

Nonparametric and Semiparametric Estimation of Additive Models with both Discrete and Continuous Variables under Dependence

Christine Camlong-Viot

Faculté de Pharmacie
Département de Biomathématique
Université Paris-Sud
92296 Châtenay-Malabry, FRANCE

Juan M. Rodríguez-Póo

Departamento de Economía
Universidad de Cantabria
39005 Santander, SPAIN

Philippe Vieu

Laboratoire de Statistique et Probabilités
Université Paul Sabatier
31062 Toulouse, FRANCE

Abstract

This paper is concerned with the estimation and inference of nonparametric and semiparametric additive models in the presence of discrete variables and dependent observations. Among the different estimation procedures, the method introduced by Linton and Nielsen, based in marginal integration, has become quite popular because both its computational simplicity and the fact that it allows an asymptotic distribution theory. Here, an asymptotic treatment of the marginal integration estimator under different mixtures of continuous-discrete variables is offered, and furthermore, in the semiparametric partially additive setting, an estimator for the parametric part that is consistent and asymptotically efficient is proposed. The estimator is based in minimizing the L_2 distance between the additive nonparametric component and its correspondent linear direction. Finally, we present an application to show the feasibility of all methods introduced in the paper. ¹

Keywords: Additive Models, Dimension reduction techniques, semiparametric models, strong mixing conditions, marginal integration

AMS Codes: 62G10, 62G05, 62G20

¹This research was financially supported by The Dirección General de Investigación del Ministerio de Ciencia y Tecnología under research grant BEC2001-1121. The authors wish to express their gratitude to the participants of the STAPH (<http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html>) group in Functional Statistics in Toulouse for their many helpful comments and suggestions

1 Introduction

This paper addresses an old problem considered here from a rather different-new perspective. The problem of how to treat discrete variables in nonparametric regression problems is already well known in the statistical literature (see among others Hall, 1981; Bierens, 1983; Grund and Hall, 1993; and Ahmad and Cerrito, 1994). When the regressors are discrete no smoothing is required to obtain root-n consistent estimators. Furthermore, if any amount of smoothing is applied, then, the discrete components do not suffer from the curse of dimensionality.

In the econometrics literature, the same problem has been traditionally approached by retreating it as a semiparametric problem. That is, the continuous variables are introduced either in a multivariate or in an additive one-dimensional nonparametric regression setting whereas discrete regressors appear in the form of linear parametric functions. These are the so called partially linear models. In this setting, Delgado and Mora (1995) show that root-n consistency of the parametric part is achieved under much weaker conditions than in the continuous case (see Robinson, 1988).

In many cases (see Horowitz, 1998) the partially linear structure does not appear to be a reasonable restriction. Racine and Li (2000) analyze the case when discrete and continuous variables are mixed within a multivariate nonparametric regression function. They provide the statistical properties of the estimator and a method to choose the different bandwidths. However, the use of multivariate nonparametric regression models presents an important problem: When many explanatory variables are available, the rate at which nonparametric smoothers converge to their true values is very slow, and the introduction of additive restrictions is recommended (Stone, 1985). In Fan, Härdle and Mammen (1998) the impact of discrete regressors in the estimation of additive models is analyzed. They also consider as a particular case a semiparametric additive partially linear model, and provide root-n consistent estimators of the parameters of interest. Their method is based on local linear regression smoothers, as they allow for components that can be either discrete or continuous. However, their estimation procedure presents some drawbacks. First, they only give the statistical properties of the nonparametric additive components that depend on absolutely continuous regressors, second the resulting estimator for the nonparametric component is created by splitting the sample in several cells. The number of cells depends on the number of categories of the discrete variables, and therefore, if the number of cells is high each may not have enough observations to estimate. Finally, the whole analysis is performed under the assumption of independent and identically distributed observations. This assumption, typically rules out regression models that contain lagged endogenous variables as regressors.

This paper addresses the problem of introducing both discrete and continuous explanatory variables into an additive nonparametric (semiparametric) regression setting that accounts for dependent data. In order to estimate the additive components marginal integration techniques (Newey, 1994; Tjostheim and Auestad, 1994 and Linton and Nielsen, 1995) are used. Here, the pilot multivariate nonparametric regression estimator is computed by using kernel methods. Discrete covariates enter in the product kernel although no smoothing is applied to them. We show that estimators of the additive components with discrete covariates exhibit root-n rates and in the mixed case, that is, estimators of the additive components that depend both on continuous and discrete covariates, the rate of convergence is the same as in the continuous case.

Further if we assume that the additive components depending on discrete regressors fall within the class of linear parametric functions, a two step method to estimate the parametric part is proposed.

The estimator is based in minimizing the L_2 distance between the additive nonparametric component and its correspondent linear direction. It is root-n consistent and achieves the semiparametric efficiency bound. An important feature of our work is to consider a strongly dependent model that allows for applications in time series situations.

The remainder of the paper is organized as follows. The statistical model and the estimator are introduced in Section 2. Its asymptotic behavior is also treated in this section. In Section 3 we present a two step root-n consistent semiparametric estimator of the partially additive linear model. In Section 4 we present an application to the estimation of a wage equation for the Spanish Economy. Finally, in the Appendix we prove the main results of the paper.

2 Additive Nonparametric Regression

Along this section we consider an additive nonparametric regression model where a subset of explanatory variables is discrete and the remaining are continuous. More precisely, let $X^c = (X_1, X_3)$ be a vector of continuous random variables valued in $\mathbb{R}^{p_1+p_2}$ and $X^d = (X_2, X_4)$ be a vector of discrete random variables valued in $\mathbb{R}^{q_1+q_2}$. That is, that there exists $\mathcal{D} \in \mathbb{R}^{q_1+q_2}$ such that

$$(1) \quad P(X^d \in \mathcal{D}) = 1,$$

$$(2) \quad \forall x^c \in \mathcal{D}, \quad P(X^d = x^c) > 0.$$

Let $X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})$. We consider a nonparametric regression model given by

$$(3) \quad Y_i = m(X_i) + \epsilon_i = \omega + m_1(X_{1i}) + m_2(X_{2i}) + m_{34}(X_{3i}, X_{4i}) + \epsilon_i,$$

where $\{(X_i, Y_i)\}_{i=1}^n$ are observations from a stationary α -mixing process, $E(\epsilon | X_i) = 0$ and $m_1(\cdot)$, $m_2(\cdot)$ and $m_{34}(\cdot, \cdot)$ are of unknown form. For identification purposes $E[m_1(X_{1i})] = E[m_2(X_{2i})] = E[m_{34}(X_{3i}, X_{4i})] = 0$. Recall that the α -mixing coefficient relative to the process $\{(X_i, Y_i)\}_{i \geq 1}$ is defined, for any $s \in \mathbb{N}$, by

$$\alpha(s) = \sup \{|P(A \cap B) - P(A)P(B)|, A \in \mathcal{F}_1^r, B \in \mathcal{F}_{r+s}^\infty\}$$

where \mathcal{F}_1^r and \mathcal{F}_{r+s}^∞ are σ -fields generated respectively by $\{(X_1, Y_1), \dots, (X_r, Y_r)\}$ and $\{(X_{r+s}, Y_{r+s}), \dots\}$. A process is strongly mixing if

$$\lim_{s \rightarrow \infty} \alpha(s) = 0.$$

See Rosenblatt (1956) for more details. Note that this model nests a broad variety of different specifications. If we set $m_1 = m_2 = 0$ then we consider the same model as in Racine and Li (2000). On the other side, if $m_2 = 0$ then we have the model analyzed in Fan, Härdle and Mammen (1998). Of course in both cases all results were obtained for the independent case.

Our interest is to estimate the unknown quantities, that is $m_1(\cdot)$, $m_2(\cdot)$ and $m_{34}(\cdot, \cdot)$ in the regression model. So far, purely additive models have been estimated using the backfitting algorithm and the so called marginal integration techniques. The first method was proposed in Hastie and Tibshirani (1990) and the second was simultaneously developed in Newey (1994); Tjostheim and Auestad (1994) and Linton and Nielsen (1995). From the computational point of view both approaches appear equally feasible. The backfitting has been mostly implemented using splines. Stone, Hansen, Kooperberg and Truong (1997) develop estimation theory using polynomial spline methods and Wahba (1992) uses smoothing splines. Also local polynomial regression has been used as in

Opsomer and Ruppert (1994). For the marginal integration techniques, series estimators (Andrews and Whang, 1990; and Newey, 1995), local constant polynomials (Linton and Nielsen, 1995) and local linear polynomials (Fan, Härdle and Mammen, 1998) have been applied. From the theoretical point of view, although the behavior of the marginal integration estimators is known better, however, important developments have been made in the theory of backfitting (see Mammen, Linton and Nielsen, 1999; Opsomer and Ruppert, 1997; and Opsomer, 2000). In the context of dependent data, to our knowledge, no results are available for the backfitting estimator whereas marginal estimators have been studied in deep by Sperlich, Tjostheim and Yang (2000) and Camlong-Viot, Sarda and Vieu (2000). On these grounds, we opt to estimate the different unknown components by marginal integration techniques.

