Tests d'hypothèses dans un modèle de régression non paramétrique

Zaher Mohdeb

Tests d'hypothèses dans un modèle de régression non paramétrique

Zaher Mohdeb

Université Mentouri
Département de Mathématiques,
Constantine, Algérie
E-mail: z.mohdeb@gmail.com

Résumé. On considère le modèle de régression non paramétrique de fonction de régression f. Une procédure de test d'hypothèse sur les coefficients de Fourier de f est proposée. On obtient le comportement asymptotique de la statistique de test proposée, on a donc ainsi le niveau et la puissance asymptotique du test. De tels tests peuvent, en particulier, être utilisés pour comparer deux signaux bruités dans une bande de fréquence. Un autre exemple est le test de l'hypothèse "f est un polynôme trigonométrique". Une étude par simulation a été menée, pour des petites tailles d'échantillon, afin de montrer la performance du test proposée.

Mots clés: Modèle de régression non linéaire, Coefficient de Fourier empirique, Test non paramétrique.

Abstract. We consider the nonparametric regression model with regression function f. A procedure for testing hypothesis on the Fourier coefficients of f is proposed. We obtain the asymptotic weak behaviour of the proposed test, then we have the level and the asymptotic power of the test. Such tests can be used in particular to compare two noisy signals in a frequency band. Another example is the test of the hypothesis that "f is a trogonometric polynomial". A simulation study is conducted, for small sample size, to demonstrate the performance the proposed test.

Key words: Nonlinear regression model, Empirical Fourier coefficient, Nonparametric test.

1 Introduction

On considère le modèle de régression non paramétrique

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

où $t_{j,n} = j/n$, $f : [0,1] \longrightarrow \mathbb{R}$ est une fonction inconnue et $\varepsilon_{j,n}$, $j = 1, \ldots, n$, forment un tableau triangulaire de variables aléatoires de variables aléatoires centrées et de variance

finie σ^2 et pour tout n les variables aléatoires $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ sont indépendantes.

L'objet de ce travail est de construire des tests d'hypothèses sur les coefficients de Fourier de f. Plus précisément, soit $c_k, \quad k \in \mathbb{Z}$, les coefficients de Fourier de f et soit $\mathcal{I} = \{i_1 < i_2 < \cdots < i_r < \cdots\}$ un sous-ensemble de $\mathbb{N}$, avec $Card(\mathcal{I}) = \infty$. On veut construire un test de l'hypothèse nulle

$$H_0 : \ c_k = 0 \ \ \forall k \in \mathcal{I} \ \text{ contre l'hypothèse } \quad H_1 : \ \exists k \in \mathcal{I} \ \ c_k \neq 0 \, . \tag{2}$$

L'hypothèse H_0 est équivalente à $c_k = 0$, $|k| \in \mathcal{I}$ et recouvre de nombreuses situations, par exemple le test de l'hypothèse $f \equiv 0$, est le test de H_0 avec $\mathcal{I} = \mathbb{N}$; 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 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 $c_k = d_k - e_k$, dans le modèle $Y = U - V = (g-h) + \xi$ où $\xi = \varepsilon - \eta$.

Dans le test de l'hypothèse (2), il est clair que H_0 est vraie si et seulement si $\sum_{|k| \in \mathcal{I}} |c_k|^2 = 0$. Le test sera donc basé sur une estimation de cette quantité. On commence par estimer c_k par l'estimation empirique $\hat{c}_k = \frac{1}{n} \sum_{j=1}^{n} Y_{j,n} e^{-2i\pi kj/n}$. Ensuite on considère une suite d'entiers $p = p(n)$ telle que $\lim\limits_{n \to \infty} p(n) = \infty$ et on pose $\mathcal{I}_p = \{i_1 < i_2 < \cdots < i_p\}$. La statistique de test est alors $\hat{T}_{n,p} = \sum_{|k| \in \mathcal{I}_p} |\hat{c}_k|^2$.

L'hypothèse H_0 est rejetée si $\hat{T}_{n,p} > t_\alpha$ où t_α est un nombre réel positif déterminé par le niveau α du test.

Nos résultats principaux énoncés ci-dessous donnent la loi asymptotique de $\hat{T}_{n,p}$ et permettent donc d'obtenir une valeur asymptotique de t_α.

La mise en oeuvre du test nécessite la connaissance de σ^2, ce qui, en pratique, n'est jamais le cas; on a donc besoin d'un estimateur. On peut utiliser l'estimateur de Rice (1984); on propose également un estimateur adapté à l'hypothèse nulle et on donne les conditions qui permettent d'utiliser cet estimateur. Enfin on peut comparer sur des simulations les effets de ces estimateurs sur la puissance empiriques du test pour les petits échantillons ($n = 100$).

La majeure partie de la littérature sur le problème des tests d'hypothèses dans le modèle (1) est basée sur la méthode des splines. L'usage des coefficients de Fourier empiriques de f pour construire des tests d'hypothèses dans le modèle (1) est abordé

par Eubank et Spiegelman (1990) et Mohdeb et Mokkadem (2001). Eubank et Spiegelman (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. Mohdeb et Mokkadem (2001) construisent des tests d'hypothèses sur les coefficients de Fourier de f dans une bande de fréquences donnée. Jayasuriya (1996) généralise l'approche de Eubank et Spiegelman (1990) pour tester l'hypothèse: f est un polynôme, dans le cas d'un modèle non normal. Cox et Koh (1989) donnent un test de l'hypothèse f est un polynôme de degré inférieur à m. Jayasuriya (1996) généralise l'approche de Eubank et Spiegelman (1990) pour tester l'hypothèse: f est un polynôme, dans le cas d'un modèle non normal.

