VERSAILLES
N° d'ordre :

UNIVERSITÉ DE VERSAILLES
SAINT-QUENTIN-EN-YVELINES
U.F.R. DE SCIENCES

THÈSE

présentée pour obtenir

Le TITRE de DOCTEUR EN SCIENCES
Spécialité : Mathématiques

par

Zaher MOHDEB

Sujet de la thèse

Tests d'hypothèses linéaires dans un modèle de régression non paramétrique

Soutenue le 22 Janvier 1999 devant le jury composé de :

Mme	Brigitte	CHAUVIN
M.	Xavier	GUYON
Mme	Sylvie	HUET
M.	Marc	LAVIELLE
M.	Abdelkader	MOKKADEM
M.	Jean-Michel	POGGI

VERSAILLES
N° d'ordre :

UNIVERSITÉ DE VERSAILLES
SAINT-QUENTIN-EN-YVELINES
U.F.R. DE SCIENCES

THÈSE

présentée pour obtenir

Le TITRE de DOCTEUR EN SCIENCES
Spécialité : Mathématiques

par

Zaher MOHDEB

Sujet de la thèse

Tests d'hypothèses linéaires dans un modèle de régression non paramétrique

Soutenue le 22 Janvier 1999 devant le jury composé de :

Mme	Brigitte	CHAUVIN
M.	Xavier	GUYON
Mme	Sylvie	HUET
M.	Marc	LAVIELLE
M.	Abdelkader	MOKKADEM
M.	Jean-Michel	POGGI

Testing for Linear Hypotheses in a Nonparametric Regression Model

Abstract

This thesis is devoted to the construction of hypotheses tests on the regression function f of a nonparametric regression model. In the first part, we construct tests for hypotheses on Fourier coefficients of f. Such tests can be used to compare two noisy signals in a given band of frequencies. The test statistics we use are function of the empirical Fourier coefficients of f. The second part deals with the test of the hypothesis "f is in E" where E is a finite dimensional vector space. We give two test statistics $\widehat{R}_n^2$ and $\widehat{\mathcal{M}}_n^2$ based upon two different approximations of the L^2 distance. The first one is obtained by estimating this distance by the empirical distance between the observations and the vector space E. The second one is constructed by using the observations suitably corrected. In this part, we assume that the functions satisfy the Hölder condition with order strictly greater than $1/2$, and we obtain the asymptotic weak behaviour of both statistics. The third part is an extension of the second one to the case the functions are Riemann-integrable; the weak behaviour of the statistic $\widehat{R}_n^2$ is then quite different from the behaviour obtained in the previous part, since a nonnegligeable quadratic term appears in the limit result. However, this additional term is explicit and allows thus the construction of several tests.

Key words: Empirical Fourier coefficients; Linear hypothesis; Nonparametric regression; Nonlinear regression; Nonparametric test.

Remerciements

Mes premiers remerciements vont à Abdelkader Mokkadem qui m'a fait bénéficier de son expérience et de sa compétence; il a su diriger mes recherches avec intuition, enthousiasme, disponibilité et rigueur. Je lui en suis extrêmement reconnaissant.

Je tiens à exprimer ma reconnaissance à Xavier Guyon et Jean-Michel Poggi qui ont bien voulu examiner ce travail et accepter d'en être les rapporteurs.

Je remercie vivement Brigitte Chauvin, Sylvie Huet et Marc Lavielle de m'avoir fait l'honneur et le plaisir de participer au jury de cette thèse.

Je remercie tous les membres du département de Mathématiques de l'Université de Versailles Saint-Quentin-en-Yvelines qui m'ont offert d'excellentes conditions de travail. Je remercie Mariane Pelletier pour toute l'attention qu'elle a porté à mon travail. Merci aussi à Fatima Aumar et Marie-France Thai pour leur soutien.

Je tiens aussi à remercier toute l'équipe du Laboratoire de Modélisation Stochastique et Statistique d'Orsay qui m'a réservé un accueil chaleureux pendant mes séjours de formation et tout particulièrement Jean Bretagnolle qui a été pour moi un soutien permanent aussi bien sur le plan humain que matériel. Mes sincères remerciements à Sabine Hoarau.

Enfin j'adresse une pensée à Salima pour son soutien et ses encouragements permanents; Ines et Yasmine pour leur patience. Qu'elles soient assurées de mon affection.

