Partial and Recombined Estimators for Nonlinear Additive Models

Nathalie Chèze[(1,2)] (nathalie.cheze@u-paris10.fr)
Jean-Michel Poggi[(2,3)] (jean-michel.poggi@math.u-psud.fr)
Bruno Portier[(2,3)] (bruno.portier@math.u-psud.fr)

(1) *Université Paris X, Modal'X, 200 avenue de la République, 92000 Nanterre*
(2) *Laboratoire de Mathématique – U.M.R. C 8628, "Probabilités, Statistique et Modélisation" Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France*
(3) *IUT Paris, Université René Descartes, 143 av. de Versailles, 75016 Paris*

Abstract. Starting from a variant of an estimator using marginal integration, this paper proposes partial and recombined estimators for nonlinear additive regression models. Partial estimators are used for data analysis purposes and recombined estimators are used to improve the estimation and prediction performances for small to moderate sample sizes. In the first part of the paper, some simulations illustrate step-by-step the principle and the value of the proposed estimators, which are finally applied to the analysis and prediction of ozone concentration in Paris area. In the second part of the paper, almost sure convergence results as well as a multivariate central limit theorem and a test for partial additivity are provided.

Keywords: Additive models; Kernel estimation; Marginal integration; Multivariate central limit theorem; Testing for partial additivity

AMS subject classification: 62G07, 62G20, 62G10, 62P99

1. Introduction

Let us consider nonlinear additive regression models of the form

$$Z_n \;=\; \sum_{i=1}^{d} f^i(X_n^i) \;+\; \mu \;+\; \varepsilon_n \tag{1.1}$$

where Z is a real-valued dependent variable, $X^1, \ldots, X^d$ are the explanatory variables, μ is a constant and ε is an unobservable noise.

Such nonlinear additive regression models have received considerable attention and have been widely used following the work of (Breiman and Friedman, 1985; Buja et al., 1989; Hastie and Tibshirani, 1990). Such models are particularly attractive. On one hand, with respect to the classical linear model $Z_n = \sum_{i=1}^{d} \theta^i X_n^i + \mu + \varepsilon_n$, considerable additional flexibility is given by the allowed nonlinear effect of each explanatory variable without losing ease of interpretation. On the other hand, with respect to the fully nonparametric model $Z_n = \Psi(X_n^1, \ldots, X_n^d) + \varepsilon_n$, the separable model (1.1) is more explicit and can be estimated without suffering from the so-called curse of dimensionality (Stone, 1985; Stone, 1986) which is the main drawback of the unstructured nonparametric regression model.

Of course in model (1.1), functions f^i are identifiable only up to an additive constant. Therefore to estimate each of them, the usual identifiability constraints $\mathbb{E}[f^i(X^i)] = 0$ for $i = 1, \ldots, d$, are assumed. The constant μ is consistently estimated by the mean of the (Z_n) with the parametric convergence rate. So, without loss of generality, μ is set equal to zero in the sequel of the paper.

An iterative estimation procedure known as the back-fitting algorithm is widely used for estimating functions f^i and is implemented in various scientific softwares (see for example Venables and Ripley, 1994). (Hastie and Tibshirani, 1990) have illustrated the efficiency of this procedure for many practical examples. The convergence results of this kind of estimators are difficult to obtain, see (Buja et al., 1989; Härdle and Hall, 1993; Opsomer and Ruppert, 1997) and more recently, (Mammen et al., 1999) for a decisive contribution.

An alternative way to back-fitting, based on marginal integration, has been independently introduced by (Auestad and Tjøstheim, 1994; Newey, 1994; Linton and Nielsen, 1995). An optimization interpretation connecting the two kinds of estimation methods is given in (Nielsen and Linton, 1998). By comparing the two estimation methods, focusing on finite sample properties, (Sperlich et al., 1999) found many similarities between associated methodologies and statistical performance, highlighting the value of the integration method in various situations, especially when the estimation of functions f^i is mainly concerned instead of the full dimensional regression function (see also Fan et al., 1998, p. 945).

