j & \text{si } |j| > \lambda \\ \beta_j - \mathbb{E}\left(\ \beta_{\lambda j}\right) & \text{si } |j| \leq \lambda \end{cases},$$

et en supposant que $2\lambda + 1 < n$.

Cette expression montre comment $R_n(\lambda)$ dépend des coefficients de *Fourier* de μ.

En utilisant la relation de *Parseval,* on montre que $\sum\limits_{|j|>\lambda} \left|\ \beta_j\right|^2$ est la norme $L_2[0,1]$ de l'erreur, lorsqu'on approxime μ par la somme partielle de sa série de *Fourier* $\sum\limits_{j=-\lambda}^{\lambda} \beta_j e^{2\pi i j t}$.

Pour n suffisamment grand, le dernier terme dans la relation (4.10) est en général relativement négligeable par rapport aux deux autres termes.

Donc, asymptotiquement, une sélection optimale de λ correspond essentiellement à un équilibre entre l'erreur stochastique (ou le terme variance $\sigma^2(2\lambda+1)/n$), et l'erreur purement déterministe de l'approximation de μ par la somme partielle de sa série de *Fourier*.

Pour étudier le comportement asymptotique de $R_n(\lambda)$, nous sommes amenés à faire quelques hypothèses sur les coefficients de *Fourier* pour μ.

-D'après la relation de *Parseval*, il est évident que $\mid \beta_j \mid^2 \to 0$ quand $j \to \infty$.

-Cependant, plus d'hypothèses plus précises concernant la vitesse de décroissance des $\mid \beta_j \mid$ peuvent être faites pour des fonctions lisses.

-Par exemple, si $\mu \in W^2_{2,per}[0,1]$, sa seconde dérivée est dans $L_2[0,1]$ et l'intégration par parties montre que

$$\begin{aligned} \beta_j &= \int_0^1 \mu(t) e^{-2\pi i j t} dt \\ &= -\frac{1}{(2\pi j)^2} \int_0^1 \mu''(t) e^{-2\pi i j t} dt \\ &= -\frac{\beta_j''}{(2\pi j)^2}. \end{aligned}$$

Puisque

$$\| \mu'' \|^2 = \sum_{j=-\infty}^{\infty} \mid \beta_j'' \mid^2 = \sum_{j=-\infty}^{\infty} (2\pi j)^4 \mid \beta_j \mid^2 < \infty,$$

il s'ensuit que $\mid \beta_j \mid^2 = o\left(\frac{1}{j^4}\right)$.

Donc, on sait au moins que le carré du module des coefficients de *Fourier* pour toutes les fonctions dans $W^2_{2,per}[0,1]$ doivent décroitre plus vite que j^{-4}.

On peut par exemple supposer que

$$\left| \beta_j \right|^2 \sim C \mid j|^{-5-\delta} \tag{4.11}$$

où C et δ sont des constantes positives.

Si les (β_j) satisfont cette condition, il s'ensuit que $\| \mu" \| < \infty$.

En utilisant également les résultats des calculs concernant la continuité et la différentiabilité de la fonction limite d'une suite de fonction uniformément convergente (voir, par exemple, *Taylor* et *Mann* (1972), chapitre 20), on montre que la condition (4.11) suffit pour assurer que $\mu \in W^2_{2,per}[0,1]$.

Sous la condition (4.11), l'ordre de grandeur du biais de $\beta_{\lambda j}$ devient comme une constante multiple de

$$\sum_{r \neq 0} \mid j + nr|^{-(5+\delta)/2} .$$

Pour étudier cette somme, il suffit de trouver l'ordre de grandeur de $\sum_{r=1}^{\infty} (\ nr \pm j)^{-(5+\delta)/2}$.

On peut, par exemple utiliser l'approximation par une intégrale pour obtenir

$$\begin{aligned} \sum_{r=1}^{\infty} (\ nr \pm j) \ ^{-(5+\delta)/2} &\leq \int_1^{\infty} (nx \pm j) \ ^{-(5+\delta)/2} \ dx \\ &= [\ 2/(3+\delta)] \, n^{-1} (n \pm j)^{-(3+\delta)/2} \\ &= O \left(\ n^{-(5+\delta)/2} \right) \end{aligned} \tag{4.12}$$

uniformément ou $\mid j| \leq (n-1)/2$.

Un calcul simple utilisant (4.12) montre que le dernier terme dans (4.10) de $R_n(\lambda)$ est de l'ordre $o\left(n^{-(5+\delta)/2}\right)$.

En utilisant une approximation par une intégrale, et la relation (4.11), on peut montrer que lorsque $\lambda \to \infty$

$$\sum_{|j|>\lambda} \mid \beta_j \mid^2 \sim 2\ C_1\ \lambda^{-4-\delta} \tag{4.13}$$

pour $C_1 = C/(4+\delta)$.

En combinant les résultats précédents avec l'expression de $R_n(\lambda)$ dans la relation (4.10), on obtient une expression asymptotique pour $R_n(\lambda)$.

Si $\lambda \to \infty, n \to \infty$ avec $\lambda/n \to 0$, alors

$$R_n(\lambda) \sim 2\ \ C_1\ \lambda^{-4-\delta} + \frac{2\lambda}{n}\ \sigma^2. \tag{4.14}$$

Pour déterminer la valeur optimale asymptotique du paramètre λ, on peut chercher la valeur qui annule la dérivée de l'expression du second membre de la relation (4.14) pour obtenir

$$\lambda_{opt} \propto n^{1/(5+\delta)}. \tag{4.15}$$

En remplaçant la valeur optimale λ_{opt} de (4.15) dans l'expression de la relation (4.14), on obtient

$$R_n(\lambda_{opt}) = O\left(\ n^{-(4+\delta)/(5+\delta)}\right) \tag{4.16}$$

Remarque 7 *Les relations (4.15) et (4.16) nous permettent de tirer un certain nombre de conclusions importantes.*

1) μ_λ *est un estimateur consistant de* μ *pourvu que* $\lambda \to \infty$ *et* $\lambda/n \to 0$ *quand* $n \to \infty$.

2) *La valeur optimale du risque décroît avec une vitesse plus petite que* $1/n$.

3) *D'après la relation (*4.15*), le nombre de termes figurant dans la somme de l'estimateur par la série de Fourier augmente très lentement par rapport à la taille n de l'échantillon. Pour avoir une idée, voici quelques exemples de l'ordre de grandeur de la valeur optimale ($n^{1/5}$),* $100^{1/5} = 2.5$; $1000^{1/5} = 4; 10000^{1/5} = 6.3$ *et* $100000^{1/5} = 10$.

Il est évident que ça ne veut pas dire que si $n = 1000$*, on doit utiliser une approximation de la série par* 4 *termes, puisque dans ce cas, ça élimine la constante de proportionnalité dans (*4.15*) et la valeur de* δ *dans (*4.11*).*

4.3 Comportement Asymptotique de l'Estimateur de la Fenêtre par la Méthode GCV

Dans cette section nous allons examiner le cas d'une fonction de régression pouvant se décomposer une série de *Fourier* généralisée. On considère donc $\{x_j\}$ un système complet orthonormé $(CONS)$; toute fonction de $L_2[0,1]$, possède un développement en série de *Fourier* généralisé de la forme $\sum_{j=1}^{\infty} \beta_j x_j(t)$.

Supposons que $\mu \in L_2[0,1]$ et que $\sum_{j=1}^{\lambda} \beta_j x_j$ converge uniformément en t vers μ quand $\lambda \to \infty$.

C'est le cas (convergence uniforme), par exemple, lorsque $\mu \in C^1[0,1]$, $\mu(0) = \mu(1)$ et $\mu'(0) = \mu'(1)$.

Sous ces conditions, le modèle de régression s'écrit

$$y_j = \sum_{k=1}^{\infty} \beta_k \; x_k(t_j) + \varepsilon_j, j = 1, ..., n. \tag{4.17}$$

Donc les données suivant un modèle linéaire, un nombre infini de coefficients inconnus.

Puisque $\mu \in L_2[0,1]$, on sait que $\sum_{j=1}^{\infty}\beta_j^2 < \infty$, on doit donc avoir $\beta_j \to 0$. Il est donc raisonnable de supposer qu'il existe un entier λ tel que

$$\mu = \sum_{j=1}^{\lambda}\beta_j \; x_j$$

et ainsi

$$y_i = \sum_{j=1}^{\lambda}\beta_j \; x_j(t_i) + \varepsilon_i, i = 1, ..., n.$$

Ce qui nous amène à un modèle de régression linéaire. En prenant $t_1, ..., t_n$ uniformément réparti sur $[0,1]$, c'est-à-dire

$$t_j = \frac{j+1}{n-1}, \; j = 1, ..., n$$

et en posant

$$X_\lambda = \left(x_j(t_i)\right) \begin{array}{l} i = 1, ..., n \\ \\ j = 1, ..., \lambda \end{array}, \tag{4.18}$$

l'estimation par la méthode des moindres carrés fournit comme estimateur de $(\; \beta_{\lambda 1}, ..., \beta_{\lambda\lambda})'$, l'estimateur suivant

$$\underline{\beta}_\lambda = (\; \beta_{\lambda 1}, ..., \beta_{\lambda\lambda})' = (\; X_\lambda^* X_\lambda)^{-1} X_\lambda^* \; \underline{y}, \tag{4.19}$$

où

$$\underline{y} = (\; y_1, ..., y_n)'$$

est le vecteur des observations et la notation "*" désigne le transposer du complexe conjugué.

Ainsi l'estimateur obtenu de la série de *Fourier* généralisé de $\mu(t)$ est de la forme

$$\mu_\lambda(t) = \sum_{j=1}^{\lambda} \beta_{\lambda j}\ x_j(t), \tag{4.20}$$

Pour estimer la variance des erreurs, on peut utiliser

$$MSE(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left[y_j - \sum_{k=1}^{\lambda}\beta_{\lambda k}\ x_k\ (t_j)\right]^2$$

ou bien

$$\sigma^2_\lambda = \frac{1}{n-\lambda}\sum_{j=1}^{n}\left[y_j - \sum_{k=1}^{\lambda}\beta_{\lambda k}\ x_k\ (t_j)\right]^2.$$

A première vue, la stratégie utilisée est attirante, cependant, beaucoup de problèmes peuvent resurgir en utilisant cette approche, par exemple, on n'est pas sûr que l'estimateur construit de μ est consistant.

Il est difficile, dans ce cas, d'étudier la consistance sans faire d'hypothèses explicites sur les éléments du système complet orthonormé (x_j). Intuitivement, le comportement asymptotique dépend des coefficients de *Fourier* non estimés et des estimateurs $\beta_{\lambda j}$ de β_j.

Shibata (1981) donne quelques résultats concernant la consistance des estimateurs des moindres carrés dans le modèle (4.17). *Li* (1984) s'est intéressé à l'existence des estimateurs consistants de quelques fonctions des β_j. *Li* (1984) donne des conditions nécessaires et suffisantes de l'existence d'un estimateur consistant et construit un estimateur consistant. D'autres estimateurs consistants proches de (4.20) sont étudiés par *Rutkowski* (1982) et *Rafajlowicz* (1987).

Supposons que nous utilisons l'estimateur de la forme (4.20) où les $(\beta_{\lambda j})$ sont donnés par (4.19). Pour cela, il est nécessaire de sélectionner une valeur de λ. Le

problème du choix optimal de la valeur de λ est discuté dans le chapitre 3. Par exemple, pour une valeur fixée de n, une valeur optimale de λ est celle qui minimise $L_n(\lambda)$. Pour estimer la valeur minimisant $L_n(\lambda)$, on peut utiliser l'un des critères étudiés dans le chapitre 3. Par exemple, l'estimation par la méthode GCV de la valeur optimale de λ est $\hat{\lambda}_n$ obtenu en minimisant

$$GCV_n(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\frac{[\ y_j - \mu_\lambda(t_j)]^2}{\left(1-\frac{\lambda}{n}\right)^2} \tag{4.21}$$

(voir *Eubank* (1988), chapitre 3).

Sous certaines conditions, la méthode GCV de sélection de λ est asymptotiquement optimale comme le montre le théorème suivant dû à *Li* (1987).

Théorème 8 ***(Li (***1987***)***

Supposons que

i) *Les ε_i sont i.i.d. vérifiant $\mathbb{E}[\ \varepsilon_1^4] < \infty$ et il existe une constante K tel que pour tout $c \geq 0$,*

$$\sup_x P\ (\ x - c \leq \varepsilon_1 \leq x + c) \leq\ K\ c. \tag{4.22}$$

ii) $\inf_{\lambda \leq n} n\ R_n(\lambda) \to \infty$ *quand* $n \to \infty$,

et

iii) $\inf_{\lambda \leq n} L_n(\lambda) \xrightarrow{p} 0$.

