

TEST DE COMPARAISON DE DEUX MODÈLES DE RÉGRESSION NON PARAMÉTRIQUES BASÉ SUR LES COEFFICIENTS DE FOURIER

Zaher Mohdeb

*Université Mentouri
Département de Mathématiques,
Constantine, Algérie
E-mail: zaher.mohdeb@umc.edu.dz*

Résumé. Dans ce travail, nous proposons une nouvelle méthode de comparaison de deux fonctions de régression f_1 et f_2 dans le cas homoscédastique et un échantillonnage fixé. Notre approche est basée sur les coefficients de Fourier empiriques des fonctions de régression respectivement de f_1 et f_2 . On obtient la distribution asymptotique de la statistique de test proposée, sous l'hypothèse nulle " $f_1 = f_2$ " ainsi que sous les alternatives globales et locales. Une étude par simulation est menée pour montrer la performance du test proposé.

Mots-clés. Régression non paramétrique, test d'égalité, coefficient de Fourier empirique.

Abstract. In this work we propose a new methodology for the comparison of two regression functions f_1 and f_2 in the case of homoscedastic error structure and a fixed design. Our approach is based on the empirical Fourier coefficients of the regression functions f_1 and f_2 respectively. We obtain the asymptotic distribution of the test statistic under the null hypothesis " $f_1 = f_2$ " and local and global alternatives. A simulation study is conducted to investigate the finite sample performance of our test.

Keywords. Nonparametric regression, test of equality, empirical Fourier coefficient.

1 Introduction

On considère les deux modèles de régression non paramétriques donnés par

$$Y_{l,j} = f_l(t_j) + \varepsilon_{l,j}, \quad l = 1, 2 \text{ et } j = 1, \dots, n, \quad (1)$$

où $t_j = j/n$; pour $l = 1, 2$, $f_l: [0, 1] \rightarrow \mathbb{R}$, est une fonction réelle inconnue et les erreurs $(\varepsilon_{l,j})_{j=1,\dots,n}$, sont des variables aléatoires i.i.d. centrées et de variance finie σ_l^2 . En outre, les erreurs aléatoires $(\varepsilon_{l,j})_{j=1,\dots,n}$, $l = 1, 2$ sont supposées indépendantes. L'objet de ce travail est de construire le test d'hypothèse d'égalité des fonctions de régression f_1 et f_2 , c'est-à-dire:

$$H_0 : f_1 = f_2 \quad \text{contre} \quad H_1 : f_1 \neq f_2. \quad (2)$$

Le problème de test d'égalité de fonctions de régression a fait l'objet de plusieurs travaux dans la littérature. On citera, sans être exhaustif, Härdle et Marron (1990), Delgado (1993), Bowman et Young (1996), Hall, Huber et Speckman (1997), Kulasekera et Wang (1998), Munk et Dette (1998), Mohdeb et Mokkadem (2001), Neumeyer et Dette (2003), Vilar-Fernández et González-Mantega (2003), Vilar-Fernández, Vilar-Fernández et González-Mantega (2007). L'objet de notre travail porte sur la construction du test d'hypothèses (2) basé sur les coefficients de Fourier des fonctions de régression f_1 et f_2 . Plus précisément, soient

$$c_{l,k} = \int_0^1 e^{-2\pi i k t} f_l(t) dt, \quad l = 1, 2 \text{ et } k \in \mathbb{Z}$$

les coefficients de Fourier de f_l , $l = 1, 2$. Le test (2) se ramène au test suivant

$$H_0 : c_{1,k} = c_{2,k}, \forall k \in \mathbb{Z} \quad \text{contre} \quad H_1 : \exists k \in \mathbb{Z} \text{ tel que } c_{1,k} \neq c_{2,k}. \quad (3)$$

Comme f_l , $l = 1, 2$, est à valeurs réelles, on a $c_{l,k} = \bar{c}_{l,-k}$ pour tout $k \in \mathbb{N}$; par conséquent tout test sur les coefficients de Fourier de f_l se ramène au test sur les $c_{l,k}$, $k \geq 0$. En particulier, l'hypothèse nulle H_0 est équivalente à $c_{1,k} = c_{2,k}$, $|k| \in \mathbb{N}$. Il s'ensuit, clairement, que H_0 est vraie si et seulement si $\sum_{|k| \in \mathbb{N}} |c_{1,k} - c_{2,k}|^2 = 0$.