2 Hypothèses et résultats

On s'intéresse au test (2) pour le modèle (1). Dorénavant, on suppose que

• (C1): f est Lipschitzienne 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): Pour tout entier n, $\varepsilon_{j,n}$, $j = 1, \ldots, n$, sont des variables aléatoires réelles i.i.d. d'espérance nulle et de variance inconnue σ^2.

La convergence en loi est notée par $\xrightarrow{\mathcal{L}}$. On note aussi $c_k = \int_0^1 e^{-2\pi ikt} f(t)\, dt$, $k \in \mathbb{Z}$ les coefficients de Fourier de f et $\mathcal{I} = \{i_1 < i_2 < \cdots < i_r < \cdots\}$ un sous-ensemble de $\mathbb{N}$, avec $Card(\mathcal{I}) = \infty$.

2.1 Résultats principaux

On considère $\mathcal{I}_p = \{i_1 < i_2 < \cdots < i_p\}$ où $p = p(n)$ est une suite croissante de limite infinie. On introduit les hypothèses suivantes:

• (A1): Pour tout entier n, $\varepsilon_{1,n} \sim \mathcal{N}(0, \sigma^2)$.

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

On a alors le théorème suivant.

Théorème 1 . *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 confiance est définie par $\hat{T}_{n,p} = \sum_{|k|\in\mathcal{I}_p} |\hat{c}_k|^2 > t'_\alpha$, le théorème précédent permet de déterminer le niveau et la puissance du test lorsque la variance est connue. 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. Suivant Eubank et Spiegelman (1990), on considère les alternatives locales $c_k = h(n)c_k^1$, $k \in \mathcal{I}$ où $\lim_{n\to\infty} h(n) = 0$ et c_k^1 est le k-ième coefficient de Fourier d'une fonction g Lipschitzienne d'ordre $\delta > \frac{1}{2}$, on obtient:

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

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

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

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

2.2 Construction du test

Le résultats du théorème 1 permet 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 Rice (1984), défini par

$$\widehat{\sigma}_1^2 = \frac{1}{2(n-1)}\sum_{i=2}^{n}(Y_{i,n} - Y_{i-1,n})^2 .$$

On peut montrer que le théorème 1 reste vrai 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

$$\hat{\sigma}_2^2 = \frac{1}{n}\sum_{j=1}^{n} |Y_{j,n} - \hat{f}(j/n)|^2 \tag{3}$$

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

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

$$\hat{f}(t) = \begin{cases} \sum_{|k|\in J} \hat{c}_k e^{2i\pi kt} & \text{si} \quad Card(J) < \infty , \\ \sum_{|k|\le q, |k|\in J} \hat{c}_k e^{2i\pi kt} & \text{si} \quad Card(J) = \infty . \end{cases} \tag{4}$$

Quand J est fini, on a le résultat suivant.

Corollaire 1 . *Si* $Card(J) < \infty$ *et si les hypothèses du théorème* 1 *sont vérifiées, alors sous* H_0*, on a:*

$$\frac{n\sum_{|k|\in\mathcal{I}_p}|\hat{c}_k|^2 - u_p\hat{\sigma}_2^2}{\hat{\sigma}_2^2\sqrt{2u_p}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

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

Quand J est infini, il nous faut introduire les hypothèses:

- (A3): $\lim_{n\to\infty}\{\sqrt{p(n)}\sum_{|k|>q(n)}|c_k|\} = 0.$
- (A4): $\lim_{n\to\infty}\{n^{-1}p(n)q^2(n)\} = 0.$

On a alors le corollaire suivant.

Corollaire 2 . *Si* $Card(J) = \infty$ *et si les hypothèses* $(A1)-(A4)$ *sont satisfaites, alors sous* H_0*, on a:*

$$\frac{n\sum_{|k|\in\mathcal{I}_p}|\hat{c}_k|^2 - u_p\hat{\sigma}_2^2}{\hat{\sigma}_2^2\sqrt{2u_p}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

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

Il apparait clairement que dans le cas où $Card(J) = \infty$, l'utilisation de $\hat{\sigma}_2^2$ exige plus d'hypothèses; il peut donc être préférable d'utiliser $\hat{\sigma}_1^2$. Par contre, dans le cas où $Card(J) < \infty$, l'utilisation de $\hat{\sigma}_2^2$ ne demande pas plus d'hypothèses.

Bibliographie

[1] Cox, D. and Koh, E. (1989), A Smoothing Spline Based Test of Model Adequacy in Polynomial Regression. *Ann. Inst. Statist. Math.,* **41**, 2, 383-400.

[2] Eubank, R. L. and Spiegelman, C. H. (1990), Testing the Goodness-of-Fit of a Linear Model Via Nonparametric Regression Techniques. *J. Amer. Statist. Assoc.* **85**, 410, 387-392.

[3] Jayasuriya, B. R. (1996), Testing for Polynomial Regression Using Nonparametric Regression Techniques. *J. Amer. Statis. Assoc.* **91**, 436, 1626-1631.

[4] Mohdeb, Z. and Mokkadem, A. (2001). Testing Hypotheses on Fourier Coefficients in Nonparametric Regression Model. *Journal of Nonparametric Statistics*, **13**, 605-629.

[5] Rice, J. (1984). Bandwidth Choice for Nonparametric Regression. *Ann. Stat.* **12**, 1215-1230.