*Alors, si $\hat{\lambda}_n$ est la valeur minimisant $GCV_n(\lambda)$ de (*4.21*), on a*

$$\frac{L_n(\ \hat{\lambda}_n)}{\inf\limits_{\lambda \leq n}\ L_n(\lambda)} \xrightarrow{p} 1, \quad \textit{quand } n \to \infty. \tag{4.23}$$

Remarque 9 1) *D'après le théorème précédent l'estimateur GCV, $\hat{\lambda}_n$ et la valeur optimale λ ont le même comportement en terme de fonction perte.*

2) *La condition i) porte sur les erreurs, par exemple si les (ε_j) ont une densité f_ε bornée, alors*

$$\begin{aligned} P(\ x-c\ \ &\leq\ \ \varepsilon \leq x+c) = \int_{x-c}^{x+c} f_\varepsilon(t)dt \\ &\leq\ \ 2\ c\left[\sup_t f_\varepsilon(t)\right], \end{aligned}$$

ce qui vérifie la condition (4.22).

3) *La condition ii) porte sur la vitesse de convergence de $R_n(\lambda)$ vers zéro et qui doit être plus petite que $1/n$.*

4) *La condition iii) exige simplement l'existence d'un estimateur consistant au sens que la fonction perte décroît vers zéro. Pour vérifier cette condition, il suffit de montrer qu'il existe une suite (λ_n) telle que $L_n(\lambda_n) \xrightarrow{p} 0$. Une condition plus forte entrainant la condition iii) est que $\inf_{\lambda \leq n} R_n(\lambda) \to 0$, quand $n \to \infty$. Cette dernière condition est en général plus facile à vérifier en pratique.*

4.4 Distribution Asymptotique

Dans cette section, nous allons étudier la distribution asymptotique de l'estimateur μ_λ de la fonction de régression μ, lorsque le système complet orthonormé $(CONS)$ est une famille exponentielle de la forme

$$x_j(t) = e^{2\pi ijt}, j \in \mathbb{Z}.$$

D'après l'étude faite dans le chapitre 2, l'estimateur μ_λ de μ est de la forme

$$\mu_\lambda(t) = \sum_{j=-\lambda}^{\lambda} \beta_{\lambda j}\ e^{2\pi i j t}$$

où

$$\underline{\beta}_\lambda = \big(\ \beta_{\lambda -\lambda}, ..., \beta_{\lambda\lambda}\big)' = (\ X_\lambda^* X_\lambda)^{-1}\, X_\lambda^*\ \underline{y},$$

avec

$$\underline{y} = (\ y_1, ..., y_n)'$$

et

$$X_\lambda = \left(e^{2\pi i j r / n}\right)_{\substack{r = 0, ..., n-1 \\ j = -\lambda, ..., \lambda}}.$$

On montre que l'estimateur μ_λ peut également se mettre sous la forme (voir chapitre 2)

$$\mu_\lambda(t) = \frac{1}{n}\sum_{j=1}^{n} y_j\ K_\lambda(\ t - t_j), \tag{4.24}$$

où

$$K_\lambda(t) = \frac{\sin[\ \pi\ (2\lambda + 1)\ t]}{\sin(\pi t)}. \tag{4.25}$$

D'après l'équation (4.24), $\mu_\lambda(t)$ est une somme pondérée de variables aléatoires non corrélées. On peut donc appliquer le théorème central limite pour étudier le comportement asymptotique de l'estimateur μ_λ.

On suppose que les (ε) sont *i.i.d.* de moments absolus d'ordre $(2 + \alpha)$ finis pour tout $\alpha > 0$.

L'outil dont on aura besoin pour déduire la distribution asymptotique de μ_λ est le théorème central limite de *Liapunov* (voir par exemple *Serfling* (1980) pour la démonstration) qui peut être formulé comme suit

Théorème 10 ***(Théorème de Liapunov)***

Soit $\{ V_{jn},\ j = 1, ..., n \text{ et } n = 1, 2, ...\}$ *un tableau triangulaire de variables aléatoires telles* $V_{1n}, ..., V_{nn}$ *soient indépendantes pour tout* $n = 1, 2, ...$

On pose

$$\sigma_n = \left[\sum_{j=1}^{n} \text{var } (V_{jn})\right]^{1/2} = \left\{\sum_{j=1}^{n} \mathbb{E}\ [\ V_{jn} - \mathbb{E}\ (V_{jn})]^2\right\}^{1/2}$$

et, pour $\theta > 2$*, on pose également*

$$C_n(\theta) = \sum_{j=1}^{n} \mathbb{E}\ \mid V_{jn} - \mathbb{E}\ (V_{jn})\mid^{\theta}.$$

Alors, si

$$\frac{C_n(\theta)}{\sigma_n^{\theta}} \to 0, \quad \textit{quand} \quad n \to \infty,$$

on a

$$\frac{1}{\sigma_n} \sum_{j=1}^{n} [\ V_{jn} - \mathbb{E}\ (V_{jn})]$$

tend en loi vers une loi normale centrée et réduite $\mathcal{N}(0, 1)$*, quand* $n \to \infty$*.*

Remarque 11 *L'avantage du théorème de Liapunov est les variables identiquement distribuées dans la somme* $\sum_{j=1}^{n} [\ V_{jn} - \mathbb{E}\,(V_{jn})]$ *n'est pas exigée.*

En effet, on montre que

$$\begin{aligned} \mu_\lambda(t) - \mathbb{E}\ [\ \mu_\lambda(t)] &= \frac{1}{n} \sum_{j=1}^{n} [\ y_j - \mathbb{E}\ (y_j)]\ K_\lambda(t - t_j) \\ &= \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j\ K_\lambda(t - t_j), \end{aligned}$$

où $K_\lambda(.)$ est donné par la relation (4.25).

Donc, $\mu_\lambda(t) - \mathbb{E}[\ \mu_\lambda(t)]$ est de la forme $\sum_{j=1}^{n} [\ V_{jn} - \mathbb{E}(\ V_{jn})]$ avec les V_{jn} indépendantes mais non identiquement distribués données par

$$V_{jn} = \frac{1}{n}\ \varepsilon_j\ \ K_\lambda(t - t_j) \tag{4.26}$$

Ce qui entraine que

$$\mathbb{E}\ (V_{jn}) = 0,\ \text{et}\ \mathbb{E}\ \left(V_{jn}^2\right) = \frac{\sigma^2}{n}\ \mid K_\lambda(t - t_j)\mid^2,$$

où $\sigma^2 = \text{var}(\varepsilon_j)$.

Ainsi

$$\begin{aligned}
\sigma_n^2 &= \sum_{j=1}^{n}\mathbb{E}\ \left(V_{jn}^2\right) = \frac{\sigma^2}{n^2}\sum_{j=1}^{n}\mid K_\lambda(t - t_j)\mid^2 \\
&= \frac{\sigma^2}{n^2}\sum_{j=1}^{n}\sum_{r=-\lambda}^{\lambda}\sum_{k=-\lambda}^{\lambda} e^{2\pi ir(t-t_j)}\ e^{-2\pi ik(t-t_j)} \\
&= \frac{2\lambda+1}{n}\ \sigma^2,
\end{aligned}$$

en utilisant le fait que

$$\frac{1}{n}\sum_{r=0}^{n-1} e^{2\pi ijr/n}\ e^{-2\pi ikr/n} = \begin{cases} 1 \text{ si } j = k \\ \\ 0 \text{ si } j \neq k \end{cases}.$$

De même, en utilisant la relation précédente, on montre que $\mid K_\lambda(u)\mid \leq (2\lambda + 1)$, ainsi

$$\begin{aligned}
\sum_{j=1}^{n}\mathbb{E}\ \mid V_{jn}\mid^{2+\alpha} &= \frac{1}{n^{2+\alpha}}\mathbb{E}\mid \varepsilon_1\mid^{2+\alpha}\sum_{j=1}^{n}\mid K_\lambda(\ t - t_j)\mid^{2+\alpha} \\
&< O\ \left[\left(\frac{\lambda}{n}\right)^{1+\alpha}\right].
\end{aligned}$$

En posant $\theta = 2 + \alpha$, comme application du théorème de *Liapunov*, alors on a $\sum_{j=1}^{n} V_{jn} \diagup \sigma_n$ converge en loi vers une loi normale standard $\mathcal{N}(0,1)$, pourvu que

$$\sum_{j=1}^{n} \frac{\mathbb{E} \mid V_{jn} \mid^{2+\alpha}}{\left(\frac{2\lambda+1}{n} \sigma^2 \right)^{1+\frac{\alpha}{2}}} = O\left[\left(\frac{\lambda}{n} \right)^{\alpha \diagup 2} \right]$$

converge vers zéro, quand $n \to \infty$.

Ce résultat peut être résumé par le théorème suivant (voir *Eubank* (1988)).

Théorème 12 *Supposons que $\lambda \to \infty$ quand $n \to \infty$, tel que $\lambda \diagup n \to 0$. Alors, pour tout t fixé, on a $\frac{n^{1 \diagup 2}}{\sigma(\, 2\lambda+1)^{1 \diagup 2}} \{\, \mu_\lambda(t) - \mathbb{E}[\, \mu_\lambda(t)]\}$ converge en loi vers une loi normale standard $\mathcal{N}(0,1)$.*

En réalité, on cherche à établir la loi asymptotique de $\frac{n^{1 \diagup 2}}{\sigma(\, 2\lambda+1)^{1 \diagup 2}} [\, \mu_\lambda(t) - \mu(t)]$, ce qui peut être fait sous certaines restrictions.

On a

$$\begin{aligned} \sqrt{n}\, \frac{\mu_\lambda(t) - \mu(t)}{\sigma\sqrt{\, 2\lambda+1}} &= \sqrt{n}\, \frac{\mu_\lambda(t) - \mathbb{E}\ [\, \mu_\lambda(t)]}{\sigma\sqrt{\, 2\lambda+1}} \\ &\quad + \sqrt{n}\, \frac{\mathbb{E}\ [\mu_\lambda(t)] - \mu(t)}{\sigma\sqrt{2\lambda+1}}. \end{aligned} \tag{4.27}$$

Le premier terme du second membre tend en loi vers une loi normale standard, d'après le théorème précédent. Donc, si le second terme tend vers zéro, d'après le théorème de *Slutsky* (voir *Serfling* (1980)), on a $\frac{n^{1 \diagup 2}}{\sigma(\, 2\lambda+1)^{1 \diagup 2}} [\, \mu_\lambda(t) - \mu(t)]$ qui tend en loi vers une variable aléatoire normale standard $\mathcal{N}(0,1)$.

Comme on a vu, d'après (4.8), on a

$$\begin{aligned} \mid \mu(t) - \mathbb{E}\ [\ \mu_\lambda(t)]\mid &= \left| \sum_{|j|>\lambda} \beta_j e^{2\pi i j t} - \sum_{|j|\leq\lambda} \left(\sum_{r\neq 0} \beta_{j+nr} \right) e^{2\pi i j t} \right| \\ &\leq \sum_{|j|>\lambda} \mid \beta_j \mid + \sum_{|j|\leq\lambda} \left| \sum_{r\neq 0} \beta_{j+nr} \right| \\ &= O\left(\frac{1}{\lambda^{(3+\delta)/2}} \right) + O\left(\frac{\lambda}{n^{(5+\delta)/2}} \right) \end{aligned} \tag{4.28}$$

pourvu que $\mid \beta_j \mid^2 \sim c \mid j \mid^{-5-\delta}$.

Si $\frac{\lambda}{n} \to 0$, quand $n \to \infty$, alors c'est le premier terme qui domine et donc

$$\sqrt{n}\ \frac{\mid \mu(t) - \mathbb{E}\ [\ \mu_\lambda(t)]\mid}{\sqrt{\ 2\lambda+1}}\ \leq\ \sqrt{n}\ O\ \left(\frac{1}{\lambda^{(4+\delta)/2}} \right), \tag{4.29}$$

uniformément en t. Ce qui donne le résultat suivant énoncé sous forme de corollaire.

Corollaire 13 *Supposons que $\mu \in W^2_{2,per}[0,1]$, avec les coefficients de Fourier vérifiant la condition $\mid \beta_j \mid^2 \sim c \mid j \mid^{-5-\delta}$, avec δ une constante positive. Alors, pour t fixé, $\sqrt{n}\ \frac{\mu_\lambda(t)\ -\mu(t)}{\sigma\sqrt{\ 2\lambda+1}}$ tend en loi vers une variable aléatoire normale standard $\mathcal{N}(0,1)$, pourvu que n, $\lambda \to \infty$ tels que $n\ \lambda^{-(4-\delta)} \to 0$.*

Remarque 14 *Le corollaire précédent permet de construire un intervalle de confiance pour $\mu(t)$ avec un coefficient de confiance $p = 1-\alpha$, donné par*

$$\left[\mu_\lambda(t) - z_{\alpha/2}\sigma\sqrt{\ \frac{2\lambda+1}{n}}, \mu_\lambda(t) + z_{\alpha/2}\sigma\sqrt{\ \frac{2\lambda+1}{n}} \right],$$

avec $z_{\alpha/2}$ vérifiant $P(\ Z > z_{\alpha/2}) = \frac{\alpha}{2}$ et $Z \sim \mathcal{N}(0,1)$.

