C. R. Acad. Sci. Paris, Ser. I 351 (2013) 143–148

Contents lists available at SciVerse ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistics

Additivity test on the nonlinear part in partially linear models

Test d'additivité de la partie non linéaire dans les modèles partiellement linéaires

Khalid Chokri [a], Djamal Louani [a,b]

[a] *LSTA, université Paris-6, 175, rue du Chevaleret, 75013 Paris, France*

[b] *Université de Reims–Champagne–Ardennes, BP 1039, 51687 Reims cedex 2, France*

ARTICLE INFO

Article history:
Received 20 December 2011
Accepted after revision 31 January 2013
Available online 16 February 2013

Presented by Paul Deheuvels

ABSTRACT

Our aim in this Note is to build an additivity test relative to the nonlinear part of the partially linear model. More precisely, considering the model of the form $Y = \mathbf{Z}^\top \beta + m(\mathbf{X}) + \varepsilon$, where $m(\mathbf{X})$ stands as the nonlinear part and $\mathbf{X}$ is a random vector taking values in the space $\mathbb{R}^d$, our goal is to construct a testing procedure that allows us to test the validity of the hypothesis according to which the function m may be written with the shape $m(\mathbf{x}) = \sum_{l=1}^d m_l(x_l)$, where the m_l's are functions defined on $\mathbb{R}$.

RÉSUMÉ

Cette Note a pour objet de construire un test d'additivité de la partie non linéaire du modèle partiellement linéaire. Plus précisément, en considérant un modèle de la forme $Y = \mathbf{Z}^\top \beta + m(\mathbf{X}) + \varepsilon$, où $m(\mathbf{X})$ correspond à la partie non linéaire et $\mathbf{X}$ est un vecteur aléatoire prenant ses valeurs dans $\mathbb{R}^d$, il s'agit alors de construire une procédure de test permettant de vérifier si l'hypothèse selon laquelle la fonction m peut s'écrire sous la forme $m(\mathbf{x}) = \sum_{l=1}^d m_l(x_l)$, où les m_i sont des fonctions définie sur $\mathbb{R}$, est vraisemblable.

Version française abrégée

Soit $(\mathbf{X}_i, Y_i, \mathbf{Z}_i)_{i\geqslant 1}$ une suite de répliques indépendantes d'un vecteur aléatoire $(\mathbf{X}, Y, \mathbf{Z})$ à valeurs dans $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p$. Il est bien connu que les méthodes non paramétriques permettent d'éliminer les biais de modélisation dans les problèmes de régression en construisant les modèles directement à partir des données. Cependant, ces méthodes sont confrontées au fléau de la dimension dans le cas multivarié caractérisé par le fait que les vitesses de convergence des estimateurs sont des fonctions décroissantes de la dimension des covariables. Une solution à cette problématique a été apportée par Stone [12], qui a introduit les modèles de regression additifs.

Les modèles partiellement linéaires permettent d'allier les techniques paramétriques aux méthodes non paramétriques pour la modélisation des données comportant une partie linéaire combinée à une partie non linéaire et se présentant sous la forme :

$$Y = \mathbf{Z}^\top \beta + m(\mathbf{X}) + \varepsilon,$$

où $\beta \in \mathbb{R}^p$ est un paramètre vectoriel inconnu, m désigne une fonction non linéaire et ε est l'erreur de modélisation, dont la variance est notée σ_ε^2. Ici, $\mathbf{Z}^\top$ indique le transposé du vecteur $\mathbf{Z}$. L'estimation par des méthodes non paramétriques de

E-mail addresses: khalid.chokri@etu.upmc.fr (K. Chokri), djamal.louani@upmc.fr (D. Louani).

1631-073X/$ – see front matter
http://dx.doi.org/10.1016/j.crma.2013.01.016

la partie non linéaire fait que, pour ce type de modèle, on est aussi confronté au fléau de la dimension. Pour pallier ce problème, nous considérons une structure additive de la fonction de régression m et nous obtenons un nouveau modèle sous la forme :

$$Y = \mathbf{Z}^\top \beta + \sum_{l=1}^{d} m_l(X_l) + \varepsilon =: \mathbf{Z}^\top \beta + m_{\text{add}}(\mathbf{X}) + \varepsilon,$$