Table des Matières

Introduction

0.1 Présentation

Cette thèse est consacrée à la construction de tests d'hypothèses linéaires sur la fonction de régression f, du modèle non paramétrique suivant

$$Y_{i,n} = f(t_{i,n}) + \varepsilon_{i,n}\,, \quad i = 1, \ldots, n\,, \tag{1}$$

où f est une fonction réelle inconnue, définie sur l'intervalle $[0,1]$ et $t_{1,n} = 0 < t_{2,n} < \cdots < t_{n,n} = 1$, est un échantillonnage fixé de l'intervalle $[0,1]$. Les erreurs $\varepsilon_{i,n}$ forment un tableau triangulaire de variables aléatoires d'espérance nulle et de variance finie σ^2 et pour tout n les variables aléatoires $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ sont indépendantes.

Le problème de test d'hypothèses dans le modèle (1) a donné lieu à de nombreux travaux ces vingt dernières années, la préocupation la plus importante étant de choisir entre une modélisation paramétrique et une modélisation non paramétrique.

Bien que les méthodes abordées soient diverses, une grande partie de la littérature sur le problème des tests d'hypothèses dans le modèle (1) est basée sur la méthode des splines. Cox et Koh (1989) donnent un test de l'hypothèse "f est un polynôme de degré inférieur à m". Eubank et Spiegelmann (1990) construisent un test de linéarité de f dans le cas d'un modèle normal en se ramenant à tester la nullité de la partie non linéaire. Jayasuriya (1996) généralise l'approche de Eubank et Spiegelmann (1990) pour tester l'hypothèse "f est un polynôme", dans le cas d'un modèle non normal.

L'utilisation de l'estimateur à noyau de f est proposée par Müller (1992) pour construire un test d'ajustement du modèle. Härdle et Mammen (1993) proposent de combiner l'estimation par la méthode du noyau avec les techniques de bootstrap pour tester l'hypothèse que le modèle est paramétrique.

La méthode du maximum de vraisemblance est également envisagée par Staniswalis et Severini (1991); pour tester l'hypothèse que le modèle est paramétrique, ils comparent l'estimateur du maximum de vraisemblance quand le modèle est paramétrique avec un estimateur du maximum de vraisemblance non paramétrique (ce dernier est construit à l'aide d'un estimateur à noyau de la vraisemblance non paramétrique).

L'usage des coefficients de Fourier empiriques de f pour construire des tests d'hypothèses dans le modèle (1) est abordé par Eubank et Spiegelmann (1990) et Eubank et Hart (1992). Nous en faisons également usage dans le chapitre 1, pour tester l'hypothèse que f appartient à un sous-espace vectoriel de dimension finie ou infinie (voir aussi Mohdeb et Mokkadem (1998)).

Plus récemment Dette et Munk (1998) considèrent le test de l'hypothèse "$f \in E$" où E est un espace vectoriel de dimension finie; ils proposent pour cela d'utiliser un estimateur empirique du carré de la distance de f à E.
Dans les chapitres 2 et 3, nous suivons une démarche similaire et proposons deux nouvelles statistiques plus simples qui sont asymptotiquement équivalentes à celle de Dette et Munk (1998) et qui, pour de petits échantillons, semblent donner de meilleurs résultats.

0.2 Description de notre travail

Avant de présenter nos résultats de façon détaillée, nous donnons tout d'abord l'organisation de ce travail. Il est constitué de trois parties.

La première (chapitre 1) est consacrée à l'étude de construction de tests d'hypothèses sur les coefficients de Fourier de f, dans une bande de fréquences donnée. De tels tests peuvent, en particulier, être utilisés pour comparer deux signaux bruités dans une bande de fréquences. Les statistiques de tests que nous utilisons, sont basées sur les coefficients de Fourier empiriques de f.

La deuxième et la troisième partie concernent l'étude du test de l'hypothèse "$f \in E$"

où E est un espace vectoriel de dimension finie.
Dans la deuxième partie (chapitre 2), nous supposons que l'hypothèse nulle et l'alternative sont dans la classe des fonctions höldériennes d'ordre $\gamma > 1/2$.
La troisième partie (chapitre 3) est une extension de la deuxième partie, au cas où les alternatives et l'hypothèse nulle sont dans la classe des fonctions Riemann-intégrables. Nous établissons la loi limite des statistiques de tests que nous proposons; cela donne une approximation du niveau et de la puissance du test pour de grands échantillons. Des simulations pour des petites tailles d'échantillons ont été menées pour vérifier la validité de l'utilisation de ces statistiques.