La statistique de test, que nous utilisons, est basée sur les estimateurs empiriques $\hat{c}_{l,k}$ de $c_{l,k}$ définis par

$$\hat{c}_{l,k} = \frac{1}{n} \sum_{j=1}^n Y_{l,j} e^{-2\pi i k j / n}, \quad l = 1, 2$$

et on considère une suite de nombres entiers $p = p(n)$ tels que $\lim_{n \rightarrow \infty} p(n) = \infty$. La statistique de test est alors basée sur

$$\hat{T}_{n,p} = \sum_{|k| \leq p} |\hat{c}_{1,k} - \hat{c}_{2,k}|^2$$

et on rejette H_0 si $\hat{T}_{n,p} > t_\alpha$ où t_α est un nombre réel positif déterminé par le niveau α du test.

Dans notre approche, la statistique de test est utilisée pour tester des hypothèses sur les coefficients de Fourier de f_l , $l = 1, 2$. Pour construire le test, nous avons besoin d'estimer la variance σ_l^2 , $l = 1, 2$. Dans le présent travail, nous utilisons l'estimateur proposé par Rice (1984). Une étude sur des simulations est menée pour montrer la performance du test proposé.

2 Hypothèses et résultats

On s'intéresse au test (3) pour le modèle (1) et on suppose que pour $l = 1, 2$

- f_l est Lipschitzienne d'ordre δ , avec $\frac{1}{2} < \delta \leq 1$.
- $\varepsilon_{l,j}$, $j = 1, \dots, n$ sont des variables aléatoires réelles i.i.d. d'espérance nulle et de variance inconnue σ_l^2 .

La convergence en loi est notée par $\xrightarrow{\mathcal{L}}$. On introduit une suite d'entiers $p = p(n)$ telle que $\lim_{n \rightarrow +\infty} p(n) = +\infty$.

2.1 Comportement asymptotique des coefficients de Fourier empiriques

On introduit les hypothèses suivantes,

- (A1): $\varepsilon_{l,1} \sim \mathcal{N}(0, \sigma_l^2)$, $l = 1, 2$.
- (A2): $\lim_{n \rightarrow +\infty} \{n^{-2\delta+1}p(n)\} = 0$.

Théorème 1 *Si (A1) et (A2) sont vérifiées, alors*

$$\frac{n \sum_{|k| \leq p} |(\hat{c}_{1,k} - \hat{c}_{2,k}) - (c_{1,k} - c_{2,k})|^2 - (2p+1)(\sigma_1^2 + \sigma_2^2)}{(\sigma_1^2 + \sigma_2^2)\sqrt{2(2p+1)}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Puisque la région de rejet est définie par $\hat{T}_{n,p} = \sum_{|k| \leq p} |\hat{c}_{1,k} - \hat{c}_{2,k}|^2 > t_\alpha$, le théorème précédent permet de déterminer le niveau et la puissance du test lorsque la variance des erreurs 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 $c_{1,k} - c_{2,k} = h(n)e_k$, $|k| \leq p$, où $\lim_{n \rightarrow \infty} h(n) = 0$ et e_k est le k -ième coefficient de Fourier d'une fonction m 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 $c_{1,k} - c_{2,k} = h(n)e_k$, $|k| \leq p$,*

$$\frac{n \sum_{|k| \leq p} |\hat{c}_{1,k} - \hat{c}_{2,k}|^2 - (2p+1)(\sigma_1^2 + \sigma_2^2)}{(\sigma_1^2 + \sigma_2^2)\sqrt{2(2p+1)}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}\left(\frac{1}{2(\sigma_1^2 + \sigma_2^2)} \sum_{|k| \in \mathbb{N}} |e_k|^2, 1\right).$$

Ce qui signifie que le test peut détecter les alternatives locales convergeant vers l'hypothèse nulle avec une vitesse plus petite que $p^{1/4}n^{-1/2}$.