In this paper we focus on the marginal integration method which is based on the following simple remark. Let us set

$$m(x) \;=\; \mathbb{E}\Big[Z/(X^1, \ldots, X^d) = x\Big] \tag{1.2}$$

where $x = (x^1, \ldots, x^d)$.

Since for a nonlinear additive model, $m(x) = \sum_{i=1}^{d} f^i(x^i)$, then we have

$$f^i(x^i) \;=\; \mathbb{E}\Big[m\Big(X^1,\dots,X^{i-1},\, x^i\,, X^{i+1},\dots,X^d\Big)\Big] \tag{1.3}$$

$$\;=\; \int_{\mathbb{R}^{d-1}} m(x)\, p^{\underline{i}}(x^{\underline{i}})\, dx^{\underline{i}} \tag{1.4}$$

where $x^{\underline{i}}$ is the vector $(x^1,\dots,x^{i-1},x^{i+1},\dots,x^d)$ and $p^{\underline{i}}$ denotes the probability density function of the vector $X^{\underline{i}}$.
Moreover, if we consider a weight function w satisfying $\mathbb{E}\left[w(X^{\underline{i}})\right] = 1$, it is also possible to obtain $f^i(x^i)$, but up to an additive constant independent of x^i. Indeed, we have

$$\mathbb{E}\Big[w(X^{\underline{i}})\, m\left(X^1,\dots,X^{i-1},\, x^i\,, X^{i+1},\dots,X^d\right)\Big] = f^i(x^i) + C_w \tag{1.5}$$

In addition, if in integral (1.4) we replace $p^{\underline{i}}$ by any given probability density function q on $\mathbb{R}^{d-1}$, then we derive a similar result:

$$\int m(x)\, q(x^{\underline{i}})\, dx^{\underline{i}} \;=\; f^i(x^i) \;+\; C_q \tag{1.6}$$

In the existing literature, two ways are taken to estimate $f^i(x^i)$: to use a strong law of large numbers to estimate the expectation in (1.3) or (1.5), or to directly estimate the integral in (1.4) or (1.6) by replacing the unknown functions by their corresponding nonparametric estimators.

Let us present previous work in more details organizing the discussion around three different approaches, all involving marginal integration. Let us also specify that, except when explicitely mentioned, the results have been established when $(X^1_n,\dots,X^d_n)_{n\geq 1}$ are i.i.d. and essentially concern pointwise central limit theorem with the one-dimensional nonparametric convergence rate.

Starting from (1.6), (Linton and Nielsen, 1995) introduce the estimator $\int \widehat{m}_n(x)\, q(x^{\underline{i}})\, dx^{\underline{i}}$ where $\widehat{m}_n$ is a kernel based estimator of m. For the case $d = 2$, they prove a pointwise central limit theorem. More recently, (Camlong, 1999; Camlong-Viot et al., 2000) establish asymptotic properties under mixing conditions: results of uniform almost sure convergence and asymptotic normality are obtained, assuming that the probability density function of $(X^1,\dots,X^d)$ is known.

To estimate $f^i(x^i)$ from (1.3), (Linton and Härdle, 1996) consider the estimator $(1/n)\sum_{j=1}^{n} \widehat{m}_n(X^1_j,\dots,X^{i-1}_j,\, x^i\,, X^{i+1}_j,\dots,X^d_j)$ and prove a pointwise central limit theorem. (Linton, 1997) proposes and studies an efficient estimator starting from the previous one.

Starting from (1.5), (Fan et al., 1998) establish the asymptotic normality and a multivariate central limit theorem for weighted estimators of the form $(1/n)\sum_{j=1}^{n} w(X_j^i)\,\widehat{m}_n(X_j^1,\dots,X_j^{i-1},x^i,X_j^{i+1},\dots,X_j^d)$. These results have been extended to nonlinear ARX models by (Masry and Tjøstheim, 1997) using kernel based methods and by (Cai and Masry, 1997) using local polynomials.