Chapitre 5

Construction et Mise en Œuvre du Test

5.1 Loi de probabilité des estimateurs

Soient $(t_j, y_j),\ j = 1, ..., n$ les n observations des deux variables indépendantes T et Y, tels que t_j et y_j sont liés par le modèle

$$y_j = \mu(t_j) + \varepsilon_j, j = 1, ..., n\ ,$$

où μ est une fonction inconnue et (ε_j) est une suite de variable aléatoire, non covariées, de même loi de probabilité, telles que

$$\mathbb{E}(\varepsilon_j) = 0, var(\varepsilon_j) = \sigma^2, j = 1, .., n,$$

On suppose aussi que les $y_j,\ j = 1, .., n$ sont obtenues comme des réponses de p variables d'entrées $X_1, ..., X_p,\ X' = [\ \underline{X}_1, ..., \underline{X}_n]$ à un rang égal à p, et β est le vecteur de dimension p de paramètres inconnus ($n >> p$).

Théorème 15 *Sous l'hypothèse* $Y \sim \mathcal{N}(\ X\beta, \sigma^2 I)$, (rang$(X) = p$,, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$,

où les ε_i sont indépendantes.). Soit $\hat{\beta}$ et $\hat{\sigma}$ les estimateurs de β et σ (respectivement) avec la méthode des moindres carrés, alors

- $\hat{\beta} \sim \mathcal{N}(\ \beta, \sigma^2(\ X'X)^{-1})$
- $(n-p)\ \hat{\sigma}^2 / \sigma^2 \sim \chi^2(n-p)$.
- $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

Preuve. Voir "*Régression Théorie et application*", de *Pierre-André Cornillon, Eric Matzner-Lober*, page 55. ■

5.2 Intervalle et Région de Confiance

Rappelons que la loi de *Student* de *d.d.l.* k est celle de $X / \sqrt{Y/k}$ où X est une variable gaussienne centrée réduite et Y suit une loi de $\chi^2(k)$ indépendante de X.

La loi de *Fisher-Snedecor* de (k, l) *d.d.l.* est celle de $(X/k)/(Y/l)$ où $X \sim \chi^2(k)$ et $Y \sim \chi^2(l)$ sont indépendantes.

On désigne par $t_k(.)$ et $f_{kl}(.)$ les fonctions quantiles de ces lois de probabilité. Comme conséquence immédiate du théorème précédent, on a les propriétés suivantes

Proposition 16 *Sous l'hypothèse $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$:*

- *Pour tout $i = 1, ...p$, la variable aléatoire*

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\ (\ \hat{\beta}_i)}$$

suit une loi de Student de paramètre $(n-p)$, $\left[\ \hat{\sigma}(\hat{\beta}_i)\right]$ est l'erreur standard de $\hat{\beta}_i$.

- *Pour tout vecteur u, la variable aléatoire*

$$T_u = \frac{u'\hat{\beta} - u'\beta}{\hat{\sigma}\ (\ u'\hat{\beta})},$$

suit une loi de Student de paramètre $(n-p)$*, où*

$$\hat{\sigma}(\ u^{'}\hat{\beta}\ )^2 = \hat{\sigma}^2 u^{'}(\ X^{'}X\ )^{-1}u.$$

- *Soit* $(q < p)$ *et* L *une matrice* $(q \times\ p)$ *de rang* q*, la variable aléatoire*

$$F = \frac{1}{q\ \hat{\sigma}^2}(\ \hat{\beta} - \beta)^{'} L^{'} \left[L(\ X^{'}X)^{-1}L^{'}\right]^{-1} L\ (\ \hat{\beta} - \beta)$$

suit une loi de Fisher-Snedecor de d.d.l. $(q, n-p)$

Pour le troisième point, notons que la variable $\left[L(\ X^{'}X)^{-1}L^{'}\right]^{-1/2} L\ (\ \hat{\beta} - \beta)$ suit la loi $\mathcal{N}(0, \sigma^2 Id_q)$, ce qui fait que le numérateur suit la loi $\sigma^2\chi^2(q)$.

5.2.1 Intervalle de confiance :

- En raison de la symétrie de la loi de *Student*, on a $P\left[\ |\ T_i| < t_{n-p}\left(\frac{1-\alpha}{2}\right)\right] = 1-\alpha.$

On obtient donc un intervalle de confiance de coefficient de confiance $(1-\alpha)$

$$[\ \hat{\beta}_i - \delta, \hat{\beta}_i + \delta],$$

où

$$\delta = \hat{\sigma}(\ \hat{\beta}_i)\ t_{n-p}\left(\frac{1-\alpha}{2}\right) = t_{n-p}\left(\frac{1-\alpha}{2}\right)\ \hat{\sigma}\sqrt{[\ (\ X'X)^{-1}]_{ii}}.$$

- Un intervalle de confiance, de niveau $(1-\alpha)$, pour σ^2 est donné par

$$\left[\frac{(n-p)\ \hat{\sigma}^2}{c_2}, \frac{(n-p)\ \hat{\sigma}^2}{c1}\right]$$

où $P(\ c_1 \leq \boldsymbol{\chi}^2(n-p) \leq c_2) = 1-\alpha$ et c_1, c_2 sont les fractiles d'un $\boldsymbol{\chi}^2(q)$ et $f_{q,n-p}\,(1-\alpha)$ est le fractile de niveau $(1-\alpha)$ d'une loi de *Fisher* admettant $(q, n-p)$ *d.d.l.*

5.2.2 Région de confiance :

On note $\|x\|_S = x^{'} S\ x$. On peut écrire la relation $P[\ F < f_{q,n-p}(1-\alpha)] = 1-\alpha$, $(q \leq p)$, sous la forme

$$P(\ L\beta \in \mathcal{R}_\alpha) = 1-\alpha$$

où $\mathcal{R}_\alpha$ donné par

- lorsque σ est connue,

$$\mathcal{R}_\alpha = \left\{ \zeta \in \mathbb{R}^q, \left\|\ L\hat{\beta} - \zeta \right\|^2_{[\ L(\ X'X)^{-1}L']^{-1}} \leq \sigma^2 \boldsymbol{\chi}^2_q\ (1-\alpha) \right\}$$

- lorsque σ est inconnue,

$$\mathcal{R}_\alpha = \left\{ \zeta \in \mathbb{R}^q : \left\|\ L\hat{\beta} - \zeta \right\|^2_{[\ L(\ X'X)^{-1}L']^{-1}} \leq q\ \hat{\sigma}^2 f_{q,n-p}(1-\alpha) \right\}$$

qui est donc une région de confiance de probabilité $(1-\alpha)$ pour le vecteur $L\beta$.

Si L est la matrice de sélection $(q \times\ p)$ telle que $L\beta = (\beta_{i1}, ...\beta_{iq})$, on obtient une région de confiance pour $(\beta_{i1}, ...\beta_{iq})$.

5.3 Test de Fisher de Signification d'une variable explicative

On veut tester $H_0 : L\beta = l$ où $l = 0$ et L est ici une matrice d'identité.

Soit $L \in \mathbb{R}^{q\times p}$, $l \in \mathbb{R}^q$, pour $l \in \mathcal{R}_\alpha$,on a

$$\left\|\ L\hat{\beta} - l \right\|^2_{[\ L(\ X^TX)^{-1}L^T]^{-1}} \leq q\ \hat{\sigma}^2 f_{q,n-p}(1-\alpha)$$

Au niveau α donné, on accepte H_0 si :

$$\frac{(\ L\hat{\beta} - l)'\ \left[\ L(\ X'X)^{-1}L\right]^{-1} (\ L\hat{\beta} - l)}{q\ \hat{\sigma}^2} > f_{q,n-p}(1-\alpha)$$

avec $f_{q,n-p}(1-\alpha)$ le quantile d'ordre $(1-\alpha)$ de la loi de $\mathcal{F}_{q,n-p}$.

5.4 Mise en Œuvre du Test d'Hypothèse

Supposons que $\underline{\varepsilon}$ est un vecteur de n variables aléatoires normales indépendantes et nous souhaitons tester une hypothèse linéaire générale de la forme $H_0 : L\underline{\beta} = \underline{\ell}$, où L est la matrice $(q \times p)$ de rang $q \leq p$ et $\underline{\ell}$ est un vecteur donné. Alors, sous l'hypothèse H_0, la statistique

$$\frac{\left(L\underline{\hat{\beta}} - \underline{\ell} \right)' \left(L(X'X)^{-1}L' \right)^{-1} \left(L\underline{\hat{\beta}} - \underline{\ell} \right)}{q\ \hat{\sigma}^2} \tag{5.1}$$

admet une loi de *Fisher* avec $(q, n-p)$ *d.d.l.* Cette statistique peut donc être utilisée pour tester la nature générale de $\underline{\beta}$. En choisissant convenablement L, (5.1) peut être utilisé pour obtenir des statistiques de test pour une variété d'hypothèses. Par exemple, en choisissant $q = 1$, L le vecteur ligne de tous les zéros à l'exception de la $i^{\text{ème}}$ composante et $\ell = \beta_{i0}$, on obtient la statistique

$$\frac{\left(\beta_{\lambda i} - \beta_{i0} \right)^2}{\sigma_\lambda^2 c_{\lambda,ii}}$$

où $c_{\lambda,ii}$ est le $i^{\text{ème}}$ élément diagonale de $c_\lambda = (X_\lambda^{'}\ X_\lambda)^{-1}$, pour tester l'hypothèse $H_0 : \beta_i = \beta_{i0}$.

Sous l'hypothèse H_0, cette statistique suit une loi de *Fisher* avec un $(1, n-p)$ *d.d.l.* Elle est, donc, équivalente à la statistique

$$Z = \frac{\left(\beta_{\lambda i} - \beta_{i0} \right)}{\left[\sigma_\lambda^2\ c_{\lambda,ii} \right]^{1/2}}$$

qui admet une loi de *Student*.

Par conséquent, l'hypothèse $H_0 : \beta_i = \beta_{i0}$ sera rejetée, si $\mid Z \mid \geq t_{n-p;\alpha/2}$, le $(\alpha/2)$ pourcentage des points supérieurs de la loi de *Student* avec $(n-p)$ *d.d.l.*

Si $\beta_i = 0$, on n'a pas besoin de la fonction x_i pour l'approximation de μ. Donc, pour vérifier si x_i est utilisé dans l'ajustement des données, on peut examiner la statistique.

$$z = \frac{\beta_{\lambda i}}{[\sigma_\lambda^2 \; c_{\lambda,\, ii}]^{1/2}}$$

Cette statistique est utilisée pour tester l'hypothèse sur les paramètres d'un modèle linéaire.

En prenant, $q = 1$, $L = (x_1(t), ..., x_p(t))$ et $\ell = \mu_0$, (5.1) peut être utilisé pour obtenir un test d'hypothèse $H_0 : \mu(t) = \mu_0$, où μ_0 est une constante spécifique. La statistique résultante est

$$Z = \frac{[\; \mu_\lambda(t) - \mu_0]}{[\; \sigma_\lambda^2 \; \underline{x}_\lambda(t)' c_\lambda \; \underline{x}_\lambda(t)]^{1/2}}$$

L'hypothèse est rejeté si $\mid Z \mid \geq t_{n-p;\alpha/2}$.

En inversant le test d'hypothèse statistique sur β_i et $\mu(t)$ donnés ci-dessus, l'intervalle de confiance de β_i et $\mu(t)$ peut être obtenu. Spécifiquement, nous trouvons que les intervalles de confiance de $(1-\alpha)$ sont

$$\beta_{\lambda i} \pm Z_{\alpha/2} \left[\; \sigma_\lambda^2 c_{\lambda,ii}\right]^{1/2}$$

et

$$\mu_\lambda(t) \pm z_{\alpha/2} \left[\sigma_\lambda^2 \; \underline{x}_\lambda(t)^{'} c_\lambda \; \underline{x}_\lambda(t)\right]^{1/2} \tag{5.2}$$

avec $\underline{x}_\lambda(t) = (x_1(t), ..., x_\lambda(t))'$ et $z_{\alpha/2}$ le $(\alpha/2)$ pourcentage des points supérieurs

de la loi normale standard. Dans le cas particulier où $t = t_i$, (5.2) est réduite à

$$\mu_\lambda(t_i) \pm Z_{\alpha/2} \left[\sigma_\lambda^2 \, h_{ii}(\lambda)\right]^{1/2}$$

où $h_{ii}(\lambda)$ est le $i^{\text{ème}}$ élément diagonale de $H(\lambda)$.