At this stage it is worth being fixed some notations. In the following, all the integrals related with continuous variables will be taken with respect to Lebesgue measure while all the integrals related with discrete variables will be taken with respect to the counting measure (the counting measure will be denoted by μ). In the following we will also make use of some functions

$$q(x) = q(x_1, x_2, x_3, x_4) = q_1(x_1)q_2(x_2)q_{34}(x_3, x_4),$$

where q_1 , q_2 and q_{34} are known density functions respectively defined on \mathbb{R}^{p_1} , \mathbb{R}^{q_1} and $\mathbb{R}^{p_2+q_2}$. Moreover, for any $\ell = 1, \dots, 4$ we will denote by f_ℓ the marginal density of X_ℓ (giving the fact that these marginal densities are either taken with respect to the Lebesgue measure for continuous X_ℓ or with respect to μ for discrete ones). Similarly, for any $\ell = 1, \dots, 4$ and for any $s > 0$ we will denote by $f_{\ell,s}$ the joint density of $(X_{\ell,j}, X_{\ell,j+s})$. Finally, we will denote by

$$f(x_1, x_2, x_3, x_4) = f_c(x_1, x_3|x_2, x_4)f_D(x_2, x_4),$$

where f_c is the conditional density (with respect to the Lebesgue measure) of (X_1, X_3) given (X_2, X_4) and where f_D is the density (with respect to the counting measure) of (X_2, X_4) .

This estimation method consists in integrating the regression function $m(\cdot)$ with respect to a suitable density function. By doing this we obtain

$$(4) \quad \int_{\mathbb{R}^{q_1+p_2+q_2}} m(x)q_2(x_2)q_{34}(x_3, x_4)\mu(dx_2)dx_3\mu(dx_4) \\ = \omega + m_1(x_1) + \int_{\mathbb{R}^{q_1}} m_2(x_2)q_2(x_2)\mu(dx_2) + \int_{\mathbb{R}^{p_2+q_2}} m_{34}(x_3, x_4)q_{34}(x_3, x_4)dx_3\mu(dx_4).$$

On the other hand, integrating $m(\cdot)$ with respect to a density function $q(x_1, x_2, x_3, x_4) = q_1(x_1)q_2(x_2) \times q_{34}(x_3, x_4)$ defined on $\mathbb{R}^{p_1+q_1+p_2+q_2}$ we obtain

$$(5) \quad \int_{\mathbb{R}^{p_1+q_1+p_2+q_2}} m(x)q_1(x_1)q_2(x_2)q_{34}(x_3, x_4)dx_1\mu(dx_2)dx_3\mu(dx_4) \\ = \omega + \int_{\mathbb{R}^{p_1}} m_1(x_1)q_1(x_1)dx_1 + \int_{\mathbb{R}^{q_1}} m_2(x_2)q_2(x_2)\mu(dx_2) + \int_{\mathbb{R}^{p_2+q_2}} m_{34}(x_3, x_4)q_{34}(x_3, x_4)dx_3\mu(dx_4).$$

Then subtracting equation (5) from (4) we obtain an expression for the additive component $m_1(x_1)$, up to an additive constant,

$$\eta_1(x_1) = m_1(x_1) - \int_{\mathbb{R}^{p_1}} m_1(x_1)q_1(x_1)dx_1 \\ = \int_{\mathbb{R}^{q_1+p_2+q_2}} m(x)q_2(x_2)q_{34}(x_3, x_4)dx_3\mu(dx_4) \\ - \int_{\mathbb{R}^{p_1+q_1+p_2+q_2}} m(x)q_1(x_1)q_2(x_2)q_{34}(x_3, x_4)dx_1\mu(dx_2)dx_3\mu(dx_4).$$

An estimator for $\eta_1(x_1)$, $\hat{\eta}_1(x_1)$, is obtained by replacing in the equation above the unknown quantities by some estimator

$$(6) \quad \hat{\eta}_1(x_1) = \int_{\mathbb{R}^{q_1+p_2+q_2}} \hat{m}_n(x) q_2(x_2) q_{34}(x_3, x_4) dx_3 \mu(dx_4) - \int_{\mathbb{R}^{p_1+q_1+p_2+q_2}} \hat{m}_n(x) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4).$$

An estimator for $m(x)$ is

$$\hat{m}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{Y_i \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_{1i}}{h_1}\right) \mathbb{I}(x_2 = X_{2i}) \frac{1}{h_3^{p_2}} L\left(\frac{x_3 - X_{3i}}{h_3}\right) \mathbb{I}(x_4 = X_{4i})}{f(X_{1i}, X_{2i}, X_{3i}, X_{4i})}.$$

This is the so-called "internal" estimator of Jones, Davies and Park (1994). $\mathbb{I}(A)$ stands for the indicator function that takes value one if A is true, and zero otherwise. In smoothing problems, the indicator function has been proposed in another contexts by Delgado and Mora (1995) and Fan, Härdle and Mammen (1998) to account for discrete variables. Further Racine and Li (2000) propose a kernel function that depends on a smoothing parameter. Delgado and Mora (1995) did not consider the case of a mixture of continuous and discrete variables, Fan, Härdle and Mammen (1998) take the indicator function over a broader set of values of X^d on its support, and finally Racine and Li (2000) face the additional problem of estimating a control parameter with no theoretical gains in doing so.

Following the marginal integration method, the component

$$\eta_2(x_2) = m_2(x_2) - \int_{\mathbb{R}^{q_1}} m_2(x_2) q_2(x_2) \mu(dx_2),$$

is estimated by

$$\hat{\eta}_2(x_2) = \int_{\mathbb{R}^{p_1+p_2+q_2}} \hat{m}_n(x) q_1(x_1) q_{34}(x_3, x_4) dx_1 dx_3 \mu(dx_4) - \int_{\mathbb{R}^{p_1+p_2+q_1+q_2}} \hat{m}_n(x) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4),$$

and the component

$$\eta_{34}(x_{34}) = m_{34}(x_{34}) - \int_{\mathbb{R}^{p_2+q_2}} m_{34}(x_3, x_4) q_{34}(x_3, x_4) dx_3 \mu(dx_4)$$

is estimated by

$$\hat{\eta}_{34}(x_3, x_4) = \int_{\mathbb{R}^{p_1+q_1}} \hat{m}_n(x) q_1(x_1) q_2(x_2) dx_1 \mu(dx_2) - \int_{\mathbb{R}^{p_1+p_2+q_1+q_2}} \hat{m}_n(x) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4).$$

In what follows we give some results about the asymptotic behavior of the estimators $\hat{\eta}_1$, $\hat{\eta}_2$ and $\hat{\eta}_{34}$. We give first some definitions and assumptions

(H.1) $m_1(x_1)$ is k -times continuously differentiable with respect to all its arguments in the support \mathcal{X}_1 of X_1 . Furthermore, $m_{34}(x_3, x_4)$ is k -times continuously differentiable with respect to $X_3 \in \mathcal{X}_3$ where \mathcal{X}_3 is the support of X_3 .

(H.2a) $\alpha(s) = O(s^{-\alpha})$, with $\alpha > \frac{2\beta}{\beta-2}$. For $\ell = 1, \dots, 4$ and $s \geq 1$ we have

$$\forall x, y, |f_{\ell,s}(x, y) - f_\ell(x) f_s(y)| \leq M < \infty.$$

(H.2b) $\alpha(s) = O(s^{-\alpha})$, with $a > \frac{2\beta}{\beta-2} \left(\frac{2k}{p_i} + 2 \right)$ for $i = 1, 3$. For $\ell = 1, \dots, 4$ and $s \geq 1$ we have

$$\forall x, y, |f_{\ell,s}(x, y) - f_{\ell}(x)f_s(y)| \leq M < \infty.$$

(H.3) $q_1(x_1)$ is bounded and $k+1$ -times continuously differentiable in the support \mathcal{X}_1 of X_1 . $q_2(x_2)$ is bounded with respect to all its arguments in the support \mathcal{X}_2 of X_2 . Furthermore, $q_{34}(x_3, x_4)$ is bounded and $k+1$ -times continuously differentiable with respect to X_3 in \mathcal{X}_3 .