Another idea, used in this paper, is to start from (1.4) and define what we call the *global estimator* of f^i, replacing m and $p^{\underline{i}}$ by their corresponding kernel-based estimators.
The scope of this paper is to define and study partial and recombined estimators deduced from the global one.
Indeed, by integrating $\widehat{m}(x)$ over a domain $\mathcal{D}$ of $\mathbb{R}^{d-1}$, instead of $\mathbb{R}^{d-1}$, we can define what we call a *partial estimator* of f^i. In addition, by partitioning $\mathbb{R}^{d-1}$, we can build a family of partial estimators of f^i.

From a non asymptotic perspective, what is the value of such an approach ? Two aspects are interesting. Firstly, the analysis of the model additivity can be based on partial estimators used as data analysis tools: the partial estimators examination can allow to diagnose and localize the lack of additivity of the model. Secondly, the recomposition of partial estimators using weights (for example, depending on x^i and the joint density of the explanatory variables) reflecting more closely the quality of the estimation of the regression function m, leads to define what we call in this paper *recombined estimators*. This strategy can improve the performance for small to moderate sample sizes, which is useful since the curse of dimensionality is not completely eliminated using integration method as mentioned by (Nielsen and Linton, 1998).

The first part of the paper (Sections 2 and 3) introduces the partial and recombined estimators with motivation, illustration by simulations and case study dealing with ozone concentration in Paris area, focusing on practical issues for analysis and prediction of time series using nonlinear additive models.

The second part of the paper (Sections 4 and 5) deal with theoretical results from an asymptotic perspective. We provide almost sure convergence results for the estimators previously introduced. A multivariate central limit theorem is established for partial estimators and an asymptotic test for partial additivity is derived. Proofs of the main results are postponed to appendices.

2. Partial and recombined estimators

2.1. Model and simulation framework

2.1.1. *Notations*

Let us consider the following additive model

$$Z_n \;=\; \sum_{i=1}^{d} f^i(X_n^i) \;+\; \varepsilon_n \;,\; n \in \mathbb{N} \tag{2.1}$$

where each function $f^i : \mathbb{R} \to \mathbb{R}$ satisfies the identifiability hypothesis $\mathbb{E}\left[f^i(X^i)\right] = 0$. The sequence of random vectors $(X_n^1, \cdots, X_n^d)_{n\geq 1}$ is strictly stationary and $(\varepsilon_n)_{n\geq 1}$ is an unobservable noise.
Since each function f^i plays the same role, without loss of generality, we can rewrite model (2.1) under the form

$$Z_n \;=\; f(X_n) \;+\; g(Y_n) \;+\; \varepsilon_n \tag{2.2}$$

where f is one of the f^i and g the sum of the others, (X_n) one of the (X_n^i) and (Y_n) the vector of the others. Of course, functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^{d-1} \to \mathbb{R}$ also satisfy the identifiability hypotheses, i.e. $\mathbb{E}\left[f(X)\right] = 0$ and $\mathbb{E}\left[g(Y)\right] = 0$. The regression function m of (1.2) is rewritten as $m(x,y) = \mathbb{E}\Big[Z/(X,Y) = (x,y)\Big]$.

To make the paper more readable, general material is formulated using form (2.2) and we use form (2.1) when actual models are concerned.

2.1.2. *Simulated models*

In this paper, simulations will illustrate step-by-step the behaviour of the different proposed estimators. The material introduced here will be useful for all the sequel of the section.

We consider two nonlinear additive models of the form (2.1), with d equal to 2 or 3, well suited for illustration purposes since the difficulties arising for large d using marginal integration method are well known. These two models, denoted by $(\mathrm{M1})_\rho$ and $(\mathrm{M2})_\rho$, are defined by

$$(\mathrm{M1})_\rho: \quad Z_n \;=\; f^1(X_n^1) \;+\; f^2(X_n^2) \;+\; \varepsilon_n$$

$$(\mathrm{M2})_\rho: \quad Z_n \;=\; f^1(X_n^1) \;+\; f^2(X_n^2) \;+\; f^3(X_n^3) \;+\; \varepsilon_n$$

where $f^1(x) = x^2 - 1$, $f^2(x) = x/2$, $f^3(x) = |\,x\,|^{1/2} - \mathbb{E}(\left|X^3\right|^{1/2})$, $\varepsilon_n \sim \mathcal{N}\left(0\,,\,0.5^2\right)$ and $(X_n^1, X_n^2, X_n^3) \sim \mathcal{N}\left(0\,,\,\Gamma\right)$ with $\Gamma(j,j) = 1$ and $\Gamma(j,k) = \rho$ for $j \neq k$. Functions f^1, f^2 and f^3 satisfy the identifiability constraint.