0.2.1 Résultats de la première partie

Dans cette partie, on considère un échantillonnage uniforme: $t_{j,n} = j/n$, $j = 1, \ldots, n$. L'objet de cette partie de la thèse, est de construire des tests d'hypothèses sur les coefficients de Fourier de f. Plus précisément, soit $c_k = \int_0^1 e^{-2\pi ikt} f(t)dt$, $k \in \mathbb{Z}$, les coefficients de Fourier de f et soit $\mathcal{I} = \{i_1 < i_2 < \cdots < i_r < \cdots\}$ un sous-ensemble de $\mathbb{N}$. On veut construire un test de l'hypothèse nulle

$$H_0 : \ c_k = c_k^0 \ \forall k \in \mathcal{I} \text{ contre l'hypothèse } \ H_1 : \ \exists k \in \mathcal{I} \ c_k \neq c_k^0 \qquad (2)$$

où les c_k^0, $k \in \mathcal{I}$, sont des coefficients de Fourier donnés d'une fonction réelle.
Comme f est à valeurs réelles, on a $c_k = \overline{c}_{-k}$, $\forall k \in \mathbb{Z}$ et donc tout test sur les c_k, $k \in \mathbb{Z}$, se ramène à un test sur les c_k, $k \geq 0$. Ainsi l'hypothèse H_0 ci-dessus est équivalente à $c_k = c_k^0$, $|k| \in \mathcal{I}$.

L'hypothèse H_0 recouvre de nombreuses situations. Par exemple, le test de l'hypothèse $f \equiv f_0$, où f_0 est une fonction donnée, est le test de H_0 avec c_k^0 coefficient de Fourier de f_0 et $\mathcal{I} = \mathbb{N}$; le test de l'hypothèse "f est une fonction trigonométrique de la forme $f(t) = \sum_{|k| \leq s} c_k e^{2i\pi kt}$, s fixé " se ramène au test de H_0 avec $c_k^0 = 0$ et $\mathcal{I} = \{s+1, s+2, \ldots\}$. Un autre exemple intéressant est celui de la comparaison de deux signaux. On a deux modèles analogues au modèle (1), $U = g + \varepsilon$ et $V = h + \eta$; on veut tester l'hypothèse "les deux signaux h et g coïncident sur une bande de fréquences". Plus précisément, soit d_k (resp. e_k) le k-ième coefficient de Fourier de g (resp. h) et soit $\mathcal{I} \subset \mathbb{N}$; on veut tester l'hypothèse "$d_k = e_k$, $\forall k \in \mathcal{I}$ ". Cela revient à tester H_0 avec $c_k = d_k - e_k$, $c_k^0 = 0$, dans le modèle $Y = U - V = (g - h) + \xi$ où $\xi = \varepsilon - \eta$.

Dans notre problème, il est immédiat que

$$H_0 \text{ est vraie si et seulement si } \sum_{|k| \in \mathcal{I}} |c_k - c_k^0|^2 = 0.$$

Le test sera donc basé sur une estimation de cette quantité. On commence par estimer c_k par l'estimation empirique $\hat{c}_k = \frac{1}{n} \sum_{j=1}^n Y_j e^{-2\pi ikj/n}$. On est ensuite amené à considérer

deux cas, suivant que $\mathcal{I} = \{i_1 < i_2 < \cdots < i_r < \cdots\}$ est fini ou non.

Cas 1: $Card(\mathcal{I}) = m < \infty$. La statistique de test sera dans ce cas

$$\hat{T}_n = \sum_{|k| \in \mathcal{I}} |\hat{c}_k - c_k^0|^2.$$

Cas 2: $Card(\mathcal{I}) = \infty$. On considère dans ce cas une suite croissante d'entiers $p = p(n)$ telle que $\lim_{n \to \infty} p(n) = \infty$ et on pose $\mathcal{I}_p = \{i_1 < i_2 < \cdots < i_p\}$. La statistique de test est alors dans ce cas $\hat{T}_{n,p} = \sum_{|k| \in \mathcal{I}_p} |\hat{c}_k - c_k^0|^2$.

Dans chacun des cas, l'hypothèse H_0 est rejetée si $\hat{T}_n > t_\alpha$ (resp. $\hat{T}_{n,p} > t'_\alpha$) où t_α (resp. t'_α) est un nombre réel positif déterminé par le niveau α du test.