Chapitre 6

La Sélection de Modèles

6.1 Introduction

On considère le modèle de régression

$$y_j = \mu(t_j) + \varepsilon_j, j = 1, ..., n \ , \tag{6.1}$$

où μ est une fonction inconnue et (ε_j) est une suite de variables aléatoires, non covariées, de même loi de probabilité, telles que

$$\mathbb{E}(\varepsilon_j) = 0, var(\varepsilon_j) = \sigma^2, j = 1, .., n, \tag{6.2}$$

On suppose aussi que les y_j, $j = 1, .., n$ sont obtenues comme des réponses de p variables d'entrées $X_1, ..., X_p$. Donc, y_j est la sortie obtenue par le vecteur d'entrées $\underline{X}_j = (X_{j_1}, ..., X_{jp})'$, $j = 1, ..., n$, et $X' = [\ \underline{X}_1, ..., \underline{X}_n]$ à un rang égal à p, et β est le vecteur de dimension p de paramètres inconnus ($n >> p$), tel que

$$y = X\beta + \varepsilon \tag{6.3}$$

où $y = (\ y_1, ..., y_n)'$, $\varepsilon = (\ \varepsilon_1, ..., \varepsilon_n)'$.

Pour choisir un meilleur modèle de régression parmi tous les modèles possibles, il faut définir un critère quantifiant la qualité du modèle. Une fois que ce critère est choisi, il faudra déterminer des procédures permettant de trouver le meilleur modèle.

Le critère le plus utiliser dans la sélection de modèles est l'ajusté R^2 (R^2_{adj}), telle que

$$R^2_{adj} = 1 - \frac{s^2_k}{\sum_{i=1}^{n} \left[\frac{(\ y_i - E(y))^2}{n-k} \right]}$$

où

$$s^2_k = \frac{SSE_k}{n-k} = \frac{1}{n-k} \sum_{j=1}^{n} (\ y_j - \hat{y}_j)^2 \,. \tag{6.4}$$

s^2_k est l'estimateur sans-biais de σ^2, SSE_k est la somme des erreurs quadratiques, $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)' = X\ \hat{\beta}$ et k le nombre de variables X_i dans le modèle candidat (voir *Rawlings*, 1988, p. 183).

D'autres travaux pour le choix du modèle sont apparus vers la fin des années 60 et au début des années 70. Les plus fréquents sont le FPE d'*Akaike* (*Akaike*, 1969), le C_p de *Mallows* (*Mallows*, 1973), et le critère de l'information d'*Akaike, AIC*, (*Akaike*, 1973, 1974), basé sur l'information de *Kullback-Leibler* (KL). Vers la fin des années 70, il y avait beaucoup de travaux dans le secteur de la théorie de l'information, où le critère de l'information bayésienne (BIC, *Akaike*, 1978), le critère de l'information de *Schwartz* (SIC, *Schwartz*, 1978), le critère de *Hannan* et de *Quinn* (HQ, *Hannan* et *Quinn*, 1979), le FPE_α (*Bhansali* et *Downham*, 1977), et le GM (*Geweke* et *Meese*, 1981) étaient exposés. Ultérieurement, vers la fin des années 80, *Hurvich* et *Tsai* (1989) ont adapté les résultats de *Sugiura* 1978 pour développer un petit-ensemble

d'estimateurs sans-biais et amélioré par la divergence de *Kullback-Leibler*, ($AICc$).

Dans de grands échantillons, un critère de la sélection de modèles qui choisit le modèle avec la répartition de l'erreur moyenne quadratique minimal serait asymptotiquement efficace (*Shibata*, 1980). FPE, AIC, $AICc$, C_p sont tous asymptotiquement efficaces. La version corrigée d'AIC la plus connue est $AICc$ (*Sugiura*, 1978 et *Hurvich* et *Tsai*, 1989). Il est important de se rappeler que tous les critères de la sélection de modèles sont eux-mêmes des variables aléatoires avec leurs propres répartitions.

Dans ce chapitre nous dérivons les moments et les probabilités pour l'AIC et ces variantes corrigés et on les lie à la performance via le concept du rapport signal-bruit. Les moments de plusieurs critères de la sélection classiques ont été étudiés par *Nishii*, 1984 et *Akaike*, 1969.

6.2 La Description du Modèle

Nous définissons le véritable modèle de régression par

$$y = \mu_* + \varepsilon_* \tag{6.5}$$

où

$$\varepsilon_* \sim \mathcal{N}(0, \sigma_*^2 I) \tag{6.6}$$

et $y = (y_1, \ldots, y_n)'$ est le vecteur $(n \times 1)$ des réponses, $\mu_* = (\mu_{*1}, ..., \mu_{*n})'$ sont les vraies valeurs de la fonction inconnue, et $\varepsilon_* = (\varepsilon_{*1}, ..., \varepsilon_{*n})'$.

Dans l'équation (6.4), nous supposons que les erreurs ε_{*i}, $i = 1, \ldots, n$ sont indépendantes et identiquement réparties, de loi $\mathcal{N}(0, \sigma_*^2)$.

Pour sélectionner un meilleur modèle, nous devons mesurer la performance du modèle en comparant les probabilités de la sélection du bon modèle pour chaque critère considéré. Pour cela, nous considérons la quantité d'information *Kullback-Leibler*, KL, (*Kullback* et *Leibler*, 1951), où elle est basée sur le rapport des vraisemblances.

Soit la norme L_2 donnée par

$$L_2 = \frac{1}{n} \parallel \mu_* - \hat{\mu}_* \parallel^2 . \tag{6.7}$$

où $\hat{\mu}_*$ est l'estimateur de la fonction μ_*.

La norme pourrait aussi être employée comme une base pour la mesure de distance. L'avantage de L_2 est qu'elle dépend seulement des moyennes des deux répartitions et non pas des densités réelles. La mesure de distance L_2, et l'information de *Kullback-Leibler* (KL) fournissent deux méthodes pour évaluer à quel point le modèle candidat approche le vrai modèle en estimant la différence entre les espérances sous le vrai modèle et le modèle candidat. Donc un critère qui est efficace dans le sens de L_2 est aussi efficace dans le sens de KL.

Pour définir l'information de *Kullback-Leibler*, nous considérons les fonctions de densité pour le vrai modèle, et le modèle candidat. Ces fonctions de vraisemblance jouent un rôle principal dans la dérivation des critères AIC et $AICc$.

On définit le KL par

$$KL = \frac{2}{n} \mathbb{E}_* \left[\log \left(\frac{f_*}{f} \right) \right] \tag{6.8}$$

où f_* et $\mathbb{E}_*$ dénotent respectivement la densité et l'espérance sous le vrai modèle et f la fonction de vraisemblance du modèle candidat. Nous avons ajusté la quantité d'information de *Kullback-Leibler* par $(2/n)$, pour l'exprimer comme un taux ou une

moyenne d'information par observation.

L'estimateur du maximum de vraisemblance de σ^2est donné par

$$\hat{\sigma}_k^2 = \frac{SSE_k}{n} = \frac{1}{n}\sum_{j=1}^{n}\left(\ y_j - \hat{y}_j\right)^2. \tag{6.9}$$

où SSE_k est la somme des erreurs quadratiques, $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)' = X\ \hat{\beta}$ et k le nombre de variables X_i dans le modèle candidat.

En introduisant la fonction log sur f et f_*, (avec l'hypothèse de la normalité des deux fonctions) et en les remplaçant dans la formule de KL, ainsi que σ^2 par $\hat{\sigma}_k^2$ donné par (6.9), et l'utilisation de la distance L_2 dans l'équation (6.7), l'information de *Kullback-Leibler* entre le modèle candidat adapté et le vrai modèle est

$$KL = \log\left(\frac{\hat{\sigma}_k^2}{\sigma_*^2}\right) + \frac{\sigma_*^2}{\hat{\sigma}_k^2} + \frac{L_2}{\hat{\sigma}_k^2} - 1 \tag{6.10}$$

6.3 Formulations du Critère de la Sélection de Modèle

Nous étudions maintenant les critères AIC et $AICc$ qui estiment l'information de *Kullback-Leibler*. AIC (*Akaike*, 1973) est asymptotiquement sans biais pour KL. Dans sa formulation *Akaike* a utilisé le fait que le vrai modèle appartient à l'ensemble des modèles candidats. En général

$$AIC = -2\ \log(\text{ probabilité}) + 2 \times \text{nombre de paramètres} \tag{6.11}$$

où la probabilité est habituellement évaluée par les paramètres estimés. La dérivation d'AIC est prévue pour créer une estimation qui est une approximation de

la quantité d'information de *Kullback-Leibler* (une dérivation plus détaillée peut être trouvée dans *Linhart* et *Zucchini*, 1986, p. 243).

En utilisant des estimateurs de maximum de vraisemblance sous la supposition d'erreurs normales et le nombre de paramètres est k pour le β et 1 pour σ^2

$$AIC = n\ \log(2\pi) + n\ \log(\hat{\sigma}_k^2) + n + 2(k+1). \tag{6.12}$$

Les constantes $(n\log(2\pi) + n)$ ne jouent aucun rôle pratique dans le choix du modèle et peut être ignoré. Nous ajustons AIC par un facteur $(1/n)$ pour l'exprimer comme un taux

$$AIC_k = \log(\hat{\sigma}_k^2) + \frac{2(k+1)}{n} \tag{6.13}$$

Hurvich et *Tsai* ont également adopté la supposition que le vrai modèle appartient à l'ensemble de modèles candidats. Ils prennent $\mathbb{E}_*[\ \hat{\sigma}_k^2] = (n-k)\ \sigma_*^2/n$, et $\mathbb{E}_*[\ 1/\hat{\sigma}_k^2] = n/\{(n-k-2)\sigma_*^2\}$ pour avoir

$$\begin{aligned}
\mathbb{E}_*[\ KL] &= \mathbb{E}_*\left[\ \log\left(\hat{\sigma}_k^2\right)\right] - \log\left(\sigma_*^2\right) + \frac{n\ \sigma_*^2}{(n-k-2)\ \sigma_*^2} + \frac{k\ \sigma_*^2}{(n-k-2)\ \sigma_*^2} - 1 \\
&= \mathbb{E}_*\left[\ \log\left(\hat{\sigma}_k^2\right)\right] - \log\left(\sigma_*^2\right) + \frac{n+k}{n-k-2} - 1
\end{aligned}$$

Notant que $\log\left(\hat{\sigma}_k^2\right)$ est sans biais pour $\mathbb{E}_*\left[\log\left(\hat{\sigma}_k^2\right)\right]$, alors $\left[\ \log\left(\hat{\sigma}_k^2\right) + \frac{n+k}{n-k-2} - \log\left(\sigma_*^2\right) - 1\right]$ est sans biais pour $\mathbb{E}_*[KL]$. La constante $(-\log\left(\sigma_*^2\right) - 1)$ ne marque aucune contribution pour la sélection du modèle et peut être ignorée. On a

$$AICc_k = \log\left(\hat{\sigma}_k^2\right) + \frac{n+k}{n-k-2} \tag{6.14}$$

Hurvich et *Tsai* (1989) ont prouvé que $AICc$ est plus performant que AIC dans les petits échantillons, mais qu'il est asymptotiquement équivalent à AIC et admettent la

même performance dans les grands échantillons. *Shibata* (1981) a prouvé que AIC et FPE sont asymptotiquement des critères efficaces, et d'autres auteurs (par exemple, *Nishii*, 1984) ont prouvé que AIC, FPE, et C_p sont asymptotiquement équivalents, ce qui implique que $AICc$ et C_p sont aussi asymptotiquement efficaces.

6.4 Les Moments des Critères de la Sélection de Modèles

6.4.1 Les Rapports Signal-Bruit des Critères

Quand on choisit parmi les modèles candidats, la méthode standard est que le meilleur modèle est celui pour lequel les valeurs du critère de la sélection de modèle utilisé atteignent leurs minimums, et les modèles sont comparés en prenant la différence entre les valeurs du critère pour chaque modèle.

Supposons que nous avons un modèle avec k variables, et un deuxième modèle avec L variables additionnelles. Nous regarderons les critères AIC et $AICc$, qui estiment l'information de *Kullback-Leibler*.

Pour AIC, nous choisissons le modèle avec k variables au lieu le modèle avec $k+L$ variables si $AIC_{k+L} > AIC_k$.