These models are examined with $\rho = 0.2$ and $\rho = 0.8$, for moderate sample sizes $n = 200$, 800 and 2000. The functions f^i considered here, have been previously examined by (Linton and Härdle, 1996).
For each model, we simulate 10 realizations of $(X_j^1, \dots, X_j^d, Z_j)_{1 \le j \le n}$ and then obtain 10 realizations of each $\widehat{f}_n^i$, denoted by $\left(\widehat{f}_n^{i(k)}\right)_{1 \le k \le 10}$.
The different estimators introduced below, involve some parameters which are chosen for the simulations as follows. The kernels are direct products of Gaussian densities and the bandwidth in the ith-direction is of the form $\widehat{\sigma}_i n^{-\alpha}$ where $\alpha = 0.2$ and $\widehat{\sigma}_i$ is an estimate of the standard deviation of X^i.

2.2. Global estimator

2.2.1. *Principle*

Using the notations of model (2.2), let us recall the principle leading to the global estimator based on marginal integration. Let us denote by p^Y the probability density function of Y. Then, integrating the regression function m with respect to p^Y leads to

$$\int m(x,y)\, p^Y(y)\, dy \;=\; f(x) \;+\; \mathbb{E}\left[g(Y)\right] \;=\; f(x)$$

Then, starting from $\widehat{m}_n(x,y)$ and $\widehat{p}_n^Y(y)$, two usual kernel-based estimators of $m(x,y)$ and $p^Y(y)$ respectively, we derive, what we call the global estimator of $f(x)$. It is given by

$$\widehat{f}_n(x) \;=\; \int \widehat{m}_n(x,y)\, \widehat{p}_n^Y(y)\, dy \tag{2.3}$$

REMARK 2.1. A slightly modified version of this global estimator converges almost surely to $f(x)$, as proved in Appendix D.

Using simulations, let us illustrate some well-known serious difficulties arising when such a global estimator is used. Let us mention that results obtained for the global estimator will only be used as a reference to help the reader to visualize typical performance (for an extensive study, see Sperlich et al., 1999).

2.2.2. *Illustration*

Let us examine two extreme situations: a model with two weakly dependent explanatory variables, for a small sample size (Fig. 1, Model $(\mathrm{M1})_{0.2}$ with $n = 200$) and a model with three strongly dependent explanatory variables, for a large sample size (Fig. 2, Model $(\mathrm{M2})_{0.8}$ with $n = 2000$). To visualize bias and variance of the estimators, Fig. 1 and Fig. 2 present four curves for each function f^i: in dotted

Figure 1. Global estimators for Model $(\mathrm{M1})_{0.2}$ with $n = 200$.

Figure 2. Global estimators for Model $(\mathrm{M2})_{0.8}$ with $n = 2000$.

line, function f^i to be estimated, restricted to $[-3\,,\,3]$; in solid line, the mean of the $\widehat{f}_n^{i(k)}$, to reduce sampling effects, and, surrounding the last curve, the curves $\widehat{f^i} \pm 2\sigma_{\widehat{f^i}}$ where $\sigma_{\widehat{f^i}}$ is the standard deviation of $\left(\widehat{f}_n^{i(k)}\right)_{1\leq k\leq 10}$.

For Model $(\mathrm{M2})_{0.8}$ the results are of bad quality even if n is quite large (for $n = 2000$, see Fig. 2), especially outside the interval $[-2\,,\,2]$.