où X_l est le l^{e} composante du vecteur $\mathbf{X}$ et m_l est une fonction réelle univariée. Notons qu'une importante revue bibliographique sur les modèles partiellement linéaires est donnée dans le travail de Chokri et Louani [3]. La question qui se pose alors est de savoir si la structure additive de la partie non linéaire du modèle peut être considérée comme vraisemblable. Cette Note propose de construire une procédure de test pour tester l'hypothèse H_0 : « $m \in \mathcal{M}$ » contre l'alternative H_1 : « $m \notin \mathcal{M}$ », où la classe de fonctions $\mathcal{M}$ est définie plus bas dans l'assertion (1.3). Le test repose sur la statistique donnée dans l'assertion (1.4) et qui est la traduction estimée de la quantité $\mathbb{E}[\mathbb{E}(Y - \mathbf{Z}^\top \beta + m_{\text{add}}(\mathbf{X}))|\mathbf{X}]^2$, qui définit un écart à l'hypothèse nulle H_0 donnée par l'observation. Notons que la loi limite obtenue pour la statistique de test convenablement normalisée est la loi normale standard, qui permet alors de construire une région de rejet pour l'hypothèse nulle H_0.

1. Introduction

Let $(\mathbf{X}_i, Y_i, \mathbf{Z}_i)_{i \geqslant 1}$ be a sequence of i.i.d. copies of the $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p$-valued random vector $(\mathbf{X}, Y, \mathbf{Z})$. Denote by f its joint density function with respect to the Lebesgue measure and by $\mathbf{g}$ the marginal density associated with the random vector $\mathbf{X}$. Parametric regression models provide powerful tools for analyzing practical data when the models are correctly specified, but may suffer from large modeling biases when structures of the models are misspecified. As an alternative, nonparametric smoothing methods ease the concerns on modeling biases. However, nonparametric models are hampered by the so-called curse of dimensionality in multivariate settings, see Stone [12] for details. One of the methods for attenuating this difficulty is to model covariate effects via a partially linear structure, a combination of linear and nonlinear parts. This results in the partially linear regression models of the form:

$$Y = \mathbf{Z}^\top \beta + m(\mathbf{X}) + \varepsilon,$$

where $\beta \in \mathbb{R}^p$ is a vector of unknown parameters, m is the nonlinear part of the model and ε is the modeling error with variance σ_ε^2. Here, $\mathbf{Z}^\top$ stands as the transpose of the vector $\mathbf{Z}$.

To reduce the dimension impact of the nonparametric part in the partially linear regression model, we consider the additive structure of the regression function m and introduce the following model:

$$Y = \mathbf{Z}^\top \beta + \sum_{l=1}^{d} m_l(X_l) + \varepsilon =: \mathbf{Z}^\top \beta + m_{\text{add}}(\mathbf{X}) + \varepsilon, \tag{1.1}$$

where X_l is the l-th component of the vector $\mathbf{X}$ and m_j is a real univariate function. Subsequently, we have to estimate the unknown quantities, that is, the vector parameters β and the univariate functions $m_j, 1 \leqslant j \leqslant d$, as well. Note that a substantial bibliographical review on the partially linear models topic is given in Chokri and Louani [3]. The goal of this Note is to build a testing procedure that allows us to test the validity of the additivity hypothesis of the nonlinear part of the model.

There exist a number of testing procedures related to nonparametric regression models with their various forms. In partially linear regression models, various tests have been proposed in the literature to test nominal hypotheses in both the linear and nonlinear parts. We refer among others to the works of González-Manteiga and Aneiros-Pérez [8] and Aneiros-Pérez et al. [1]. Li et al. [10] proposed a linearity test in partially linear models. In the generalized nonparametric regression models with estimated parameters, Gozalo and Linton [9] developed several kernel-based tests testing an additivity hypothesis. Dette and von Lieres und Wilkau [6] investigated properties of various tests for the additive regression hypothesis in the common nonparametric regression models. Derbort et al. [5] proposed a simple consistent test of additivity in a multiple nonparametric regression model where data are observed on a lattice. Camlong-Viot [2] proposed a test statistics to test additivity hypotheses of the regression function for β-mixing data. In regression models on functional explanatory data, Delsol et al. [4] proposed a theoretical framework for structural testing procedures. Note that our work is specific to partially linear models and, on the light of the recent works of Müller et al. [11] and Ferraty et al. [7], may be adapted to build an additivity test procedure in regression models embedding a functional data part.

1.1. Construction of the test statistic

In order to set out our testing procedure, let us introduce first some notations. For any $1 \leqslant l \leqslant d$, set $\mathbf{x}_{-l} = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_d)$, $\mathbf{q}_{-l}(\mathbf{x}_{-l}) = \prod_{j=1, j \neq l}^{d} q_j(x_j)$ and $\mathbf{q}(\mathbf{x}) = \prod_{l=1}^{d} q_l(x_l)$, where q_l, $1 \leqslant l \leqslant d$, are univariate densities.