2.2 Construction du test

Le résultat du théorème 1 permet de construire le test quand les variances σ_l^2 , $l = 1, 2$, sont connues; cependant en pratique les σ_l^2 sont inconnues et il faut donc les estimer. On peut utiliser l'estimateur donné par Gasser, Sroka et Jennen-Steinmetz (1986); dans le présent travail, on utilise l'estimateur proposé par Rice (1984), défini par

$$\hat{\sigma}_l^2 = \frac{1}{2(n-1)} \sum_{j=2}^n (Y_{l,j} - Y_{l,j-1})^2, \quad l = 1, 2. \quad (4)$$

On peut montrer que le théorème 1 reste vrai en remplaçant σ_l^2 par $\hat{\sigma}_l^2$, $l = 1, 2$.

Corollaire 1 *Si (A1) et (A2) sont vérifiées, alors sous $H_0 : f_1 = f_2$, on a :*

$$\frac{n \sum_{|k| \leq p} |\hat{c}_{1,k} - \hat{c}_{2,k}|^2 - (2p+1)(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)}{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) \sqrt{2(2p+1)}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Ces résultats donnent le niveau de signification lorsque les fonctions de régression sont Lipschitziennes sous l'hypothèse nulle, ainsi que la puissance du test pour les alternatives également Lipschitziennes.

3 Simulation

Une étude par simulation est menée afin de montrer la performance du test proposé. On étudie les puissances empiriques du test d'hypothèses (2), avec $f_1(t) = t$ et deux formes différentes respectivement pour les alternatives f_2 : $f_2^{(1)}(t) = \beta t e^{-2t}$ et $f_2^{(2)}(t) = \beta t^2$, pour plusieurs choix de β sur l'intervalle $[0, 2]$.

Bibliographie

- [1] Bowman A. and Young, S. (1996). Graphical comparison of nonparametric curves. *Appl. Statist.*, **45**, 83-98.
- [2] Delgado, M. A. (1993). Testing the equality of nonparametric regression curves, *Statist. Probab. Lett.*, **17**, 199-204.
- [3] Eubank R. L. and Spiegelman, 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] Gasser, T., Sroka L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625-633.

- [5] Hall, P., Huber, C. and Speckman, P. L. (1997). Covariate-matched one-sided tests for the difference between functional means. *J. Amer. Statist. Assoc.*, **92**, 1074-1083.
- [6] Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *Ann. Statist.*, **18**, 63-89.
- [7] Kulasekera K. B. and Wang, J. (1998). Bandwidth selection for power optimality in a test of equality of regression curves. *Stat. Probab. Lett.*, **37**, 287-293.
- [8] Mohdeb, Z. and Mekkadem, A. (2001). Testing Hypotheses On Fourier Coefficients in Nonparametric Regression Model. *Journal of Nonparametric Statistics*, **13**, 605-629.
- [9] Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: Exact and asymptotic theory. *Ann. Statist.*, **26**, 2339-2368.
- [10] Neumeyer N. and Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *Ann. Statist.*, **31**, 3, 880-920.
- [11] Rice, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215-1230.
- [12] Vilar-Fernández J. M. and González-Manteiga, W. (2003). Nonparametric comparison of curves with dependent errors. *Statistics*, **38**, 2, 81-99.
- [13] Vilar-Fernández, J. M., Vilar-Fernández J. A. and González-Manteiga, W. (2007). Bootstrap tests for nonparametric comparison of regression curves with dependent errors. *Test*, **16**, 123-144.