For Model $(M1)_{0.2}$, the estimation is quite satisfactory as soon as $n = 200$ (see Fig. 1). In other words, multiplying the sample size by ten does not suffice to compensate the addition of one explanatory variable and the switch from weak to strong dependence. This illustrates that estimation performances depend heavily on the dependence of the explanatory variables and the number of them (see Nielsen and Linton, 1998, p. 221).

In addition to evaluate estimation and prediction errors, we compute two quadratic criteria, defined in Appendix E. The first one is the estimation error criterion calculated for each function f^i and denoted by Sf^i. Quantities Sf^i allow to appreciate the quality of the estimation of f^i: the smaller Sf^i, the better is the estimation. The second one is the prediction error criterion, denoted by SZ. It estimates $\mathbb{V}\mathrm{ar}(\varepsilon)$ and quantifies the quality of the prediction of Z: the closer to $\mathbb{V}\mathrm{ar}(\varepsilon) = 0.5^2$, the better is the prediction.

From Table VI of Appendix E, as expected the larger n, the better are the convergence and prediction results (see quantities Sf^i and SZ). The larger the correlation, the slower is the convergence rate. The results obtained for Models $(M1)_{0.8}$ and $(M2)_{0.8}$ are clearly less satisfactory (for the same sample size) than the results for models $(M1)_{0.2}$ and $(M2)_{0.2}$ respectively. From $d = 2$ to $d = 3$ explanatory variables, the performance are dramatically altered for the high correlation case, but are of the same order of magnitude for the small correlation case.

2.3. Partial estimators

2.3.1. *Principle*

Using the notations of model (2.2), let us present the idea leading to define what we call partial estimators. Let $\mathcal{D}$ be a compact set of $\mathbb{R}^{d-1}$ such that $\mathbb{P}\,[Y \in \mathcal{D}] \neq 0$. Then, integrating $m(x, y)$ with respect to $p^Y(y)$ over $\mathcal{D}$, instead of $\mathbb{R}^{d-1}$, and normalizing the integral by $\mathbb{P}\,[Y \in \mathcal{D}]$, we obtain $f(x)$ up to an additive constant, independent of x. Indeed, we have:

$$\int \frac{\mathbf{1}_{\mathcal{D}}(y)}{\mathbb{P}\,[Y \in \mathcal{D}]}\; m(x, y)\, p^Y(y)\, dy \;=\; f(x) \;+\; C_{\mathcal{D}} \tag{2.4}$$

Then, replacing m and p^Y by their corresponding kernel-based estimators $\widehat{m}_n(x, y)$ and $\widehat{p}_n^{\,Y}(y)$, we define what we call a partial estimator by:

$$\int \frac{\mathbf{1}_{\mathcal{D}}(y)}{\mathbb{P}\,[Y \in \mathcal{D}]}\; \widehat{m}_n(x, y)\, \widehat{p}_n^{\,Y}(y)\, dy \tag{2.5}$$

REMARK 2.2. This estimator almost surely converges to $f(x) + C_{\mathcal{D}}$ (see Theorem 4.1 in Paragraph 4.3).

In order to build an "unbiased" version of the partial estimator (2.5), we can use the following remark: since the constant $C_{\mathcal{D}}$ does not depend on x, we have for any given real x_0,

$$C_{\mathcal{D}} \;=\; \int \frac{\mathbf{1}_{\mathcal{D}}(y)}{\mathbb{P}\left[Y \in \mathcal{D}\right]}\; m(x_0, y)\, p^Y(y)\, dy \;-\; f(x_0) \tag{2.6}$$

Therefore, using this relation combined with (2.5) and estimating $f(x_0)$ by $\widehat{f}_n(x_0)$ and $\mathbb{P}\left[Y \in \mathcal{D}\right]$ by $\int_{\mathcal{D}} \widehat{p}_n^Y(y)\, dy$, an "unbiased" partial estimator of function f is given by:

$$\widehat{f}_n^{\mathcal{D}}(x) = \int \frac{\mathbf{1}_{\mathcal{D}}(y)}{\int_{\mathcal{D}} \widehat{p}_n^Y(y) dy} \Big(\widehat{m}_n(x, y) - \widehat{m}_n(x_0, y)\Big) \widehat{p}_n^Y(y) dy + \widehat{f}_n(x_0) \tag{2.7}$$