(H.4) The kernel functions $K(\cdot)$ and $L(\cdot)$ are compactly supported, bounded, continuous and they integrate to one. Furthermore,

$$\begin{aligned} \forall (i_1, \dots, i_{p_1}) \in \mathbb{R}^{*p_1}, \quad (\forall j, i_j < k) &\Rightarrow \int_{\mathbb{R}^{p_1}} u_1^{i_1} \cdots u_{p_1}^{i_{p_1}} K(u_1, \dots, u_{p_1}) du_1 \cdots du_{p_1} = 0 \\ \forall (i_1, \dots, i_{p_1}) \in \mathbb{R}^{*p_1}, \quad \forall j, \int_{\mathbb{R}^{p_1}} u_j^k K(u_1, \dots, u_{p_1}) du_1 \cdots du_{p_1} &\in \mathbb{R}^* \end{aligned}$$

and

$$\begin{aligned} \forall (i_1, \dots, i_{p_2}) \in \mathbb{R}^{*p_2}, \quad (\forall j, i_j < k) &\Rightarrow \int_{\mathbb{R}^{p_2}} u_1^{i_1} \cdots u_{p_2}^{i_{p_2}} L(u_1, \dots, u_{p_2}) du_1 \cdots du_{p_2} = 0 \\ \forall (i_1, \dots, i_{p_2}) \in \mathbb{R}^{*p_2}, \quad \forall j, \int_{\mathbb{R}^{p_2}} u_j^k L(u_1, \dots, u_{p_2}) du_1 \cdots du_{p_2} &\in \mathbb{R}^* \end{aligned}$$

(H.5) The bandwidth satisfy $h_1 = c_1 n^{-\frac{1}{2k+p_1}}$ and $h_3 = c_1 n^{-\frac{1}{2k+p_3}}$.

(H.6) The functions $f(\cdot)$ and f_{ℓ} , for $\ell = 1, 2, 3, 4$ are such that

$$\exists b, B \text{ such that } 0 < b \leq f(x) \leq B < \infty \text{ and } 0 < b \leq f_{\ell}(x_{\ell}) \leq B < \infty.$$

Let $x_{-\ell} = (x_1, \dots, x_{\ell-1}, x_{\ell+1}, \dots, x_4)$. Then the conditional density $f(x_{-\ell}|x_{\ell})$ exists and it is bounded away from zero on the support of $f(\cdot)$.

(H.7) The conditional variance $\sigma_0^2(x) = \text{Var}(Y|X = x)$ is continuous.

(H.8) $\forall i, j, E \left[|Y_i Y_j|^{\beta/2} \middle| X \right] \leq M < \infty, \beta > 2$.

Assumptions (H.1), (H.4), (H.5) and (H.6) are standard in nonparametric regression techniques. In fact (H.4) assumes higher order kernels (see Vieu, 1991). Note that as expected the number of derivatives allowed in (H.1) matches the order of the kernels in (H.4). This is needed to control the bias in the multivariate estimator. The bandwidth rates in (H.5) are chosen according the previous conditions on kernels and densities. (H.6) introduces a strong assumption: The densities must be compactly supported. This is done without loss of generality. In fact we could allow for unbounded support using trimming techniques (Robinson, 1988), but this would complicate the analysis unnecessarily. (H.2a) and (H.2b) are mixing conditions. Note that we have considered separately the discrete and the continuous covariates case. In this condition it is assumed that mixing coefficients decay at a algebraic rate. This is the weakest condition it can be imposed for the rate of decay of the mixing coefficients (see Bosq, 1998).

Now with the previous assumptions in hand we provide two results that characterize the asymptotic properties of the different components. The proofs are relegated to the Appendix. We start by the estimators of the component that depend respectively on continuous explanatory variables, $\hat{\eta}_1(x_1)$, and a mixture of continuous and discrete regressors, $\hat{\eta}_{34}(x_3, x_4)$.

Theorem 1 *i)* Consider assumptions (H.1), (H.2b), (H.3), (H.4), (H.5), (H.6), (H.7) and (H.8) hold, then as $n \rightarrow \infty$, we have

$$(7) \quad \sqrt{nh_1^{p_1}} (\hat{\eta}_1(x_1) - \eta_1(x_1)) \rightarrow_d \mathcal{N}(b(x_1), v^2(x_1))$$

$$b(x_1) = \frac{1}{k!} \sum_{j=1}^{p_1} \int u_j^k K(u) du \left[(-1)^k \frac{\partial^k m_1}{\partial x_{1j}^k}(x_1) + \int m_1(z_1) \frac{\partial^k q_1}{\partial z_{1j}^k}(z_1) dz_1 \right],$$

$$v^2(x_1) = \int K^2(u) du \int \int \int [\sigma_0^2(x_1, x_2, x_3, x_4) + m^2(x_1, x_2, x_3, x_4)] \\ \times \frac{[q_2(x_2)q_{34}(x_3, x_4)]^2}{f(x_1, x_2, x_3, x_4)} \mu(dx_2) dx_3 \mu(dx_4).$$

ii) Furthermore, as n grows up to infinity, we have

$$(8) \quad \sqrt{nh_3^{p_2}} (\hat{\eta}_{34}(x_3, x_4) - \eta_{34}(x_3, x_4)) \rightarrow_d \mathcal{N}(b(x_3, x_4), v^2(x_3, x_4))$$

with

$$b(x_3, x_4) = \frac{1}{k!} \sum_{j=1}^{p_2} \int u_j^k L(u) du \left[(-1)^k \frac{\partial^k m_{34}}{\partial x_{3j}^k}(x_3, x_4) + \int m_{34}(z_3, z_4) \frac{\partial^k q_{34}}{\partial z_{3j}^k}(z_3, z_4) dz_3 \mu(dz_4) \right],$$

and

$$v^2(x_3, x_4) = f_4(x_4) \int L^2(u) du \int \int [\sigma_0^2(x_1, x_2, x_3, x_4) + m^2(x_1, x_2, x_3, x_4)] \\ \times \frac{[q_1(x_1)q_2(x_2)]^2}{f(x_1, x_2, x_3, x_4)} dx_1 \mu(dx_2).$$

our result in (7) is a generalization of the one obtained in Theorem 1 from Fan, Härdle and Mammen (1998) to dependent observations. Furthermore, the result in (8) remarks that in the case of mixture between continuous and discrete variables, the asymptotic variance of the marginal integration estimator suffers only from the dimensionality of the continuous variables. That is, the dimension of the discrete variables does not affect the rate of convergence of the estimator. Finally, we provide also an interesting result for the marginal integration estimator with all discrete covariables, $\hat{\eta}_2(x_2)$. The statistical properties of this estimator are given in the next result

Theorem 2 Consider assumptions (H.1), (H.2a), (H.3), (H.4), (H.5), (H.6), (H.7) and (H.8) hold, then

$$\sqrt{nh} (\hat{\eta}_2(x_2) - \eta_2(x_2)) \rightarrow_d \mathcal{N}(0, v^2(x_2))$$

$$v^2(x_2) = (f_2(x_2) - q_2(x_2))^2 \int \int \int [\sigma_0^2(z_1, z_2, z_3, z_4) + m^2(z_1, z_2, z_3, z_4)] \\ \times \frac{[q_1(z_1)q_{34}(z_3, z_4)]^2}{f(z_1, z_2, z_3, z_4)} dz_1 dz_3 \mu(dz_4) - \eta_2^2(x_2),$$

as n tends to infinity.

Note that although the multivariate nonparametric estimator contains some smoothing, the bias of $\hat{\eta}_2(x_2)$ is exactly equal to zero. This is because the marginal integration estimator of $\eta_2(x_2)$ is obtained by integrating out all directions that contain some smoothness.

3 A Semiparametric Estimator of a Partially Linear Model

As already indicated in the Introduction, the presence of discrete explanatory in nonparametric regression problems can be approached by rewriting the model as a semiparametric one. This semiparametric model combines a linear parametric part (with discrete covariates) plus a nonparametric term that contains the continuous variables. The partially linear model has long tradition in the econometrics literature and it was fully analyzed in an i.i.d. context in Robinson (1988) among others. Furthermore, if an additional restriction of additivity in the nonparametric part is added, then we obtain the so called additive partially linear model. Examples of this model have been considered in Opsomer (1999) and Li (2000). Although, as explained in Section 2, many econometric problems of interest do not admit the partially additive linear decomposition in this Section we adopt it and we obtain a root-n consistent semiparametric estimator of the parametric part. This estimator can be compared with other previous in the literature.