En appliquant

$$\mathbb{E}\,[\ \log(SSE_k)] = \log(\sigma_*^2) + \log(n-k) - \frac{1}{n-k} \tag{6.15}$$

où $SSE_k = \sum_{j=1}^{n} (\ y_j - \hat{y}_j)^2 \sim \sigma_*^2\ \boldsymbol{\chi}^2\ (n-k)\ ,\ \ \hat{y} = (\hat{y}_1, ..., \hat{y}_n)'$, et $\boldsymbol{\chi}^2(n-k)$ est la loi du *Chi-deux* à $(n-k)$ degré de liberté.

On suppose $\triangle AIC = AIC_{k+L} - AIC_k$, le signal est

$$\mathbb{E}[\ \triangle AIC] = \log\left(\frac{n-k-L}{n-k}\right) - \frac{L}{(n-k-L)(n-k)} + \frac{2L}{n}. \tag{6.16}$$

Puisque

$$\log\left(\frac{SSE_{k+L}}{SSE_k}\right) \sim \log\text{-Beta}\left(\frac{n-k-L}{2}, \frac{L}{2}\right), \tag{6.17}$$

le bruit est donné par

$$\boldsymbol{sd}\ [\ \triangle AIC] = \boldsymbol{sd}\ [\ \triangle \log SSE_k] \doteq \frac{\sqrt{2L}}{\sqrt{(n-k-L)(n-k+2)}},$$

et le rapport signal-bruit est

$$\begin{aligned}\frac{\mathbb{E}\ [\ \triangle AIC]}{\boldsymbol{sd}\ [\ \triangle AIC]} &= \frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}} \\ &\quad \times \left[\log\left(\frac{n-k-L}{n-k}\right) - \frac{L}{(n-k-L)(n-k)} + \frac{2L}{n}\right]. \end{aligned} \tag{6.18}$$

Quand L tend vers $n-k$, on a

$$\frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}} \log\left(\frac{n-k-L}{n-k}\right) \to 0$$

et

$$\frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}}\ \frac{L}{(n-k-L)(n-k)} \to -\infty$$

Les comportements du premier et du deuxième terme sont obtenus par l'utilisation du $\log(\hat{\sigma}_k^2)$ où $\log(SSE_k/n)$. Le dernier terme,

$$\frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}}\ \frac{2L}{n} \to 0.$$

Ce troisième terme augmente pour un petit L, alors il décroit vers 0 lorsque L croit. Ainsi, lorsque L devient plus grand, le rapport signal-bruit d'AIC devient faible, et

un rapport signal-bruit négatif indésirable résulte. Typiquement, le rapport signal-bruit d'AIC croit pour un petit L, mais comme $L \to n-k$, le rapport signal-bruit de $AIC \to -\infty$.

Quand L et k sont fixés et $n \to \infty$, on a $\log(1-L/n) \simeq -L/n$ quand $L << n$, alors le rapport signal-bruit asymptotique pour AIC est

$$\lim_{n\to\infty} \frac{\mathbb{E}\ [\ \triangle AIC]}{\boldsymbol{sd}\ [\ \triangle AIC]} = \lim_{n\to\infty} \frac{n}{\sqrt{2L}} \left[\log\left(1-\frac{L}{n}\right) - \frac{L}{n^2} + \frac{2L}{n}\right] = \sqrt{\frac{L}{2}} \tag{6.19}$$

De l'équation (6.17), le rapport signale-bruit pour $AICc$ est

$$\begin{aligned} \frac{\mathbb{E}\ [\ \triangle AICc]}{\boldsymbol{sd}\ [\ \triangle AICc]} &= \frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}} \times \\ &\quad \left[\log\left(\frac{n-k-L}{n-k}\right) - \frac{L}{(n-k-L)(n-k)} + \frac{2L(n-1)}{(n-k-2)(n-k-L-2)}\right] \end{aligned} \tag{6.20}$$

qui croit quand L croit.

Le rapport signal-bruit asymptotique pour $AICc$ est $\sqrt{L/2}$.

Pour $AIC = \log(\hat{\sigma}_k^2) + 2(k+1)/n$, le rapport signal-bruit associé a $\log(\hat{\sigma}_k^2) \to -\infty$ quand L augmente. L'exécution des critères de la sélection de modèles avec des rapports signal-bruit faibles pourrait être améliorée si leurs rapports signal-bruit étaient renforcés. Pour cette raison, nous présenterons la variante corrigée signal-bruit $AICu$ (*McQuarrie*, *Shumway* et *Tsai*, 1997),

$$AICu_k = \log(\ s_k^2) + \frac{n+k}{n-k-2} \tag{6.21}$$

où l'indice "u" indique que nous avons employé l'estimateur sans-biais s_k^2 au lieu de l'estimateur de vraisemblance $\hat{\sigma}_k^2$. Cette formule est obtenu par le changement de

$\log(\hat{\sigma}_k^2)$ par $\log(s_k^2)$, et le changement de la fonction de pénalité de AIC [$2(k+1)/n$] par $2(k+1)/(n-k-2)$. En suite on additionne 1 pour plus de cohésion car les constantes ne jouent aucun rôle pratique dans le choix du modèle.

Nous appliquons notre algorithme et l'équation (6.15) pour obtenir le rapport signal-bruit de $AICu$ donné par,

$$\frac{\mathbb{E}\ [\ \triangle AICu]}{\boldsymbol{sd}\ [\ \triangle AICu]} = \frac{\sqrt{(n-k-L)\,(n-k+2)}}{\sqrt{2L}} \times \left[-\frac{L}{(n-k-L)(n-k)}+\frac{2L(n-1)}{(n-k-2)(n-k-L-2)}\right] \quad (6.22)$$

Le rapport signal-bruit asymptotique pour $AICu$ est $\sqrt{2L}$.

Les petits ou négatifs rapports signal-bruit ont comme conséquence une probabilité élevée, et les grands signal-bruit mènent à des petites probabilités.

6.4.2 Les Probabilités des Critères

Supposons qu'il y a un modèle d'ordre k, et nous ajustons un modèle candidat d'ordre $k+L$, où $L > 0$. Aussi, on suppose que les deux derniers modèles sont comparés et qu'ils forment des modèles emboités. Nous calculons la probabilité de sélectionner un modèle avec $k+L$ variables au lieu du vrai modèle avec k variables.

Pour un n fini, la probabilité que AIC préfère le modèle avec $k+L$ variables est

$$\begin{aligned} P\{\ AIC_{k+L} &< AIC_k\} \\ &= P\left\{\log\left(\hat{\sigma}^2_{k+L}\right)+\frac{2(k+L+1)}{n}<\log\left(\hat{\sigma}^2_k\right)+\frac{2(k+1)}{n}\right\} \\ &= P\left\{\log\left(SSE_{k+L}\right)-\log\left(n\right)+\frac{2(k+L+1)}{n}<\log\left(SSE_k\right)-\log\left(n\right)+\frac{2(k+1)}{n}\right\} \\ &= P\left\{\frac{SSE_k-SSE_{k+L}}{SSE_{k+L}}>\exp\left(\frac{2L}{n}\right)-1\right\} \\ &= P\left\{\frac{G}{Q}>\exp\left(\frac{2L}{n}\right)-1\right\} \\ &= P\left\{F>\frac{n-k-L}{L}\left[\exp\left(\frac{2L}{n}\right)-1\right]\right\}. \end{aligned} \tag{6.23}$$

où $G = SSE_k - SSE_{k+L} \sim \boldsymbol{\chi}^2(L)$, $Q = SSE_{k+L} \sim \boldsymbol{\chi}^2(n-k-L)$ (loi de *Chi-deux* à $(n-k-L)$ degré de liberté), $F = \frac{G}{Q} \sim \mathcal{F}_{L,n-k-L}$ (loi de *Fisher* à $(L, n-k-L)$ degrés de liberté).

Pour $AICc$ on a

$$\begin{aligned} P\{\ AICc_{k+L} &< AICc_k\} \\ &= P\left\{F>\frac{n-k-L}{L}\left(\exp\left[\frac{2L(n-1)}{(n-k-L-2)\,(n-k-2)}\right]-1\right)\right\} \end{aligned} \tag{6.24}$$

et, aussi pour $AICu$, on a

$$\begin{aligned} P\{\ AICu_{k+L} &< AICu_k\} \\ &= P\left\{F>\frac{n-k-L}{L}\left[\exp\left(\frac{\log(n)L}{n}\right)-1\right]\right\}. \end{aligned} \tag{6.25}$$

Quand $n \to \infty$, et si k et L sont fixés, il est clair que la loi $\boldsymbol{\chi}^2_{n-k-L}/(n-k-L) \to 1$ presque sûrement, et du théorème de *Slutsky*, nous avons $\mathcal{F}^2_{n-k-L} \to \boldsymbol{\chi}^2_L/L$, et la loi $\mathcal{F}$ de *Fisher* d'un petit échantillon est remplacé par une loi *Chi-deux* ($\boldsymbol{\chi}^2$). Les probabilités asymptotiques sont les dérivées de la loi $\boldsymbol{\chi}^2$. Nous employons l'expression

de l'exponentielle ($\exp(z) = 1 + z + \circ(z^2)$) dans les formules des probabilités précédentes, on obtient pour le critère AIC,

$$\begin{aligned} \frac{n-k-L}{L}\left[\exp\left(\frac{2L}{n}\right)-1\right] &= \frac{n-k-L}{L}\left[\frac{2L}{n}+O\left(\frac{1}{n^2}\right)\right] \\ &= \frac{n}{L}\frac{2L}{n}+O\left(\frac{1}{n}\right)\to 2 \end{aligned}$$

avec

$$\mathcal{F}_{L,n-k-L} \to \frac{\boldsymbol{\chi}^2(L)}{L}, \text{ quand } n \to \infty$$

Donc $P\{\ AIC_{k+L} < AIC_k\} = P\{\ G > 2L\}$, où $G = SSE_k - SSE_{k+L} \sim \boldsymbol{\chi}^2(L)$ (loi de *Chi-deux* à L degré de liberté). Le $O(1/n^2)$ sera ignoré dans toutes autres dérivations.

$P\{\ AICc_{k+L} < AICc_k\} = P\{\ G > 2L\}$,

et $P\{\ AICu_{k+L} < AICu_k\} = P\{\ G > 3L\}$.

Chapitre 7

Simulation

Soient $(t_j, y_j)_{j=1,...,n}$ les n observations des deux variables indépendantes T et Y, tels que t_j et y_j sont liés par le modèle

$$y_j = \mu(t_j) + \varepsilon_j, j = 1, ..., n \ , \tag{7.1}$$

où μ est une fonction inconnue et (ε_j) est une suite de variables aléatoires, non covariées, de même loi de probabilité, telles que

$$\mathbb{E}(\varepsilon_j) = 0, var(\varepsilon_j) = \sigma^2, j = 1, .., n, \tag{7.2}$$

Supposons que les y_j, $j = 1, .., n$ sont obtenues comme des réponses de p variables d'entrées $X_1, ..., X_p$.

Comme un exemple nous allons traiter le cas de la fonction $f(t) = 40\ t\ (1 - t)$, c'est-à-dire les observations sont obtenues par

$$y_j = 40\ t_j(1 - t_j) + \varepsilon_j,\ \ j = 1, ..., 50,$$

où les points t_j, $j = 1, ..., 50$ de réalisations sont uniformément espacés dans $[0, 1]$, donné par

$$t_j = \frac{j-1}{50}, \ j = 1, ..., 50. \tag{7.3}$$

et les ε_j sont des erreurs aléatoires de loi normale de moyenne zéro, non corrélées, avec une variance commune σ^2.

(Source : '*Spline Smoothing and Nonparametric Regression*' de *Eubank Randall, L.* p. 104.)

Dans cette étude, les calculs sont faits pour trois cas de sigma : $\sigma = 0.25$, $\sigma = 0.5$, et $\sigma = 1$.

La matrice des réalisations est définie par

$$X_\lambda = \left(e^{(\ 2\pi ijr/n)}\right) \begin{matrix} r = 0, n-1 \\ j = -\lambda, \lambda \end{matrix}. \tag{7.4}$$

où $n = 50$, et $\lambda \in \{1, ..., p\}$.

Premier cas : $\sigma = 0.25$

Les valeurs des y_j, et t_j, $j = 1, ..., 50$, pour $\sigma = 0.25$ sont données dans la table

7.1.1.