REMARK 2.3. As proved by Lemma A.1 (see Appendix A), when $\mathcal{D}$ is a compact set of $\mathbb{R}^{d-1}$, we have

$$\int_{\mathcal{D}} \widehat{p}_n^Y(y)\, dy \;\xrightarrow[n\to\infty]{a.s.}\; \mathbb{P}\left[Y \in \mathcal{D}\right] \tag{2.8}$$

Therefore, combining the almost sure convergence result of (2.5) with result (2.8), we derive that $\widehat{f}_n^{\mathcal{D}}(x) \xrightarrow[n\to\infty]{a.s.} f(x)$, as soon as $\widehat{f}_n(x_0)$ almost surely converges to $f(x_0)$.

REMARK 2.4. This strategy to obtain "unbiased" partial estimator depends on the choice of x_0. An other way can also be used to obtain constant $C_{\mathcal{D}}$: as $\mathbb{E}\left[f(X)\right] = 0$, we infer from (2.4) that

$$C_{\mathcal{D}} \;=\; \int \Big(\int \frac{\mathbf{1}_{\mathcal{D}}(y)}{\mathbb{P}\left[Y \in \mathcal{D}\right]}\; m(x, y)\, p^Y(y)\, dy \Big) p^X(x)\, dx \tag{2.9}$$

where p^X is the probability density function of X. Then, estimating the integral leads to a non-localized estimation of the additive constant. This alternative procedure can be useful when the actual data are far from the realizations of an additive model.

Finally, let us mention that by partitioning $\mathbb{R}^{d-1}$, we can define not only one but a family of partial estimators of f, in order to recombine them from a non-asymptotic perspective (see Paragraph 2.4). Let us be a little bit more precise. Let $\mathcal{D}_1, \ldots, \mathcal{D}_q$ be a partition of $\mathbb{R}^{d-1}$ such that for any j, $\mathbb{P}\left[Y \in \mathcal{D}_j\right] \neq 0$. Then, integrating over each $\mathcal{D}_j$ instead of $\mathcal{D}$, we can build q "unbiased" partial estimators of $f(x)$, denoted by $\widehat{f}_n^{\mathcal{D}_j}(x)$ and defined by (2.7), replacing $\mathcal{D}$ by $\mathcal{D}_j$.

The terminology of partial estimators comes from the following reconstruction property :

$$\sum_{j=1}^{q} \Big(\int_{\mathcal{D}_j} \widehat{p}_n^{\,Y}(y)\, dy \Big)\, \widehat{f}_n^{\mathcal{D}_j}(x) \;=\; \widehat{f}_n(x) \tag{2.10}$$

REMARK 2.5. More generally, for any given positive real numbers $(\alpha_j)_{1\leq j\leq q}$ summing to 1, $\sum_{j=1}^{q} \alpha_j\, \widehat{f}_n^{\mathcal{D}_j}(x)$ is an estimator of $f(x)$.

REMARK 2.6. The basic definition of partial estimators allows to handle arbitrary complex domains $\mathcal{D}_j$ of $\mathbb{R}^{d-1}$. We restrict our attention on domains of the form $\mathcal{D}_j = \bigotimes_{k=1}^{d-1} (a_{j,k}\,,\, b_{j,k})$ for obvious computational reasons. On one hand, this choice leads to simple evaluations of integrals over $\mathcal{D}_j$. On the other hand, another attractive feature of this choice, of course more relevant for a real world problem, is that a partial estimator associated with such a domain $\mathcal{D}_j$ has a simple interpretation in terms of the initial variables taken separately instead of a complicated one involving new variables, more difficult to interpret (for example, linear combinations of the set of the covariables). The choice of the thresholds $(a_{j,k}\,,\, b_{j,k})$ for a given k is based on the analysis of the kth covariable. For simulations, we select those leading to the same number of observations within each interval, for a prescribed small number of intervals. For real world problems, more meaningful thresholds could be derived from the knowledge of the case studied (see paragraph 3.2 for an example).