If in the econometric model introduced in Section 2 we impose the additional restrictions $m_2(x_2) = \sum_{l=1}^{q_1} m_{2l}(x_{2l})$ and, without loss of generality, $m_{2l}(x_{2l}) = \theta_l + \gamma_l x_{2l}$, then (3) has the following expression

$$(9) \quad Y_i = \omega + \sum_{l=1}^{q_1} \theta_l + m_1(X_{1i}) + \sum_{l=1}^{q_1} \gamma_l X_{2li} + m_{34}(X_{3i}, X_{4i}) + \epsilon_i.$$

Note that in this context, the identification restriction $E[m_2(x_2)] = 0$ implies that $\theta_l = -\gamma_l E(X_{2l})$, for $l = 1, \dots, q_1$. If we rewrite (9) under the previous restriction we obtain

$$(10) \quad Y_i = \omega + m_1(X_{1i}) + \sum_{l=1}^{q_1} \gamma_l (X_{2li} - E(X_{2l})) + m_{34}(X_{3i}, X_{4i}) + \epsilon_i.$$

In this model, it is of interest to estimate the components $\gamma_1, \gamma_2, \dots, \gamma_{q_1}$ at root-n rate. Furthermore, in order to make inference it is interesting to obtain its asymptotic distribution. One problem is that the previous identification restriction introduces in the estimating equation quantities that are unknown for the researcher as the expected values $E(X_{2l}), \dots, E(X_{2q_1})$. One way to solve this problem is to introduce the following assumption

(H.9) $q_2(x_2) = 1$ in the support of X_2 .

Note that other identification strategies are possible. For example, in Fan, Härdle and Mammen (1998), p. 952, for the sake of identification they make $\theta = \sum \omega + \sum_{l=1}^{q_1} \theta_l$ and they overestimate the quantities $m_1(\cdot)$ and $m_{34}(\cdot, \cdot)$ by an amount of θ .

Let $\{x_{2lj}\}_{j=1}^J$ be the set of all possible values that X_{2l} can take such that $f_{2l}(x_{2lj}) = P(X_{2l} = x_{2lj}) > 0$ for $j = 1, \dots, J$. Then, the easiest way to define an estimator seems to us to choose the value of γ_l that minimizes the L_2 distance between the model estimated nonparametrically, $\hat{\eta}_{2l}(x_{2l})$, and its corresponding linear direction, $\gamma_l(x_{2l} - \bar{X}_{2l})$, i. e.

$$\hat{\gamma}_l = \operatorname{argmin} \sum_{j=1}^J (\gamma_l(x_{2lj} - \bar{X}_{2l}) - \hat{\eta}_{2l}(x_{2lj}))^2,$$

where $\bar{X}_{2l} = \frac{1}{J} \sum_{j=1}^J x_{2lj}$. This idea was already explored in another context by Cristobal, Faraldo and Gonzalez-Manteiga (1987). Compared to others our estimator presents some advantages. First, its asymptotic properties are obtained under much weaker conditions. Mainly, lagged endogenous

variables may appear as regressors. Second, the estimator is unique, and it does not depend on cells or predetermined sets of values that can take the discrete variable. The following result is shown in the Appendix

Theorem 3 Consider assumptions (H.1), (H.2b), (H.3), (H.4), (H.5), (H.6), (H.7), (H.8) and (H.9) hold, then

$$\sqrt{n}(\hat{\gamma}_l - \gamma_l) \rightarrow_d \mathcal{N}(0, v_l^2)$$

with

$$v_l^2 = \frac{1}{J^2} \sum_{j=1}^J \left\{ f_{2l}(x_{2lj}) (x_{2lj} - \bar{X}_{2l})^2 \int \int \int [\sigma_0^2(z_1, x_{2lj}, z_3, z_4) + m^2(z_1, x_{2lj}, z_3, z_4)] \right. \\ \left. \times \frac{[q_1(z_1)q_{34}(z_3, z_4)]^2}{f(z_1, x_{2lj}, z_3, z_4)} dz_1 dz_3 \mu(dz_4) \right\} - \eta_2^2(x_2),$$

4 An application: Estimation of a wage equation in a labor supply model

Our aim here is to estimate a wage equation of the Spanish labor market. In order to do this, note that in any microeconomic study of the labor market, two facts are readily apparent: many individuals do not work, and wages are not available to nonworking people. This introduces a serious bias in the estimation of many behavioral equations since only a nonrandomly chosen subsample is available to estimate the parameters of interest. This is pointed out in Gronau (1974) and Heckman (1974). In their papers a sample selection model is introduced, consisting of two equations: a wage equation, explaining the potential log-wage rate of every individual, including non-workers, and a selection equation indicating whether or not someone is employed and therefore the wage is observed. Taking into account the above restrictions, we propose the following sample selection model, also referred in Amemiya (1985) as the Type II Tobit model,

$$(11) \quad S_i = \mathbb{I} \left(\sum_{l=1}^{k_1} \beta_l Z_{1il} + U_{1i} > 0 \right) \quad i = 1, \dots, n$$

$$(12) \quad W_i = \mathbb{I}(S_i = 1) \times \left(\omega + \sum_{l=1}^{k_2} m_l(Z_{2il}) + U_{2i} \right) \quad i = 1, \dots, n$$

Here, the parameters $\beta_1, \dots, \beta_{k_1}$, and the functions $m_1(Z_{21}), \dots, m_{k_2}(Z_{k_2})$ are unknown and need to be estimated. (U_1, U_2) are random variables whose realizations are unobserved by the researcher. The observed variables are W, S, Z_1 and Z_2 . Z_1 and Z_2 might contain common variables. S denotes a dummy variable indicating whether the individual has a paid job or not, and W is the wage someone receives if he/she is employed. It is only observed iff $S = 1$. Equation (12) is the so called market wage equation. The explanatory variables in this equation, Z_2 , are the standard ones in this type of models (see Vella, 1998), i.e. four dummy variables associated with age, and three dummy variables referring to education level. We also used the unemployment rate in the area of residence since participation may depend on cyclical conditions of the economy.

Equation (11) reflects the difference between the market and the reservation wage. It is a reduced form participation equation. Therefore, among the explanatory variables in this equation, Z_1 , we can find variables related to both market and individual characteristics: One dummy variable for the gender differential effect, three dummy variables referring to education level. Education level is used as an indicator of potential earnings of individuals. We decided also to include a dummy variable that indicates marital status. This last variable approximates the reservation wage.

In order to estimate a wage equation for the Spanish labor market we have available data obtained from the *Encuesta de Población Activa (EPA)*, the Spanish quarterly Labor Force Survey. This survey has taken place every quarter since 1975 and is collected by the National Bureau of Statistics (INE). It covers approximately 60,000 households and contains information about 150,000 individuals aged over 16. It provides information at different levels of disaggregation at both national and regional level. From these surveys, in the second quarter of 1990 the National Bureau of Statistics randomly selected a cross-section of 4,989 individuals (1,010 are unemployed looking for work) and provided additional information about some variables that were considered relevant for labor market participation analysis.

The variables included in this data set are defined in Table 1, where we also include some descriptive statistics.

Variable	Description	Whole Sample	Worker Sample
AGE16-19	dummy, 1 if age 16 to 19	0.1317 (0.3383)	0.1111 (0.3145)
AGE20-25	dummy, age 20 to 25	0.2653 (0.4417)	0.2565 (0.4371)
AGE26-35	dummy, age 26 to 35	0.2782 (0.4483)	0.2614 (0.4398)
AGE>45	dummy, older than 45	0.1386 (0.3457)	0.1437 (0.3511)
ELEMENTARY	dummy elementary school	0.3550 (0.4773)	0.3399 (0.4740)
H.SCHOOL	dummy, high school	0.1158 (0.3202)	0.1062 (0.3083)
UNIVERSITY	dummy, university	0.0643 (0.2455)	0.0392 (0.1943)
U-RATE	unemployment rate	0.1718 (0.0693)	0.1714 (0.0710)
NOT HEAD OF HOUSE	dummy, 1 if person is not head of household	0.7039 (0.4567)	0.6160 (0.4867)
SEXF	dummy, 1 if female	0.6802 (0.4666)	0.6258 (0.4843)
SINGLE	dummy, 1 if single	0.6891 (0.4631)	0.7255 (0.4466)
PARTICIPATING	dummy, 1 if participating	0.6059 (0.4888)	... (...)
SIZE		1010	612

Table 1: *Comparative Statistics of explanatory variables, mean and standard deviation (in brackets).*