Table 7.1.1. Les valeurs de y pour l'intervalle $[0, 1]$.

t	y	t	y	t	y	t	y	t	y
0.00	-0.0934	0.20	6.6308	0.40	9.7332	0.60	9.3726	0.80	6.3933
0.02	0.5760	0.22	6.8194	0.42	10.0022	0.62	9.4226	0.82	5.9474
0.04	1.6077	0.24	7.1656	0.44	9.5930	0.64	8.7851	0.84	5.5965
0.06	1.8013	0.26	8.0540	0.46	10.0265	0.64	9.2918	0.86	4.8616
0.08	2.5507	0.28	7.8465	0.48	9.9748	0.68	8.5539	0.88	4.4128
0.10	4.1039	0.30	8.6019	0.50	9.6931	0.70	7.8840	0.90	3.7270
0.12	4.2060	0.32	8.5763	0.52	9.9152	0.72	8.0917	0.92	2.9770
0.14	5.4732	0.34	9.1619	0.54	9.8959	0.74	8.0679	0.94	2.3260
0.16	5.3152	0.36	9.4280	0.56	9.5851	0.76	7.3093	0.96	1.2903
0.18	5.9473	0.38	9.2165	0.58	9.2554	0.78	6.9045	0.98	0.5480

La régression des données nécessite la sélection d'un meilleur modèle d'ajustement, et pour cela on va appliquer les méthodes discutées dans le chapitre six.

Le calcul des rapports signal-bruit pour les critères de sélections AIC, $AICc$, $AICu$ pour choisir un modèle avec k variables au lieu un modèle avec $k + L$ variables donne les résultats présentés dans la table 7.1.2. Rappelons que les rapports signal-bruit pour les critères AIC, $AICc$, et $AICu$ sont donnés par les formules suivantes

$$\frac{\mathbb{E}\ [\ \triangle AIC]}{\boldsymbol{sd}\ [\ \triangle AIC]} = \frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}} \times \left[\log\left(\frac{n-k-L}{n-k}\right) - \frac{L}{(n-k-L)(n-k)} + \frac{2L}{n}\right], \qquad (7.5)$$

$$\frac{\mathbb{E}\ [\ \triangle AICc]}{\boldsymbol{sd}\ [\ \triangle AICc]} = \frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}} \times \left[\log\left(\frac{n-k-L}{n-k}\right) - \frac{L}{(n-k-L)(n-k)} + \frac{2L(n-1)}{(n-k-2)(n-k-L-2)}\right], \tag{7.6}$$

$$\frac{\mathbb{E}\ [\ \triangle AICu]}{\boldsymbol{sd}\ [\ \triangle AICu]} = \frac{\sqrt{(n-k-L)\,(n-k+2)}}{\sqrt{2L}} \times \left[-\frac{L}{(n-k-L)(n-k)} + \frac{2L(n-1)}{(n-k-2)(n-k-L-2)}\right] \tag{7.7}$$

où $\triangle AIC = AIC_{k+L} - AIC_k$.

Table 7.1.2. Les rapports signal-bruit pour les critères de sélections AIC, $AICc$, $AICu$.

L	AIC	$AICc$	$AICu$
1	−28.93180	2700.69996	2701.42451
2	−41.47918	3665.96202	3666.98660
3	−51.51852	4304.49442	4305.74909
4	−60.34984	4759.28491	4760.73337
5	−68.47614	5088.33138	5089.95036
6	−76.15746	5322.75938	5324.53225
7	−83.55129	5481.92665	5483.84073
8	−90.76381	5579.04968	5581.09482
9	−97.87266	5623.75890	5625.92673

Les résultats dans la table 7.1.2 montrent que les rapports signal-bruit pour le critère de sélection de modèles AIC décroit quand la taille de l'échantillon L croit,

et les rapports signal-bruit pour les critères de sélection de modèles $AICc$ et $AICu$ croit quand la taille de l'échantillon L croit.

Les rapports signal-bruit asymptotiques pour les critères de sélections AIC, $AICc$, $AICu$ pour choisir un modèle avec k variables au lieu un modèle avec $k+L$ variables donne les résultats présentés dans la table 7.1.3 tels que

$$\lim_{n\to\infty} \frac{\mathbb{E}\ [\ \triangle AIC]}{\boldsymbol{sd}\ [\ \triangle AIC]} = \lim_{n\to\infty} \frac{\mathbb{E}\ [\ \triangle AICc]}{\boldsymbol{sd}\ [\ \triangle AICc]} = \sqrt{\frac{L}{2}}, \tag{7.8}$$

$$\text{et} \qquad \lim_{n\to\infty} \frac{\mathbb{E}\ [\ \triangle AICu]}{\boldsymbol{sd}\ [\ \triangle AICu]} = \sqrt{2L}. \tag{7.9}$$

Table 7.1.3. Les rapports signal-bruit asymptotiques pour les critères de sélections AIC, $AICc$, $AICu$.

L	AIC	$AICc$	$AICu$
1	0.70711	0.70711	1.41421
2	1.00000	1.00000	2.00000
3	1.22474	1.22474	2.44949
4	1.41421	1.41421	2.82843
5	1.58114	1.58114	3.16228
6	1.73205	1.73205	3.46410
7	1.87083	1.87083	3.74166
8	2.00000	2.00000	4.00000
9	2.12132	2.12132	4.24264

Les résultats de la table 7.1.3 montrent que les critères efficaces AIC et $AICc$

ont des rapports signal-bruit équivalents, et que $AICu$ a des rapports signal-bruit beaucoup plus grands que AIC et $AICc$.

Le calcul des probabilités pour les critères de sélections AIC, $AICc$, $AICu$ pour choisir un modèle avec $k+L$ variables au lieu un modèle avec k variables donne les résultats présentés dans la table 7.1.4, tels que

$$P\{\ AIC_{k+L} < AIC_k\} = P\left\{F > \frac{n-k-L}{L}\left[\exp\left(\frac{2L}{n}\right) - 1\right]\right\}, \tag{7.10}$$

$$\begin{aligned} P\{\ AICc_{k+L} &< AICc_k\} \\ &= P\left\{F > \frac{n-k-L}{L}\left(\exp\left[\frac{2L(n-1)}{(n-k-L-2)\,(n-k-2)}\right] - 1\right)\right\}, \end{aligned} \tag{7.11}$$

$$\begin{aligned} \text{et} \qquad P\{\ AICu_{k+L} &< AICu_k\} \\ &= P\left\{F > \frac{n-k-L}{L}\left[\exp\left(\frac{\log(n)L}{n}\right) - 1\right]\right\}. \end{aligned} \tag{7.12}$$

où $F = \frac{SSE_k - SSE_{k+L}}{SSE_{k+L}} \sim \mathcal{F}_{L,n-k-L}$ (loi de *Fisher* à $(L, n-k-L)$ degrés de liberté), $SSE_k = \sum_{j=1}^{n} (\ y_j - \hat{y}_j)^2$, et $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)' = X\ \hat{\beta}$.

Table 7.1.4. Le calcul des probabilités pour les critères de sélections AIC, $AICc$, $AICu$.

L	AIC	$AICc$	$AICu$
1	0.21459	0.00000	0.08261
2	0.04986	0.00000	0.00345
3	0.00534	0.00000	0.00004
4	0.00036	0.00000	0.00000
5	0.00002	0.00000	0.00000
6	0.00000	0.00000	0.00000
7	0.00000	0.00000	0.00000
8	0.00000	0.00000	0.00000
9	0.00000	0.00000	0.00000

Les résultats dans la table 7.1.4 montrent que les probabilités pour le critère de sélection de modèles AIC et $AICu$ décroient quand la taille de l'échantillon L croit.

Maintenant, le calcul des probabilités asymptotique pour les critères de sélection AIC, $AICc$, $AICu$ pour choisir un modèle avec $k+L$ variables au lieu un modèle avec k variables donne les résultats présentés dans la table 7.1.5, tels que

$$P\{\ AIC_{k+L} \ < \ AIC_k\} = P\{\ G > 2L\}, \tag{7.13}$$

$$P\{\ AICc_{k+L} \ < \ AICc_k\} = P\{\ G > 2L\}, \tag{7.14}$$

et

$$P\{\ AICu_{k+L} \ < \ AICu_k\} = P\{\ G > 3L\}. \tag{7.15}$$

où $G = SSE_k - SSE_{k+L} \sim \boldsymbol{\chi}^2(L)$ (loi de *Chi-deux* à L degré de liberté), $SSE_k =$

$\sum_{j=1}^{n} (\ y_j - \hat{y}_j)^2$, et $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)' = X\ \hat{\beta}$.

Table7.1.5. Les probabilités asymptotiques pour les critères de sélections AIC, $AICc$, $AICu$.

L	AIC	$AICc$	$AICu$
1	0.15730	0.15730	0.08326
2	0.13534	0.13534	0.04979
3	0.11161	0.11161	0.02929
4	0.09158	0.09158	0.01735
5	0.07524	0.07524	0.01036
6	0.06197	0.06197	0.00623
7	0.05118	0.05118	0.00377
8	0.04238	0.04238	0.00229
9	0.03517	0.03517	0.00140

D'après la table 7.1.5, il est clair que AIC et $AICc$ ne sont pas des critères consistants, puisque leurs probabilités ne sont pas nulles et on peut voir aussi que AIC et $AICc$ sont asymptotiquement équivalents. Le signal-bruit de la variante corrigée $AICu$ a des plus petites probabilités asymptotiques qu'AIC et $AICc$.

Pour estimer la fonction de régression μ du modèle 7.1, nous calculons maintenant les valeurs des estimateurs du risque de prédiction $MSE(\lambda)$, $GCV(\lambda)$, et $\hat{P}(\lambda)$ qui

sont données par les formules suivantes

$$MSE(\lambda) = \frac{1}{n}\, \underline{y}'\, [\, I - H(\lambda)]^2\, \underline{y}, \tag{7.16}$$

$$\hat{P}(\lambda) = MSE(\lambda) + 2\, \hat{\sigma}^2 \left(\frac{\#(\lambda)}{n} \right), \tag{7.17}$$

$$\text{et} \qquad GCV(\lambda) = \frac{MSE(\lambda)}{\left(1 - \frac{\#(\lambda)}{n}\right)}\ . \tag{7.18}$$

où $H(\lambda) = X_\lambda\ (X_\lambda'\ X_\lambda)^{-1} X_\lambda'$, X_λ donné par la formule 7.4, $\#(\lambda) = \text{tr}[H(\lambda)]$, $n = 50$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (\ y_j - \hat{y}_j)^2$, $\underline{y} = (y_1, ..., y_n)$, et $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)' = X\ \hat{\beta}$.

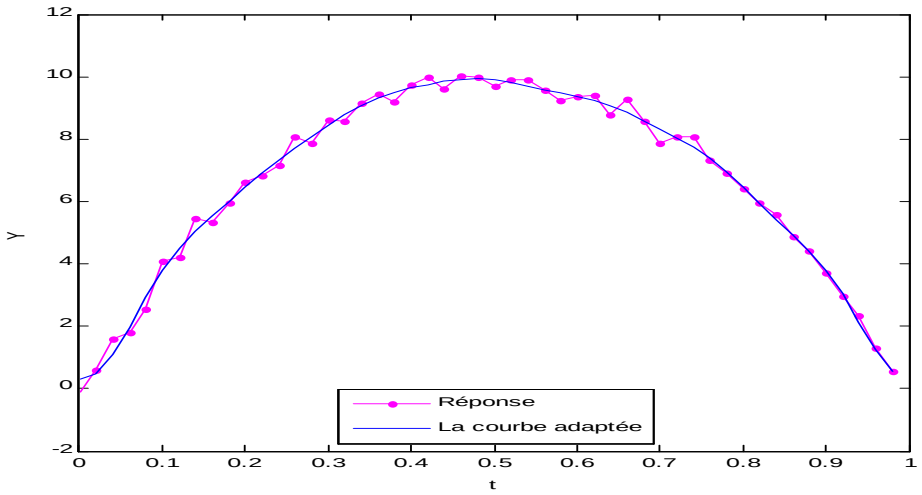
Table 7.1.6. Les valeurs de $MSE(\lambda)$, $GCV(\lambda)$, et $\hat{P}(\lambda)$.

λ	$MSE(\lambda)$	$GCV(\lambda)$	$\hat{P}(\lambda)$
1	0.89672	0.95396	0.90422
2	0.27145	0.30161	0.28395
3	0.12934	0.15039	0.14684
4	0.07741	0.09440	0.09991
5	0.05710	0.07321	0.08460
6	0.05052	0.06828	0.08302
7	0.04783	0.06833	0.08533
8	0.04708	0.07133	0.08958
9	0.04680	0.07548	0.09430
10	0.04080	0.07035	0.09330

La table 7.1.6 donne le sommaire des valeurs de MSE, $\hat{P}(\lambda)$ et $GCV(\lambda)$ pour tous $\lambda \in \{1, ..., p\}$.