2.3.2. *Illustration*

Let us briefly consider a single sample generated from Model $(\mathrm{M1})_{0.2}$ with $n = 800$. Fig. 3 displays the partial estimators of function f^1 and contains six plots. At the top left, X^2 versus X^1: the horizontal lines are defined by the four intervals $\mathcal{D}_j$. At the top right, the four partial estimators are superimposed. Each of the four plots located at the bottom of the figure contains a partial estimator.

The intervals $(\mathcal{D}_j)_{1\leq j\leq 4}$ defining the partial estimators of f^1 are chosen in such a way that the estimations of probabilities $\mathbb{P}\left[X^2 \in \mathcal{D}_j\right]$ are equal to $1/4$. The additive constant defined in (2.6) is estimated by taking x_0 equal to 0.

The four partial estimators are close to each other and each of them is as satisfactory than the global estimator.

Of course, from a practical point of view, the partial estimators examination can allow to diagnose and localize the lack of additivity of the modelized phenomenon. This point is illustrated in Section 3,

Figure 3. Partial estimators of f^1 for model $(\mathrm{M1})_{0.2}$ with $n = 800$.

dedicated to air pollution analysis. In addition, partial estimators could also be used to build a test for partial additivity (see Section 5).
Next, we shall see that, when the convenient model is additive, the partial estimators can be used to improve the estimation for small to moderate sample size, by defining what we call *recombined estimators*.

2.4. Recombined estimators

2.4.1. *Motivation*

The inspection of the partial estimators on a simple situation presented below, suggests a new estimator based on a better way of taking the joint distribution of the explanatory variables into account.

Let us consider a single sample generated from Model $(\mathrm{M1})_{0.8}$ with $n = 800$. Fig. 4 displays the global estimators of f^1 and f^2 and shows that these functions are not well estimated except within $[-1\,,\,1]$.

This unsatisfactory behaviour of the global estimators is explained by the relatively small sample size and the strong dependence between the explanatory variables. Since the variables X^1 and X^2 are strongly

Figure 4. Global estimators for model $(\mathrm{M1})_{0.8}$ with $n = 800$.

Figure 5. Model $(\mathrm{M1})_{0.8}$ with $n = 800$. Explanatory variables X^2 versus X^1.

positively correlated, the observations are concentrated around the first bisecting line (see Fig. 5). So, one cannot expect an estimate of m of good quality far from this line. Then, the estimators of f^1 and f^2, built by integrating $\widehat{m}_n$ with respect to the estimated marginal density of X^2 for f^1 and of X^1 for f^2, are of good quality only within the square $[-1\,,\,1] \times [-1\,,\,1]$.

This illustrates once again, one of the well-known limitations of the marginal integration estimators for highly dependent explanatory variables and moderate sample size.

Nevertheless, the inspection of the partial estimators of each function (see Fig. 6 for f^1), leads to a more subtle diagnosis: each partial estimator is globally not satisfactory but is of very good quality on a proper domain.

For example, let us focus on the first partial estimator of f^1 displayed on the left of the second row of Fig. 6. This estimator behaves correctly on the interval $[-2\,,\,0.4]$ but is of very bad quality otherwise. This can be explained by recalling that this partial estimator comes from the integration of $\widehat{m}_n$ over the domain $[-3\,,\,-0.8]$ of X^2, for which, as it can be seen on the top left of Fig. 6, the observations of X^1 are mainly concentrated on $[-2\,,\,0.4]$.

Figure 6. Partial estimators of f^1. Model $(\mathrm{M1})_{0.8}$ with $n = 800$.

Therefore, as the global estimator of f^1 is obtained by combining linearly the four partial estimators, with equal weights and without reference to the joint distribution of X^1 and X^2, the less accurate partial estimators contaminate the sum. This phenomenon leads to an unsatisfactory global estimator.
To improve the quality of the global estimator, a basic idea is to differently combine the partial estimators using weights reflecting the quality of each of them.