Nos résultats principaux donnent la loi asymptotique de $\hat{T}_n$ (resp. $\hat{T}_{n,p}$) et permettent donc d'obtenir une valeur asymptotique de t_α (resp. t'_α).

Les hypothèses sont:

Hypothèses

• (C1): f est höldérienne d'ordre δ, avec $\frac{1}{2} < \delta \leq 1$, i.e. il existe une constante M telle que $|f(s) - f(t)| \leq M|s - t|^\delta$, pour tout $s, t \in [0, 1]$.

• (C2): $\varepsilon_{j,n}$, $j = 1, \ldots, n$, sont des variables aléatoires réelles i.i.d. d'espérance nulle et de variance inconnue σ^2.

La convergence en loi est notée: $\xrightarrow{\mathcal{L}}$.

Résultats principaux

Quand $\mathcal{I}$ est fini, on obtient le résultat suivant:

Théorème 0.2.1 . *Si* $Card(\mathcal{I}) = m < \infty$, *on a:*

$$\frac{n}{\sigma^2} \sum_{|k| \in \mathcal{I}} |\hat{c}_k - c_k|^2 \xrightarrow{\mathcal{L}} \chi^2(s) \quad quand \quad n \to +\infty$$

où $s = 2m - 1$ *si* $0 \in \mathcal{I}$ *et* $s = 2m$ *si* $0 \notin \mathcal{I}$.

Dans le cas $Card(\mathcal{I}) = \infty$ et $\mathcal{I}_p = \{i_1 < i_2 < \cdots < i_p\}$ où $p = p(n)$ est une suite croissante de limite infinie, on introduit les hypothèses supplémentaires suivantes:

- (A1): $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$.

- (A2): $\lim_{n \to +\infty} \{n^{-2\delta+1} p(n)\} = 0$.

On obtient alors:

Théorème 0.2.2 . *Si* $Card(\mathcal{I}) = \infty$ *et si* (A1) *et* (A2) *sont vérifiées, alors*

$$\frac{n \sum_{|k| \in \mathcal{I}_p} |\hat{c}_k - c_k|^2 - u_p \sigma^2}{\sigma^2 \sqrt{2u_p}} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

où $u_p = 2p - 1$ *si* $0 \in \mathcal{I}$ *et* $u_p = 2p$ *si* $0 \notin \mathcal{I}$.

Puisque la région de rejet est définie par $\hat{T}_n = \sum_{|k| \in \mathcal{I}} |\hat{c}_k - c_k^0|^2 > t_\alpha$ si $Card(\mathcal{I}) < \infty$ ou $\hat{T}_{n,p} = \sum_{|k| \in \mathcal{I}_p} |\hat{c}_k - c_k^0|^2 > t'_\alpha$ si $Card(\mathcal{I}) = \infty$, les théorèmes 0.2.1 et 0.2.2 permettent de déterminer le niveau et la puissance du test lorsque la variance est connue. Le théorème 0.2.1 est un résultat paramétrique alors que le théorème 0.2.2 est non paramétrique. La question qui se pose dans le cas non paramétrique est de savoir quelles sont les alternatives proches de H_0 qui peuvent être distinguées de l'hypothèse nulle. En suivant la démarche de Eubank et Spiegelmann (1990), on considère les alternatives locales $c_k = c_k^0 + h(n) c_k^1$, $k \in \mathcal{I}$ où $\lim_{n \to \infty} h(n) = 0$ et c_k^1 est le k-ième coefficient de Fourier d'une fonction g höldérienne d'ordre $\delta > \frac{1}{2}$; on obtient:

Proposition 0.2.1 . *On se place sous les hypothèses du théorème* 0.2.2 *et on considère* $h(n) = p^{1/4} n^{-1/2}$. *On a alors sous les alternatives locales* $c_k = c_k^0 + h(n) c_k^1$, $k \in \mathcal{I}$,

$$\frac{n \sum_{|k| \in \mathcal{I}_p} |\hat{c}_k - c_k^0|^2 - u_p \sigma^2}{\sigma^2 \sqrt{2u_p}} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(\frac{1}{2\sigma^2} \sum_{|k| \in \mathcal{I}} |c_k^1|^2, 1 \right)$$

où $u_p = 2p - 1$ *si* $0 \in \mathcal{I}$ *et* $u_p = 2p$ *si* $0 \notin \mathcal{I}$.

La proposition 0.2.1 signifie que le test peut détecter les alternatives locales convergeant vers l'hypothèse nulle avec une vitesse inférieure ou égale à $p^{1/4} n^{-1/2}$.