Here, we estimate the wage equation using the two step method proposed in Heckman (1979). Since we are only interested in estimating the wage equation we focus our attention in the specification and estimation of such equation, and therefore, we skip all details about the estimation of the

participation equation. More details about the estimation procedure and correspondent results can be found in Fernandez and Rodriguez-Poo (2001). Taking into account all previous considerations and following Heckman's two step procedure the wage equation has the following expression

$$(13) \quad W_i = \omega + m_1(Z_{2i1}) + m_2(Z_{2i2}) + m_3(Z_{2i3}) + \theta \lambda \left(\sum_{l=1}^{k_1} \hat{\beta}_l Z_{1il} \right) + \xi_i, \quad \text{for } S_i = 1,$$

where

$$\xi_i = W_i - E[W|S = 1, Z_{1i}, Z_{2i}] - \theta \left\{ \lambda \left(\sum_{l=1}^{k_1} \hat{\beta}_l Z_{1il} \right) - \lambda \left(\sum_{l=1}^{k_1} \beta_l Z_{1il} \right) \right\},$$

θ is a nuisance parameter and $\lambda = \phi(\cdot)/\Phi(\cdot)$ is the inverse of the Mill's ratio. $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the density and the distribution function. $\hat{\beta}_1, \dots, \hat{\beta}_{k_1}$ are probit maximum likelihood estimators of the parameters of the selection equation (Amemiya, 1985). Since Z_{21} and Z_{22} are discrete variables that stand for age (four values) and education (four values) and Z_{23} is a continuous variable (unemployment rate) the marginal integration estimators of Section 2 are used, and the results are presented in Figure 1

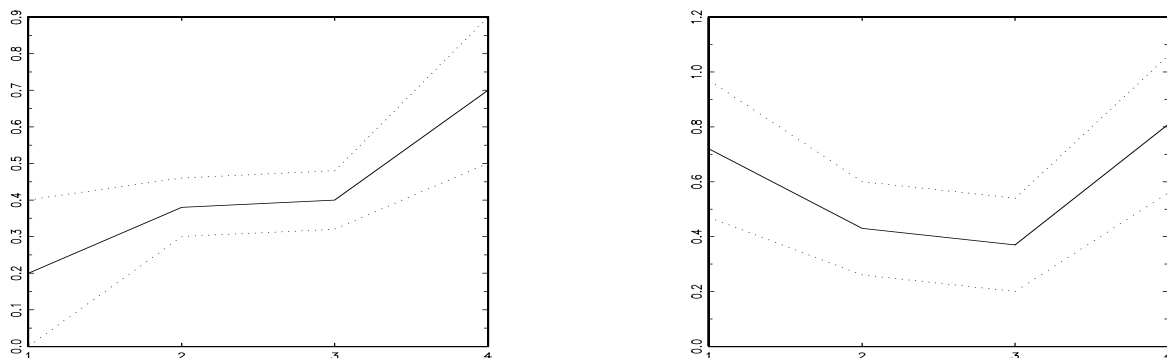


Figure 1: In the x-axis we represent respectively the education level (l.h.s.) and the age (r.h.s.). In the y-axis are shown the estimated nonparametric functions (solid line), and their confidence bands with 5% of significance level (dotted lines)

The estimated values have been computed using gaussian kernels and the bandwidth has been chosen by over-smoothing. The reason is that as we have learned in Theorem 3, the bias of the estimators is exactly equal to zero, and therefore, we can choose the bandwidth in such a way that minimizes the variance. Now, following the method suggested in Section 3, we use the estimates that have been obtained above to compute γ_1 and γ_2 in

$$(14) \quad W_i = \omega + \gamma_1 Z_{2i1} + \gamma_2 Z_{2i2} + m_3(Z_{2i3}) + \theta \lambda \left(\sum_{l=1}^{k_1} \hat{\beta}_l Z_{1il} \right) + \xi_i, \quad \text{for } S_i = 1.$$

Parameter estimates and their estimated standard errors are shown in Table 2 In order to obtain the confidence bands in Figure 1 and the estimated standard deviations in Table 2, the unknown quantities given in Theorems 2 and 3 are replaced by consistent estimators.

The results obtained are standard in this type of problems. Then, wages are directly related to education level, and there exists the U-shape relationship between wages and age.

Parameter	Estimated value	Estimated St. dev.
γ_1	0.472	0.039
γ_2	0.624	0.021

Table 2: *Parameter estimates and estimated standard deviations calculated according to the method proposed in Section 3*

Appendix

Proof of Theorem 1.i.

We first state some notations. Let

$$\begin{aligned}\alpha_1(x_1) &= \int \int \int m(x_1, x_2, x_3, x_4) q_2(x_2) q_{34}(x_3, x_4) \mu(dx_2) dx_3 \mu(dx_4); \\ \hat{\alpha}_1(x_1) &= \int \int \int \hat{m}_n(x_1, x_2, x_3, x_4) q_2(x_2) q_{34}(x_3, x_4) \mu(dx_2) dx_3 \mu(dx_4); \\ C_n &= \mu + \int \int \int (m_2(z_2) + m_{34}(z_3, z_4)) g_n(z_2, z_3, z_4) \mu(dz_2) dz_3 \mu(dz_4); \\ \hat{C}_n &= \int \int \int \int \hat{m}_n(x_1, x_2, x_3, x_4) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4); \\ C &= \int m_1(x_1) q_1(x_1) dx_1; \\ g_n(z_2, z_3, z_4) &= \int \int \int \mathbb{I}(x_2 = z_2) \frac{1}{h_3^{p_3}} L\left(\frac{x_3 - z_3}{h_3}\right) \mathbb{I}(x_4 = z_4) q_2(x_2) q_{34}(x_3, x_4) \mu(dx_2) dx_3 \mu(dx_4).\end{aligned}$$

Remark: By (H.3) we have $g_n(z_2, z_3, z_4) = q_2(z_2) q_{34}(z_3, z_4) + o(1)$.

We can also write

$$\hat{\alpha}_1(x_1) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_{1i}}{h_1}\right) \frac{\tilde{Y}_{ni}}{f_1(X_{1i})}$$

with

$$\tilde{Y}_{ni} = \frac{Y_i f_1(X_{1i})}{f(X_{1i}, X_{2i}, X_{3i}, X_{4i})} g_n(X_{2i}, X_{3i}, X_{4i}).$$

Then, we have written $\hat{\alpha}_1$ as a nonparametric estimator of $\tilde{m}_n(\cdot) = E\left(\tilde{Y}_{ni} | X_{1i} = \cdot\right)$, and we have

$$\begin{aligned}\tilde{m}_n(x_1) &= m_1(x_1) + C_n, \\ \eta_1(x_1) &= m_1(x_1) - C, \\ \hat{\eta}_1(x_1) &= \hat{\alpha}_1(x_1) - \hat{C}_n.\end{aligned}$$

The proof of the asymptotic normality of $\hat{\eta}_1 - \eta_1$ is obtained by the proof of the three following points:

$$(15) \quad \sqrt{nh_1^{p_1}} (\hat{\alpha}_1(x_1) - \tilde{m}_n(x_1)) \rightarrow_d \mathcal{N}(b_1(x_1), v^2(x_1)),$$

$$(16) \quad E\left(\hat{C}_n - C_n - C\right) = h_1^k b_1 + o\left(h_1^k\right),$$

$$(17) \quad \text{Var}\left(\hat{C}_n\right) = o\left(\frac{1}{nh_1^{p_1}}\right),$$

where

$$b_1 = \frac{(-1)^k}{k!} \sum_{j=1}^{p_1} \int u_j^k K(u) du \int m_1(z_1) \frac{\partial^k q_1}{\partial z_{1j}^k}(z_1) dz_1,$$

and where $b_1(x_1) = b(x_1) - b_1$.

Proof of (15)

For the bias part, integrating by substitution and using a Taylor expansion of m_1 , we have

$$\begin{aligned} E\hat{\alpha}_1(x_1) - \tilde{m}_n(x_1) &= E \left(\frac{1}{h_1^{p_1}} K \left(\frac{x_1 - X_1}{h_1} \right) \frac{\tilde{Y}_{n1}}{f_1(X_1)} \right) - \tilde{m}_n(x_1) \\ (18) \quad &= h_1^k \frac{(-1)^k}{k!} \sum_{j=1}^{p_1} \int u_j^k K(u) du \frac{\partial^k m_1}{\partial x_{1j}^k}(x_1) + o(h_1^k). \end{aligned}$$

Now we have to compute the variance of $\hat{\alpha}_1(x_1)$.

$$(19) \quad \text{Var}(\hat{\alpha}_1(x_1)) = \frac{1}{nh_1^{2p_1}} \text{Var}(\Delta_i) + \frac{2}{(nh_1)^{2p_1}} \sum_{1 \leq i < j \leq n} \text{Cov}(\Delta_i, \Delta_j),$$

where

$$\Delta_i = K \left(\frac{x_1 - X_{1i}}{h_1} \right) \frac{\tilde{Y}_{ni}}{f_1(X_{1i})} - E \left\{ K \left(\frac{x_1 - X_{1i}}{h_1} \right) \frac{\tilde{Y}_{ni}}{f_1(X_{1i})} \right\}.$$