En examinant les valeurs dans la table 7.1.6, on trouve que $\hat{P}(\lambda)$ et $GCV(\lambda)$

Figure 7.1.7. L'adjustement des données par la courbe adaptée.



donnent la valeur $\lambda = 6$ comme une valeur optimale qui minimise $P(\lambda)$. Du point de vue pratique, il n'y a pas une grande différence entre ces trois critères. Donc, on peut utiliser $\lambda = 6$ pour construire l'estimateur μ_λ qui donne la courbe adaptée, où

$$\mu_\lambda(t) = \sum_{j=-\lambda}^{\lambda} \beta_{\lambda j}\, e^{2\pi i j t}. \tag{7.19}$$

$$\text{et} \qquad \beta_{\lambda j} = \frac{1}{n}\sum_{r=1}^{n} y_r\, e^{[\,-2\pi i j(r-1)/n]},\ j = -\lambda, ..., \lambda. \tag{7.20}$$

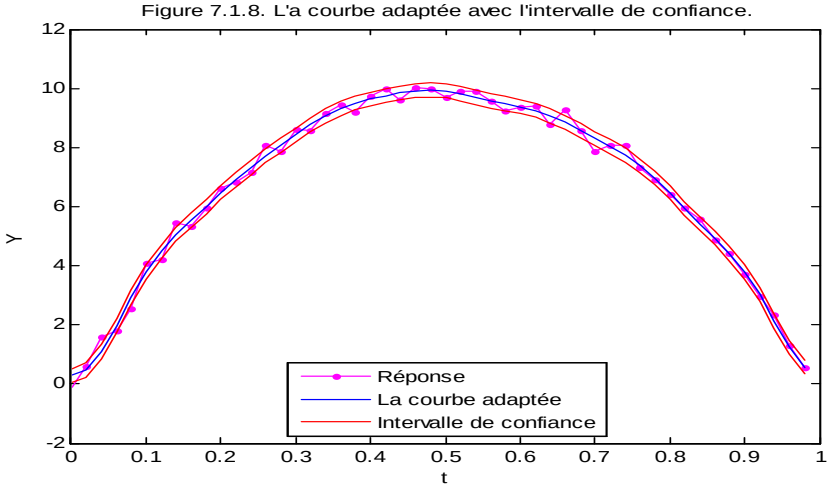
La figure 7.1.7 montre l'ajustement des données par la courbe adaptée.

Par une simple manipulation on peut utiliser l'estimateur μ_λ ($\lambda = 6$) pour construire l'intervalle de confiance estimé pour un niveau de confiance égal a $(1 - \alpha)$ avec $\alpha = 0.05$. La figure 7.1.8 donne les intervalles de confiance de β_i et $\mu(t_i)$ obtenus par

$$\beta_{\lambda i} \pm Z_{\alpha/2}\left[\ \sigma_\lambda^2 c_{\lambda,ii}\right]^{1/2} \tag{7.21}$$

$$\text{et} \qquad \mu_\lambda(t_i) \pm Z_{\alpha/2}\left[\sigma_\lambda^2\, h_{ii}(\lambda)\right]^{1/2} \tag{7.22}$$

Figure 7.1.8. L'a courbe adaptée avec l'intervalle de confiance.



où $h_{ii}(\lambda)$ est le $i^{\text{ème}}$ élément diagonale de $H(\lambda) = X_\lambda\ (X'_\lambda\ X_\lambda)^{-1}X'_\lambda$, X_λ donné par la formule 7.4, $c_{\lambda,ii}$ est le $i^{\text{ème}}$ élément diagonale de $c_\lambda = (\ X'_\lambda\ X_\lambda)^{-1}$, $\sigma^2_\lambda = n\ MSE(\lambda)/n - \lambda$, $MSE(\lambda)$ donné par la formule 7.16, et $Z_{\alpha/2}$ le $(\alpha/2)$ pourcentage des points supérieurs de la loi normale standard.

Deuxième cas : $\sigma = 0.5$

Le calcul des y_j, et t_j, $j = 1, ..., 50$, pour $\sigma = 0.5$ sont données dans la table 7.2.1.

Table 7.2.1. Les valeurs de y pour l'intervalle $[0, 1]$.

t	y	t	y	t	y	t	y	t	y
0.00	-0.0065	0.20	6.1446	0.40	9.59599	0.60	9.2966	0.80	5.8348
0.02	0.9612	0.22	6.9815	0.42	10.1732	0.62	9.7674	0.82	5.7577
0.04	1.0886	0.24	6.9971	0.44	10.2434	0.64	9.2260	0.84	5.0847
0.06	2.6621	0.26	7.7064	0.46	10.5890	0.64	9.5079	0.86	4.3678
0.08	2.9988	0.28	8.2737	0.48	10.5998	0.68	8.0335	0.88	4.3483
0.10	4.9658	0.30	8.9956	0.50	10.4793	0.70	8.6398	0.90	2.8552
0.12	4.4295	0.32	9.0896	0.52	9.1567	0.72	7.2470	0.92	3.1008
0.14	4.1626	0.34	7.6539	0.54	9.4408	0.74	6.9747	0.94	1.2435
0.16	5.5679	0.36	9.3587	0.56	10.1986	0.76	7.4429	0.96	1.8005
0.18	6.1538	0.38	9.8370	0.58	9.2566	0.78	6.7938	0.98	0.9557

Les résultats des rapports signal-bruit et des probabilités pour les critères de sélection de modèles AIC, $AICc$, et $AICu$ reste les même dans les trois cas de sigma $\sigma = 0.25$, $\sigma = 0.5$, et $\sigma = 1$. Donc on passe au calcul des valeurs des estimateurs du risque de prédiction $MSE(\lambda)$, $GCV(\lambda)$, et $\hat{P}(\lambda)$ définis par les formues 7.16, 7.17, et 7.18. Les calculs sont donnés dans la table 7.2.2.

Table 7.2.2. Les valeurs de $MSE(\lambda)$, $GCV(\lambda)$, et $\hat{P}(\lambda)$.

λ	$MSE(\lambda)$	$GCV(\lambda)$	$\hat{P}(\lambda)$
1	0.80466	0.85602	0.83466
2	0.39799	0.44221	0.44799
3	0.28898	0.33602	0.35898
4	0.24881	0.30343	0.33881
5	0.21415	0.27455	0.32415
6	0.18535	0.25047	0.31535
7	0.18159	0.25941	0.33159
8	0.17989	0.27256	0.34989
9	0.17277	0.27866	0.36277
10	0.15190	0.26189	0.36190

En examinant les valeurs dans la table 7.2.2, on trouve que $\hat{P}(\lambda)$ et $GCV(\lambda)$ donnent la valeur $\lambda = 6$ comme une valeur optimale qui minimise $P(\lambda)$. Du point de vue pratique, il n'y a pas une grande différence entre ces trois critères. Donc, on peut utiliser $\lambda = 6$ pour construire l'estimateur μ_λ qui donne la courbe adaptée, en utilisant les formules 7.19, 7.20. La figure 7.2.3 montre l'ajustement des données par la courbe adaptée.

Si on pose $\lambda = 6$, on peut utiliser l'estimateur μ_λ pour construire l'intervalle de confiance estimé pour un niveau de confiance égal a $(1 - \alpha)$ avec $\alpha = 0.05$. Les intervalles sont calculés d'après les formules 7.21, 7.22, et qui sont représenter dans la figure 7.2.4.

Figure 7.2.3. L'adjustement des données par la courbe adaptée.

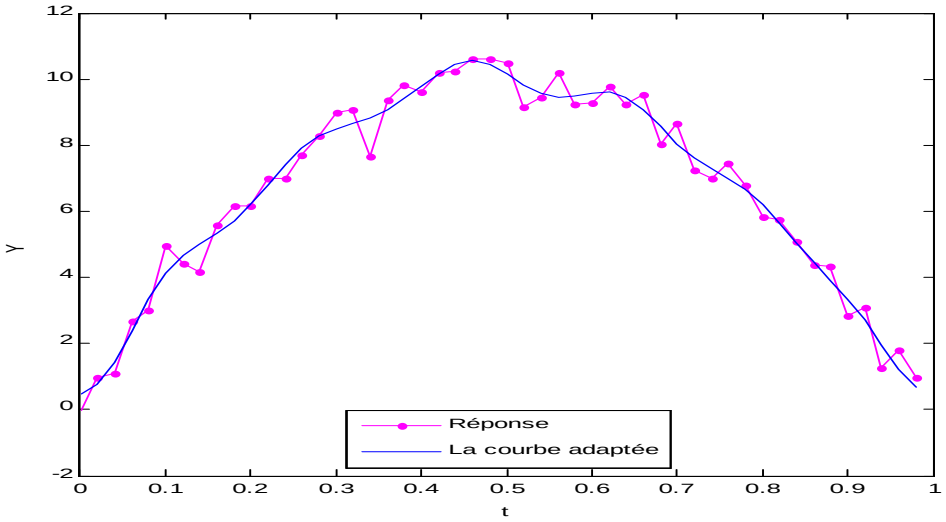
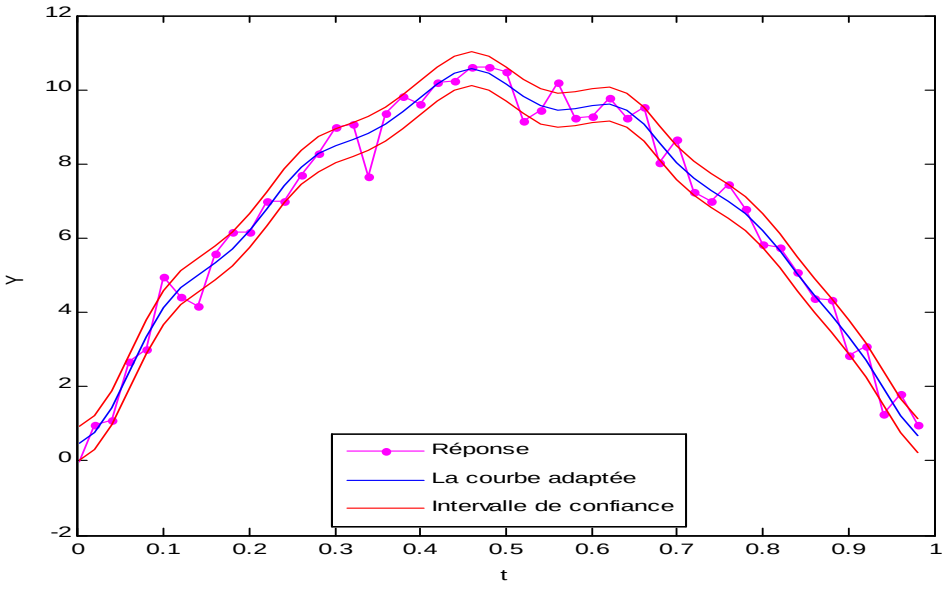


Figure 7.2.4. L'a courbe adaptée avec l'intervalle de confiance.



Troisième cas : $\sigma = 1$

Le calcul des y_j, et t_j, $j = 1, ..., 50$, pour $\sigma = 1$ sont données dans la table 7.3.1.

Table 7.3.1. Les valeurs de y pour l'intervalle $[0, 1]$.

t	y	t	y	t	y	t	y	t	y
0.00	0.7582	0.20	6.0632	0.40	8.2282	0.60	11.0484	0.80	7.1599
0.02	0.0921	0.22	7.8348	0.42	9.0581	0.62	8.3989	0.82	4.1911
0.04	2.2162	0.24	7.1888	0.44	10.1877	0.64	9.4214	0.84	6.9130
0.06	1.1835	0.26	8.7095	0.46	8.9383	0.64	9.5649	0.86	3.2062
0.08	3.8438	0.28	7.5887	0.48	10.2754	0.68	8.4400	0.88	5.3335
0.10	1.4769	0.30	8.4689	0.50	11.1071	0.70	10.8953	0.90	2.4903
0.12	4.5087	0.32	9.1026	0.52	10.2290	0.72	8.9199	0.92	3.3295
0.14	4.0827	0.34	10.0923	0.54	10.1010	0.74	6.8450	0.94	3.2212
0.16	4.6026	0.36	9.8365	0.56	10.2622	0.76	8.1079	0.96	2.3543
0.18	6.0558	0.38	9.1363	0.58	10.9600	0.78	7.5642	0.98	0.8210

Les valeurs des estimateurs du risque de prédiction $MSE(\lambda)$, $GCV(\lambda)$, et $\hat{P}(\lambda)$ définis par les formues 7.16, 7.17, et 7.18 sont données dans la table 7.3.2.

Table 7.3.2. Les valeurs de $MSE(\lambda)$, $GCV(\lambda)$, et $\hat{P}(\lambda)$.