Construction du test

Les résultats des théorèmes 0.2.1 et 0.2.2 permettent de construire le test quand la variance σ^2 est connue; cependant, en pratique, σ^2 est inconnue et il faut donc l'estimer. On peut utiliser l'estimateur de Gasser, Sroka et Jennen-Steinmetz (1986),

$$\hat{\sigma}_1^2 = \frac{2}{3(n-2)} \sum_{j=2}^{n-1} \left(\frac{1}{2} Y_{j-1} + \frac{1}{2} Y_{j+1} - Y_j \right)^2 .$$

Gasser, Sroka et Jennen-Steinmetz (1986) montrent que sous les hypothèses (C1) et (C2), $\hat{\sigma}_1^2$ converge vers σ^2 à la vitesse $n^{-1/2}$; on peut alors facilement montrer que les théorèmes 0.2.1 et 0.2.2 restent vrais en remplaçant σ^2 par $\hat{\sigma}_1^2$.

On peut aussi considérer un estimateur $\hat{\sigma}_2^2$ de σ^2 qui converge sous H_0. Plus précisément, soit

$$\hat{\sigma}_2^2 = \frac{1}{n} \sum_{j=1}^{n} |Y_j - \hat{f}(j/n)|^2$$

où $\hat{f}$ est défini de la manière suivante.

Notons J le complémentaire de $\mathcal{I}$ dans $\mathbb{N}$;

si $Card(J) = \nu < \infty$, on pose

$$\hat{f}(t) = \sum_{|k| \in \mathcal{I}} c_k^0 e^{2\pi i k t} + \sum_{|k| \in J} \hat{c}_k e^{2\pi i k t} ;$$

si $Card(J) = \infty$, on considère une suite croissante d'entiers $q = q(n)$ telle que $\lim_{n \to \infty} q(n) = +\infty$ et on pose

$$\hat{f}(t) = \sum_{|k| \in \mathcal{I}} c_k^0 e^{2\pi i k t} + \sum_{|k| \leq q, |k| \in J} \hat{c}_k e^{2\pi i k t} .$$

Introduisons les hypothèses suivantes:

- (A3): $\lim_{n \to \infty} \{n^{-1/2} q(n)\} = 0$ et $\sum_{k \in \mathbb{Z}} |c_k| < \infty$.

Bibliographie

[1] Dette, H., and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 778-800.

[2] Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression using nonparametric via order selection criteria. *Ann. Stat.*, **20**, 1412-1425.

[3] Eubank, R. L. and Spiegelmann, C. H. (1990). Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Assoc.* **85**, 410, 387-392.

[4] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, **21**, 1926-1947.

Tests d'hypothèses linéaires dans un modèle de régression non paramétrique

Zaher MOHDEB

RESUME: Cette thèse est consacrée à la construction de tests d'hypothèses sur la fonction de régression f, d'un modèle de régression non paramétrique. Dans une première partie, on construit des tests d'hypothèses sur les coefficients de Fourier de f. De tels tests peuvent être utilisés pour comparer deux signaux bruités dans une bande donnée de fréquences. Les statistiques de test que nous utilisons, s'expriment en fonction des coefficients de Fourier empiriques de f. La deuxième partie porte sur le test de l'hypothèse "f est un élément de E" où E est un espace vectoriel de dimension finie. Nous proposons deux statistiques de test $\widehat{R}_n^2$ et $\widehat{\mathcal{M}}_n^2$ basées sur deux approximations différentes de la distance dans L^2. La première est obtenue en estimant cette distance par la distance empirique des observations à l'espace E. La seconde est construite à l'aide des observations convenablement corrigées. Dans cette partie, nous supposons que les fonctions considérées sont höldériennes d'ordre strictement plus grand que $1/2$ et nous obtenons le comportement asymptotique en loi de chacune des deux statistiques proposées. La troisième partie est une extension de la deuxième au cas où les fonctions sont Riemann-intégrables; le comportement en loi de la statistique $\widehat{R}_n^2$ est alors sensiblement différent de celui obtenu dans la partie précédente, puisque l'on constate, dans le résultat limite, l'apparition d'un terme non négligeable. Cependant, ce terme supplémentaire est explicite et permet donc la construction de différents tests.

MOTS CLES: Coefficients de Fourier empiriques; Hypothèse linéaire; Régression non linéaire; Régression non paramétrique; Test non paramétrique.