Integrating by substitution and by (H.3) and (H.4), we have $E\Delta_i = \mathcal{O}(h_1^{p_1})$ and then

$$(20) \quad \lim_{n \rightarrow \infty} nh_1^{p_1} \left[\frac{1}{nh_1^{2p_1}} E\Delta_i^2 \right] = 0.$$

Integrating by substitution and using (H.3), (H.4) and (H.5), we obtain that

$$(21) \quad E\Delta_i^2 = v^2(x_1) + o\left(\frac{1}{nh_1^{p_1}}\right),$$

with

$$\begin{aligned} v^2(x_1) &= \int K^2(u) du \int \int \int [\sigma_0^2(x_1, x_2, x_3, x_4) + m^2(x_1, x_2, x_3, x_4)] \\ &\quad \times \frac{[q_2(x_2)q_{34}(x_3, x_4)]^2}{f(x_1, x_2, x_3, x_4)} \mu(dx_2) dx_3 \mu(dx_4). \end{aligned}$$

Now we will look at the covariance terms. Integrating by substitution and using (H.3), (H.4) and (H.5), we have

$$(22) \quad \text{Cov}(\Delta_i, \Delta_j) = \mathcal{O}(h_1^{2p_1}).$$

On the other hand, by (H.8), $E|\tilde{Y}_{ni}|^\beta \leq M < \infty$, and then $E|\Delta_i|^\beta \leq M < \infty$, that allows us to use the covariance inequality for strongly mixing processes (see e.g. Bosq, 1998, Corollary 1.1, p. 21). Then we have

$$|\text{Cov}(\Delta_i, \Delta_j)| \leq M\alpha^{\frac{\beta-2}{\beta}} (|i-j|).$$

Now we proceed as in Bosq (1996, p. 43) and we introduce a sequence u_n of integers that allows to write

$$\begin{aligned} \sum_{1 \leq i < j \leq n} \text{Cov}(\Delta_i, \Delta_j) &= \sum_{|i-j| \leq u_n} \text{Cov}(\Delta_i, \Delta_j) + \sum_{|i-j| > u_n} \text{Cov}(\Delta_i, \Delta_j) \\ &= \mathcal{O}\left(h_1^{2p_1} n u_n + n^2 \alpha^{\frac{\beta-2}{\beta}}(u_n)\right). \end{aligned}$$

Choosing $u_n = (h_1^{p_1} \log n)^{-1}$ gives with (H.2b)

$$(23) \quad \lim_{n \rightarrow \infty} n h_1^{p_1} \left[\frac{1}{n h_1^{2p_1}} \sum_{1 \leq i < j \leq n} \text{Cov}(\Delta_i, \Delta_j) \right] = 0.$$

Because of (H.2b) we can now apply a CLT for mixing random variables (see e. g. Rio, 2000, Theorem 4.2., p. 64). So, the relations (18), (19), (20), (21), (22), and (23) lead directly to (15).

Proof of (16)

Computing $E\{\hat{m}_n(x_1, x_2, x_3, x_4)\}$ in a standard way, and since the regression function m is additive, we arrive at

$$E\left(\hat{C}_n - C_n\right) = \int m_1(z_1) \int \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - z_1}{h_1}\right) q_1(x_1) dx_1 dz_1.$$

A Taylor expansion of q_1 leads directly to (16).

Proof of (17)

We have to compute

$$\text{Var}\left(\hat{C}_n\right) = \frac{1}{n} \text{Var}(U_1) + \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \text{Cov}(U_i, U_j),$$

where

$$U_i = \frac{Y_i}{f(X_{1i}, X_{2i}, X_{3i}, X_{4i})} p_n(X_{1i}) g_n(X_{2i}, X_{3i}, X_{4i})$$

and

$$p_n(X_{1i}) = \int \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_{1i}}{h_1}\right) q_1(x_1) dx_1.$$

By (H.3), (H.4), (H.5) and integrating by substitution, we can see that $\text{Var}(U_1) = \mathcal{O}(1)$ and $E|U_i|^\beta \leq M < \infty$. Then, the covariance terms can be treated exactly as we did before for getting (23) by Rio's inequality and by condition (H.2b). This is enough to see that the relation (17) is proved.

Proof of Theorem 1.ii.

It remains now to prove the second part of Theorem 1, namely the equation (8). The proof follows the same lines as for the estimation of the additive component m_1 , because m_{34} depends on some

continuous random variable. So we will just give the main steps. Introduce the notations:

$$\begin{aligned}\alpha_{34}(x_3, x_4) &= \int \int m(x_1, x_2, x_3, x_4) q_1(x_1) q_2(x_2) dx_1 \mu(dx_2); \\ \hat{\alpha}_{34}(x_3, x_4) &= \int \int \hat{m}_n(x_1, x_2, x_3, x_4) q_1(x_1) q_2(x_2) dx_1 \mu(dx_2); \\ D_n &= \mu + \int \int (m_1(z_1) + m_2(z_2)) g_n(z_1, z_2) dz_1 \mu(dz_2); \\ \hat{D}_n &= \int \int \int \int \hat{m}_n(x_1, x_2, x_3, x_4) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4); \\ D &= \int \int m_{34}(x_3, x_4) q_{34}(x_3, x_4) dx_3 \mu(dx_4); \\ g_n(z_1, z_2) &= \int \int \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - z_1}{h_1}\right) \mathbb{I}(x_2 = z_2) q_1(x_1) q_2(x_2) dx_1 \mu(dx_2).\end{aligned}$$

Remark: By (H.3), we have $g_n(z_1, z_2) = q_1(z_1)q_2(z_2) + o(1)$.

We can also write

$$\hat{\alpha}_{34}(x_3, x_4) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_3^{p_2}} L\left(\frac{x_3 - X_{3i}}{h_3}\right) \mathbb{I}(x_4 = z_4) \frac{\tilde{Y}_{ni}}{f_4(X_{4i}) f_c(X_{3i}|X_{4i})}$$

with

$$\tilde{Y}_{ni} = \frac{Y_i f_3(X_{3i})}{f_c(X_{1i}, X_{3i}|X_{2i}, X_{4i}) f_2(X_{2i})} g_n(X_{1i}, X_{2i}).$$

Then, we have rewritten $\hat{\alpha}_{34}$ as a nonparametric estimator of $\tilde{m}_n(\cdot, \cdot) = E\left(\tilde{Y}_{ni} \mid (X_{3i}, X_{4i}) = (\cdot, \cdot)\right)$, and we have

$$\begin{aligned}\tilde{m}_n(x_3, x_4) &= m_{34}(x_3, x_4) + D_n, \\ \eta_{34}(x_3, x_4) &= m_{34}(x_3, x_4) - D, \\ \hat{\eta}_{34}(x_3, x_4) &= \hat{\alpha}_{34}(x_3, x_4) - \hat{D}_n.\end{aligned}$$

The proof of the asymptotic normality of $\hat{\eta}_{34} - \eta_{34}$ will be obtained from the three following points that can be proved exactly as results (15), (16) and (17):

$$(24) \quad \sqrt{nh_3^{p_2}} (\hat{\alpha}_{34}(x_3, x_4) - \tilde{m}_n(x_3, x_4)) \rightarrow_d \mathcal{N}(b_{34}(x_3, x_4), v^2(x_3, x_4)),$$

$$(25) \quad E(\hat{D}_n - D_n - D) = h_3^k b_3 + o(h_3^k),$$

$$(26) \quad \text{Var}(\hat{D}_n) = o\left(\frac{1}{nh_3^{p_2}}\right),$$

with

$$\begin{aligned}b_{34}(x_3, x_4) &= h_3^k \frac{(-1)^k}{k!} \sum_{j=1}^{p_2} \int u_j^k L(u) du \frac{\partial^k m_{34}}{\partial x_{3j}^k}(x_3, x_4) + o(h_3^k) \\ b_3 &= \frac{1}{k!} \sum_{j=1}^{p_2} \int u_j^k L(u) du \int m_{34}(z_3, z_4) \frac{\partial^k q_{34}}{\partial x_{3j}^k}(x_3, x_4) dz_3 \mu(dz_4).\end{aligned}$$

Proof of Theorem 2

To prove the asymptotic normality of $\eta_2 - \hat{\eta}_2$ we have to show the following relationships

$$(27) \quad \sqrt{n} (\hat{\eta}_2(x_2) - E\hat{\eta}_2(x_2)) \rightarrow_d \mathcal{N}(0, v^2(x_2))$$

$$(28) \quad E\hat{\eta}_2(x_2) = \eta_2(x_2).$$

Proof of (27)