λ	$MSE(\lambda)$	$GCV(\lambda)$	$\hat{P}(\lambda)$
1	1.45855	1.55165	1.48855
2	0.84348	0.93721	0.89348
3	0.78371	0.91129	0.85371
4	0.78281	0.95464	0.87281
5	0.73625	0.94391	0.84625
6	0.73067	0.98740	0.86067
7	0.71290	1.01842	0.86290
8	0.66448	1.00679	0.83448
9	0.63880	1.03032	0.82880
10	0.63410	1.09328	0.84410

En examinant les valeurs dans la table 7.3.2, on trouve que $\hat{P}(\lambda)$ et $GCV(\lambda)$ donnent la valeur $\lambda = 5$ comme une valeur optimale qui minimise $P(\lambda)$. En pratique, il n'y a pas une grande différence entre ces trois critères. Donc, on peut utiliser $\lambda = 5$ pour construire l'estimateur μ_λ qui donne la courbe adaptée et cela en appliquant les formules 7.19, 7.20. La figure 7.3.3 montre l'ajustement des données par la courbe adaptée.

Si on pose $\lambda = 5$, on peut utiliser l'estimateur μ_λ pour construire l'intervalle de confiance estimé pour un niveau de confiance égal a $(1 - \alpha)$ avec $\alpha = 0.05$. Les intervalles sont calculés d'après les formules 7.21, 7.22, et représentés dans la figure 7.3.4.

Figure 7.3.3. L'adjustement des données par la courbe adaptée.

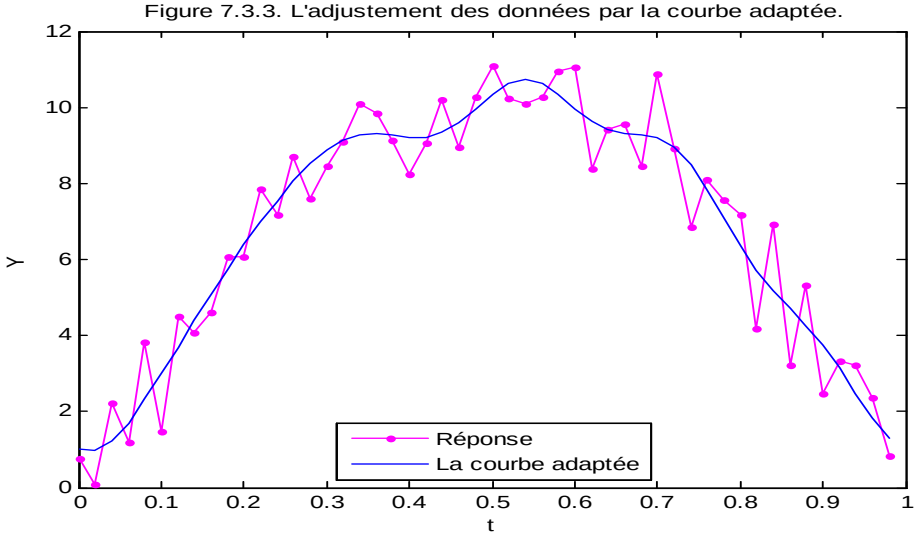
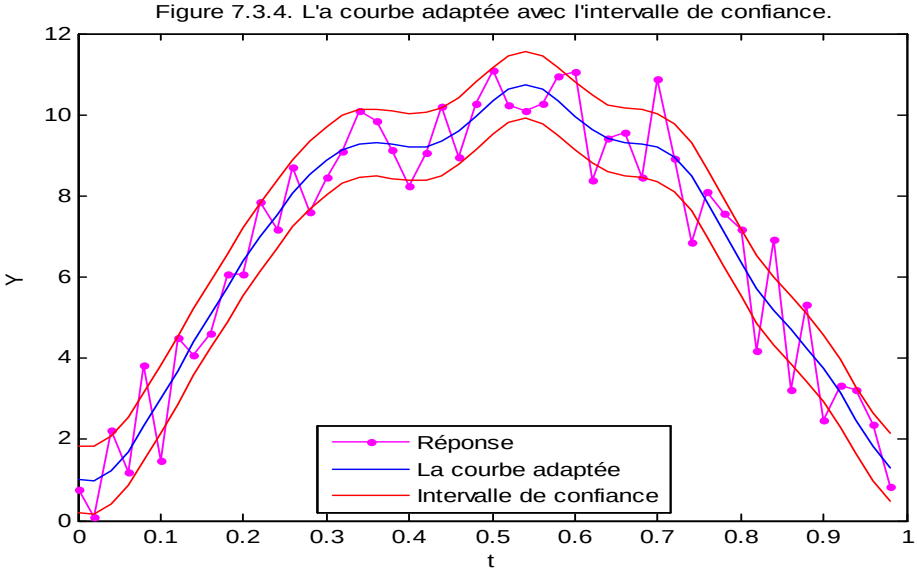


Figure 7.3.4. L'a courbe adaptée avec l'intervalle de confiance.



REFERENCES

[1] Akaike, H. A. (1978). Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math. A* 30 : 9 − 14.

[2] Akaike, H. A. (1974). New look at the statistical model identification. *I.E.E.E. Trans. Auto. Control* 19 : 716 − 723.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In 2nd International Symposium on Information Theory* (B. N. Petrov and F. Csàki, eds.), Budapest : Akademia Kiadó : 267 − 281.

[4] Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math* 22 : 203 − 217.

[5] Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics* 16 : 125 − 127.

[6] Berry, S. M., Carroll, R. J., et Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* 97 : 160 − 9.

[7] Burnham, K. P., et Anderson, D. (2002). *Model Selection and multi-Model Inference*, 2nd ed. New York : Springer-Verlag.

[8] Cai, Z., Fan, J., et Li, R. Z. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* 95 : 888 − 902.

[9] Cornillon, P. A. Eric Matzner-Lober, "Régression Théorie et application".

[10] Craven, P. et Wahba, G. (1979). Smoothing noisy data with spline functions : estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math* 31 : 377 − 403.

[11] Davis, P. J. (1975). *Interpolation and Approximation*. New York : Dover Publications, Inc.

[12] Delecroix, M., et Thomas-Agnan, C. (2000). Spline and kernel regression under shape restriction. In M. G. Schimek (Ed.), *Statistical Theory and Computational Aspects of Smoothing*, pp. 109 − 33. Heidelberg : Physica-Verlag.

[13] Efron, B. (1982). The *Jackknife*, the *Bootstrap and Other Resampling Plans*. SIAM monograph no. 38, CBMS-NSF. Philadelphia : SIAM.

[14] Eubank, R. L. "Spline Smoothing and Nonparametric Regression", Department of Statistics, Southern Methodist University, Dallas, Texas. Mercel Dekker, INC. New York and Basel.

[15] Fahrmeir, L., et Tutz, G. (2002). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed. New York : Springer-Verlag.

[16] Geisser, S. (1975). The predictive sample reuse method with application. *J. Amer. Statist. Assoc.* 70 : 320 − 328.

[17] Golub, G. Heath, M. et Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21 : 215 − 223.

[18] Gunst, R. F. et Mason, R. L. (1980). *Regression Analysis and Its Application : A Data-Oriented Approach*. New York : Marcel-Dekker.

[19] Hannan, E. J. et Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. B* 41 : 190 − 195.

[20] Hansen, M.H., Huang, J. Z., Kooperberg, C., Stone, C. J., et Truong, Y. K. (2003). Statistical Modeling with Spline Functions : *Methodology and Theory*. New York : Spriger-Verlag.

[21] Li, K. C. (1987). Asymptotic optimality for $\boldsymbol{C}_P$, $\boldsymbol{C}_L$, cross-validation and generalized cross-validation : discrete index set. *Ann. Statist.*, to appear.

[22] Li, K. C. (1984*a*). Regression models with infinitely many parameters : consistency of bounded linear functions. *Ann. Statist.* 12 : 601 − 611.

[23] Mallows, C. L. (1973). Some comments on $\boldsymbol{C}_P$. *Technometrics* 15 : 661−675.

[24] McCulloch, C. E., et Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York : Wiley.

[25] Myers, R. H., Montgomery, D. C., et Vining, G. G. (2001). *Generalized Linear Models : With Applications in Engineering and the Sciences*. New York : Wiley.

[26] Pinheiro, J. C., et Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York : Springer-Verlag.

[27] Rafajlowicz, E. (1987). Nonparametric orthogonal series estimators of regression : a class attaining the optimal convergence rate in L_2. *Statist. and Prob. Letters* 5 : 219 − 224.

[28] Raudenbush, S. W., Yang, M.-L., et Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* 9 : 141 − 57.

[29] Rutkowski, L. (1982). On system identification by nonparametric function fitting. *IEEE Trans. Automat. Control* 27 : 225 − 227.

[30] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 6 : 461 − 464.

[31] Searle, S. R. (1971). *Linear Models*. New York : John Wiley.

[32] Serfling, R. J. (1980). *Approximation Theorems of Mathematical statistics*. New York : John Wiley.

[33] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* 68 : 45 − 54.

[34] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (With discussion). *J. roy. Statist. Soc. B* 36 : 111 − 147.

[35] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. A* 7 : 13 − 26.

[36] Taylor, A. E. et Mann, W. R. (1972). *Advanced Calculus*. Lexington Mass : Xerox College Publishing.

[37] Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev.* 53 : 141 − 170.

[38] Wahba, G. (1977*c*). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (P. R. Krishnaiah, ed.) : 507 − 523. Amsterdam : North holland.

[39] Wahba, G. (1979). Convergence rates of "thin plate" smoothing splines when the data are noisy. *In Smoothing Techniques for Curve Estimation* (Gasser, T. et Rosenblatt, M. eds.) : 233 − 245. Lecture Notes in Math. No. 757. New York : Springer-Verlag.

[40] Wahba, G. (1984*a*). Cross-validated spline methods for the estimation of multivariate functions from data on functionals. In *statistics : An Appraisal, proceeding* 50*th Anniversary Conference Iowa State Statistical Laboratory* (David, H. A. et David, H. T. eds.) : 205 − 235. Ames : Iowa State Univ. Press.

[41] Wahba, G. (1977*b*). Optimal smoothing of density estimates. In *Classification and Clustering* (Van Ryzin, J. ed.) : 423 − 458. New York : Academic Press.

[42] Wahba, G. (1984*b*). Partial spline models for the semi-parametric estimation of functions of several variables. In *Statistical Analysis of Time Series :* 319 − 329. Tokyo : Institute of Statistical Mathematics.

[43] Wegman, E. J. et Wright, I. W. (1983). Splines in statistics. *J. Amer. Statist. Assor.* 78 : 351 − 365.

Abstract

The nonparametric regression is a statistical tool making it possible to describe the relation between a dependent variable and one or more explanatory variables, without specifying of strict form for this relation. In this memory, we present an analysis of Fourier series who are employed in many sciences and technology. Thus, we stress a nonparametric model of regression. To study this type of model, one writes it in the form of a parametric model. For this reason, one supposes that the function of regression is linear. The estimate of the parameters is made by applying the method of validation cross which is followed by an asymptotic study and a nonparametric test of assumption.

To choose a better model of adjustment, one applies the selection with the criterion of the information of Akaike, where these corrected alternatives are described, and one derives their moments and probabilities and also one binds the latter to the performance via the concept of the signal-to-noise ratio.

At the end, we present an example of simulation for study the consistence of our estimators.

Key words : *Nonparametric regression, multiple linear Regression, cross validation, Akaike information criterion.*

ملخص

التسوية دون ثوابت هي وسيلة إحصائية تسمح بإعطاء علاقة بين متغير مستقل و أخر أو عدة متغيرات أخرى تفسيرية، دون تحديد شكل مدقق لهذه العلاقة. ضمن هذه المذكرة نقوم بدراسة تحليلية لسلسلة فوريه (*Fourier*) و التي ذكرت في عدة علوم و تقنيات. نعطي الضوء على نموذج لتسوية دون ثوابت. لدراسة هذا النموذج، نكتبه على شكل نموذج بثوابت.من اجل هذا الأخير، نفرض أن دالة التسوية خطية وعلى شكل سلسلة فوريه. لتقدير الثوابت في هذه المذكرة استعملنا طريقة التصديق المتقاطع (*La validation croisé*) متبوعة بدراسة تقريبية و فحص فرضية دون ثوابت.

لاختيار أحسن نموذج تسوية استعملنا معيار *Akaike*، أين نعطي صيغ مختلفة له، و نشتق تباينها و احتمالا□ا، و كذلك نربط هته الأواخر بمعنى نسبي إشارة–تشتت.

أخيرا نعطي مثالا عدديا لدراسة صلابة التقدير.

كلمات مفتاحيه: تسوية خطية مضاعفة، تسوية دون ثوابت، التصديق المتقاطع، معيار *Akaike*.