We write

$$\hat{\eta}_2(x_2) = \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathbb{I}(x_2 = X_{2i}) - q_2(X_{2i})}{f_2(X_{2i})} \right) \tilde{Z}_{ni} \equiv \frac{1}{n} \sum_{i=1}^n \Delta_i$$

where

$$\begin{aligned} \tilde{Z}_{ni} &= Y_i \int \int \int \frac{1}{h_1^{p_1}} K \left(\frac{x_1 - X_{1i}}{h_1} \right) \frac{1}{h_3^{p_2}} L \left(\frac{x_3 - X_{3i}}{h_3} \right) \mathbb{I}(x_4 = X_{4i}) \\ &\quad \times \frac{q_1(x_1)q_{34}(x_3, x_4)}{f_c(X_{1i}, X_{3i}|X_{2i}, X_{4i})f_4(X_{4i})} dx_1 dx_3 \mu(dx_4). \end{aligned}$$

Let us first compute the variance term

$$(29) \quad \text{Var}(\hat{\alpha}_2(x_2)) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(\Delta_i) + \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \text{Cov}(\Delta_i, \Delta_j),$$

Using (28) we directly have

$$(30) \quad E\Delta_i = E\hat{\eta}_2(x_2) = \eta_2(x_2).$$

Integrating by substitution and using (H.3), we obtain

$$(31) \quad \begin{aligned} E\Delta_i^2 &= (f_2(x_2) - q_2(x_2))^2 \int \int \int [\sigma_0^2(z_1, x_2, z_3, z_4) + m^2(z_1, x_2, z_3, z_4)] \\ &\quad \times \frac{[q_1(z_1)q_{34}(z_3, z_4)]^2}{f(z_1, x_2, z_3, z_4)} dz_1 dz_3 \mu(dz_4) + o(1). \end{aligned}$$

Now, for the computation of the covariance terms, by using (H.8) we obtain that $E|\tilde{Z}_{ni}|^\beta \leq M < \infty$ and then $E|\Delta_i|^\beta \leq M < \infty$, that allows us to use the covariance inequality for strongly mixing processes (see e.g. Bosq, 1998, Corollary 1.1, p. 21). Then we have

$$|\text{Cov}(\Delta_i, \Delta_j)| \leq M\alpha^{\frac{\beta-2}{\beta}} (|i-j|).$$

By a simple computation, and using (H.2a), we obtain

$$(32) \quad \left| \frac{1}{n^2} \sum_{1 \leq i < j \leq n} \text{Cov}(\Delta_i, \Delta_j) \right| \leq Mn^{1-\alpha\frac{\beta-2}{\beta}}.$$

Finally, using a Central Limit Theorem for strongly mixing processes (Rio, 2000, Theorem 4.2., p. 64) with relations (29), (30), (31), (32) and with (H.2a) we get directly (27).

Proof of (28)

We first compute the expectation of $\hat{m}_n(x_1, x_2, x_3, x_4)$:

$$E\{\hat{m}_n(x_1, x_2, x_3, x_4)\} = m(z_1, x_2, z_3, x_4) \frac{1}{h_1^{p_1}} K \left(\frac{x_1 - z_1}{h_1} \right) \frac{1}{h_3^{p_2}} L \left(\frac{x_3 - z_3}{h_3} \right) dz_1 dz_3,$$

and then, since the regression function is additive we easily obtain that

$$E\{\hat{\eta}_2(x_2)\} = m_2(x_2) - \int m_2(x_2)q_2(x_2)\mu(dx_2) = \eta_2(x_2),$$

and (28) is proved.

Proof of Theorem 3

Note that just to simplify notations we have removed the index l from X_{2l} . That is, along the proof we will use X_{2i} instead of X_{2li} and f_2 instead of f_{2l} . This is done just for notational convenience and without loss of generality. Let us define

$$\bar{X}_2 = \frac{1}{J} \sum_{j=1}^J x_{2j},$$

and

$$\sigma_{\bar{X}_2}^2 = \frac{1}{J} \sum_{j=1}^J (x_{2j} - \bar{X}_2)^2.$$

The estimator $\hat{\gamma}_l$ of γ_l is defined as follows:

$$\hat{\gamma}_l = \frac{\sum_{j=1}^J \hat{\eta}_2(x_{2j}) (x_{2j} - \bar{X}_2)}{\sum_{j=1}^J (x_{2j} - \bar{X}_2)^2}$$

The bias term is not difficult to compute. Because of Theorem 2, we have

$$\forall x_{2j}, E \{ \hat{\eta}_2(x_{2j}) \} = \eta_2(x_{2j}),$$

while, by assumption (H.9), the choice made for q_2 allows to see that

$$\eta_2(x_{2j}) = \gamma_l (x_{2j} - \bar{X}_2).$$

Clearly, this implies that we have

$$E \hat{\gamma}_l = \gamma_l.$$

So the only remaining question is to compute the variance term, $\text{Var}(\hat{\gamma}_l)$. It can be written as

$$\sigma_{\bar{X}_2}^4 \text{Var}(\hat{\gamma}_l) = \frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J \text{Cov}(\hat{\eta}_2(x_{2j}) (x_{2j} - \bar{X}_2), \hat{\eta}_2(x_{2j'}) (x_{2j'} - \bar{X}_2)).$$

Let, as in the proof of (27), introduce the quantity

$$\Delta_i(x_{2j}) = \left(\frac{\mathbb{I}(x_{2j} = X_{2i}) - q_2(X_{2i})}{f_2(X_{2i})} \right) \tilde{Y}_{ni}.$$

Then,

$$\begin{aligned} & \text{Cov}(\hat{\eta}_2(x_{2j}) (x_{2j} - \bar{X}_2), \hat{\eta}_2(x_{2j'}) (x_{2j'} - \bar{X}_2)) \\ &= \text{Cov} \left(\frac{1}{n} \sum_{i=1}^n \Delta_i(x_{2j}) (x_{2j} - \bar{X}_2), \frac{1}{n} \sum_{k=1}^n \Delta_k(x_{2j'}) (x_{2j'} - \bar{X}_2) \right) \\ (33) \quad &= \frac{1}{n^2} \sum_{i=1}^n \text{Cov}(\Delta_i(x_{2j}) (x_{2j} - \bar{X}_2), \Delta_i(x_{2j'}) (x_{2j'} - \bar{X}_2)) \end{aligned}$$

$$(34) \quad + \frac{1}{n^2} \sum_{i \neq k} \text{Cov}(\Delta_i(x_{2j}) (x_{2j} - \bar{X}_2), \Delta_k(x_{2j'}) (x_{2j'} - \bar{X}_2)).$$

Let us now look at the computation of (33). Note first that we have

$$\begin{aligned} & \frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J \frac{1}{n} \text{Cov} (\Delta_i (x_{2j}) (x_{2j} - \bar{X}_2), \Delta_i (x_{2j'}) (x_{2j'} - \bar{X}_2)) \\ &= \frac{1}{n} \frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J [E \Delta_i (x_{2j}) (x_{2j} - \bar{X}_2) \Delta_i (x_{2j'}) (x_{2j'} - \bar{X}_2) \\ & \quad - E \Delta_i (x_{2j}) (x_{2j} - \bar{X}_2) E \Delta_i (x_{2j'}) (x_{2j'} - \bar{X}_2)] \end{aligned}$$

Using the calculations of the proof of (27), we easily obtain

$$\begin{aligned} & \frac{1}{n} \frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J E \Delta_i (x_{2j}) (x_{2j} - \bar{X}_2) E \Delta_i (x_{2j'}) (x_{2j'} - \bar{X}_2) \\ &= \frac{1}{n} \frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J m_2(x_{2j}) (x_{2j} - \bar{X}_2) m_2(x_{2j'}) (x_{2j'} - \bar{X}_2) \\ &= \frac{1}{n} \gamma_l^2 \left[\frac{1}{J} \sum_{j=1}^J m_2(x_{2j}) (x_{2j} - \bar{X}_2) \right]^2 \\ (35) \quad &= \frac{1}{n} \gamma_l^2 \sigma_{X_2}^4. \end{aligned}$$

On the other hand we have

$$\begin{aligned} & \frac{1}{n} \frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J E \Delta_i (x_{2j}) (x_{2j} - \bar{X}_2) \Delta_i (x_{2j'}) (x_{2j'} - \bar{X}_2) \\ &= \frac{1}{n} E \left[\frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J \left(\frac{\mathbb{I}(x_{2j} = X_{2i}) - q_2(X_{2i})}{f_2(X_{2i})} \right) \tilde{Y}_{ni} (x_{2j} - \bar{X}_2) \right. \\ & \quad \times \left. \left(\frac{\mathbb{I}(x_{2j'} = X_{2i}) - q_2(X_{2i})}{f_2(X_{2i})} \right) \tilde{Y}_{ni} (x_{2j'} - \bar{X}_2) \right] \\ &= \frac{1}{n} E \left[\frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J \frac{\tilde{Y}_{ni}^2}{f_2^2(X_{2i})} (x_{2j} - \bar{X}_2) (x_{2j'} - \bar{X}_2) \right. \\ & \quad \times \left. \{ \mathbb{I}(x_{2j} = X_{2i}) \mathbb{I}(x_{2j'} = X_{2i}) - \mathbb{I}(x_{2j} = X_{2i}) q_2(X_{2i}) - \mathbb{I}(x_{2j'} = X_{2i}) q_2(X_{2i}) + q_2^2(X_{2i}) \} \right] \\ &= \frac{1}{n} E \left[\frac{1}{J^2} \sum_{j=1}^J \sum_{j'=1}^J \frac{\tilde{Y}_{ni}^2}{f_2^2(X_{2i})} (x_{2j} - \bar{X}_2) (x_{2j'} - \bar{X}_2) \mathbb{I}(x_{2j} = X_{2i}) \mathbb{I}(x_{2j'} = X_{2i}) \right] \\ &= \frac{1}{n} \frac{1}{J^2} \sum_{j=1}^J E \left[\frac{\tilde{Y}_{ni}^2}{f_2^2(X_{2i})} (x_{2j} - \bar{X}_2)^2 \mathbb{I}(x_{2j} = X_{2i}) \right]. \end{aligned}$$

Integrating by substitution and using (H.3), (H.4) and (H.5) give

$$\begin{aligned} & \frac{1}{n} \frac{1}{J^2} \sum_{j=1}^J E \left[\frac{\tilde{Y}_{ni}^2}{f_2^2(X_{2i})} (x_{2j} - \bar{X}_2)^2 \mathbb{I}(x_{2j} = X_{2i}) \right] \\ &= \frac{1}{n} \frac{1}{J^2} \sum_{j=1}^J \left\{ f_2(x_{2j}) (x_{2j} - \bar{X}_2)^2 \int \int \int [\sigma_0^2(z_1, x_{2j}, z_3, z_4) + m^2(z_1, x_{2j}, z_3, z_4)] \right. \\ & \quad \times \left. \frac{[q_1(z_1) q_{34}(z_3, z_4)]^2}{f(z_1, x_{2j}, z_3, z_4)} dz_1 dz_3 \mu(dz_4) + o(1) \right\}. \end{aligned}$$

Finally,

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Var}(\hat{\gamma}_l) &= \frac{1}{J^2} \sum_{j=1}^J \left\{ f_2(x_{2j}) (x_{2j} - \bar{X}_2)^2 \int \int \int [\sigma_0^2(z_1, x_{2j}, z_3, z_4) + m^2(z_1, x_{2j}, z_3, z_4)] \right. \\ &\quad \left. \times \frac{[q_1(z_1)q_{34}(z_3, z_4)]^2}{f(z_1, x_{2j}, z_3, z_4)} dz_1 dz_3 \mu(dz_4) \right\} - \gamma_l^2. \end{aligned}$$

It remains just to look at the computation of (34). Proceeding as in (32), we have

$$\frac{1}{n^2} \sum_{i \neq k} \text{Cov}(\Delta_i(x_{2j})(x_{2j} - \bar{X}_2), \Delta_k(x_{2j'})(x_{2j'} - \bar{X}_2)) = o\left(\frac{1}{n}\right).$$

We can write $\hat{\gamma}_l$ as

$$\hat{\gamma}_l = \frac{1}{n} \sum_{i=1}^n \delta_i,$$

where

$$\delta_i = \frac{1}{J\sigma_{X_2}^2} \Delta_i(x_{2j})(x_{2j} - \bar{X}_2),$$

and then, applying the central limit theorem for strongly mixing processes (Rio, 2000, Theorem 4.2., p. 64), our result is proved.

References

- Ahmad, I. A. and Cerrito, P. B. (1994) Nonparametric estimation of joint discrete-continuous probability densities with applications. *Journal of Statistical Planning and Inference*, **41**, 349-364.
- Amemiya T. (1985) *Advanced Econometrics*. Harvard University Press, Cambridge.
- Andrews, D. W. K. and Whang, Y.-J. (1990) Additive interactive regression models: circumvention of the curse of dimensionality. *Econometric Theory*, **6**, 466-479.
- Bierens, H. J. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78**, 383, 699-707.
- Bosq, D. (1998) *Nonparametric statistics for stochastic processes: Estimation and prediction*. Lecture Notes in Statistics, 110. Springer-Verlag, New York.
- Camlong-Viot, Ch., Sarda, P. and Vieu, Ph. (2000) Additive time series: the kernel integration method. *Mathematical Methods of Statistics*, **9**, 358-375.
- Cristóbal, J. A., Faraldo, P. and Gonzalez-Manteiga, W. (1987) A class of linear regression parameter estimators constructed by nonparametric estimation. *Annals of Statistics*, **15**, 603-609.
- Delgado, M. A. and Mora, J. (1995) Nonparametric and semiparametric estimation with discrete regressors. *Econometrica*, **63**, 1477-1484.
- Fan, J., Härdle, W. and Mammen, E. (1998) Direct estimation of low-dimensional components in additive models. *The Annals of Statistics*, **26**, 943-971.
- Fernandez, A. and Rodriguez-Poo, J. M. (2001) An empirical investigation of parametric and semi-parametric estimation methods in sample selection models. Universidad de Cantabria. Preprint.

- Gronau R. (1974) Wage comparisons—a selectivity bias. *Journal of Political Economy*, **82**, 1119-1143.
- Grund, B. and Hall, P. (1993) On the performance of kernel estimators for high-dimensional, sparse binary data, *Journal of Multivariate Analysis*, **44**, 321-344.
- Hall, P. (1981) On nonparametric multivariate binary discrimination. *Biometrika*, **68**, 287-294.
- Hastie, T. J. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman and Hall, London.
- Heckman J (1974) Shadow prices, market wages and labor supply. *Econometrica*, **42**, 679-694.
- Heckman J (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153-161.
- Horowitz, J. (1998) *Semiparametric methods in Econometrics*. Lecture Notes in Statistics, Springer Verlag, New York.
- Jones, M. C., Davies, S. J. and Par, B. U. (1994) Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.*, **89**, 825-832.
- Li, Q. (2000) Efficient estimation of partially linear models. *International Economic Review*, **41**, 1073-1091.
- Linton, O. and Nielsen, J. P. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93-100.
- Mammen, E. Linton, O. and Nielsen, J. P. (1995) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, **27**, 1443-1490.
- Newey, W. K. (1994) Kernel estimation of partial means. *Econometric Theory*, **10**, 233-253.
- Newey, W. K. (1995) Convergence for series estimators. In G. S. Maddala, P. C. B. Phillips and T. N. Srinivasan (eds.), *Statistical methods of Economics and Quantitative Economics: Essays in Honor of C. R. Rao*, 254-275.
- Opsomer, J. D. and Ruppert, D. (1994) Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, **25**, 186-211.
- Opsomer, J. D. and Ruppert, D. (1998) A fully automated bandwidth selection method for fitting additive models. *J. Amer. Statist. Assoc.*, **93**, 605-619.
- Opsomer, J. D. (2000) Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**, 166-179.
- Racine, J. and Li, Q. (2000) Nonparametric estimation of regression functions with both categorical and continuous data. Preprint.
- Rio, E. (2000) *Théorie asymptotique des processus aléatoires faiblement dépendants*. Mathématiques et Applications. Springer Verlag, New York.
- Robinson, P. M. (1988) Root-n consistent semiparametric regression. *Econometrica*, **56**, 931-954.
- Rosenblatt, M. (1956) A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U.S.A.*, **42**, 43-47.
- Sperlich, S., Tjostheim, D. and Yang, L. (2000) Nonparametric estimation and testing of interactions in additive models. *Econometric Theory*, **18**, 197-251.

Stone, C. J. (1985) Additive regression and other nonparametric models. *Annals of Statistics*, **14**, 592-606.

Stone, C. J., Hansen, M. H., Kooperberg, C. and Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics*, **25**, 1371-1470.

Tjøstheim, D. and Auestad, B. (1994) Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.*, **89**, 1398-1409.

Vella F. (1998) Estimating models with sample selection bias: a survey. *Journal of Human Resources*, **33**, 127-169.

Vieu, Ph. (1991) Quadratic errors for nonparametric estimators under dependence. *Journal of Multivariate Analysis*, **39**, 324-347.

Wahba, G. (1990) *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics, **59**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, .