2.4.2. *Definition*
Let us use the notations of model (2.2). Let us consider $\mathcal{D}_1, \dots, \mathcal{D}_q$ a partition of $\mathbb{R}^{d-1}$ such that for any j, $\mathbb{P}[Y \in \mathcal{D}_j] \neq 0$. Starting from q partial estimators of f denoted by $\widehat{f_n}^{\mathcal{D}_j}$ and given by (2.7), we define the estimator $\sum_{j=1}^{q} \alpha_j(x)\, \widehat{f_n}^{\mathcal{D}_j}(x)$ where $\alpha_j(x)$ are data driven weights such that $\sum_{j=1}^{q} \alpha_j(x) = 1$.

In order to improve the quality of the global estimator, we propose to choose weights depending on the joint density of the explanatory variables as well as the location of the estimation point x. A convenient choice is to set $\alpha_j(x) = \mathbb{P}[Y \in \mathcal{D}_j | X = x]$, since the greater $\alpha_j(x)$, the better is (locally) the estimation of $\widehat{m}_n$ and then, the better is the estimation of $\widehat{f}_n^{\mathcal{D}_j}(x)$.

The weights $\alpha_j(x)$ are estimated by $\widehat{\alpha}_j(x)$ defined by

$$\widehat{\alpha}_j(x) = \Big(\int_{\mathcal{D}_j} \widehat{p}_n^{XY}(x,y)\, dy \Big) \Big/ \widehat{p}_n^X(x)$$

where $\widehat{p}_n^{XY}$ and $\widehat{p}_n^X$ are usual kernel-based estimators of p^{XY} and p^X the probability density functions of (X, Y) and X respectively. This leads to what we call a *recombined estimator*:

$$\widehat{f}_n^{re}(x) \;=\; \sum_{j=1}^{q} \widehat{\alpha}_j(x)\, \widehat{f}_n^{\mathcal{D}_j}(x) \tag{2.11}$$

REMARK 2.7. As a consequence of Remark 2.3 to obtain an almost sure convergence result for the recombined estimator, the domains $\mathcal{D}_j$ must be compact subsets of $\mathbb{R}^{d-1}$. This constraint is of course not restrictive from a practical point of view, but an obvious modification of the weights $\widehat{\alpha}_j(x)$ is needed. Let us choose q arbitrary compact domains $\mathcal{D}_j$ and consider the weights $\widehat{\beta}_j(x)$ defined by

$$\widehat{\beta}_j(x) = \int_{\mathcal{D}_j} \widehat{p}_n^{XY}(x,y)\, dy \Big/ \int_{\bigcup_{j=1}^{q} \mathcal{D}_j} \widehat{p}_n^{XY}(x,y)\, dy \tag{2.12}$$

From Lemma A.1 of Appendix A, $\widehat{\beta}_j(x)$ converges to the deterministic counterpart of the right hand side of (2.12), and then, the slightly modified recombined estimator $\displaystyle\sum_{j=1}^{q} \widehat{\beta}_j(x)\, \widehat{f}_n^{\mathcal{D}_j}(x) \xrightarrow[n\to\infty]{a.s.} f(x)$ as soon as $\widehat{f}_n(x_0)$ is a consistent estimator of $f(x_0)$.

2.4.3. *Illustration*

The results obtained using the recombined estimator on the example considered in paragraph 2.4.1, can be visualized in Fig. 7 and are to be compared with Fig. 4. Functions f^1 and f^2 are now very well estimated in this hard context of high dependence and moderate sample size.

To illustrate the performance of the recombined estimators, one can find in Table VII of Appendix E the results obtained for the models previously examined for the global estimator. It has to be compared with Table VI of Appendix E.

Figure 7. Recombined estimators for model $(\mathrm{M1})_{0.8}$ with $n = 800$.

Figure 8. Recombined estimator. Model $(\mathrm{M2})_{0.8}$ with $n = 2000$.

Firstly, the recombined estimator does not alter the good performance obtained by the global one in the weakly dependent case, ie. for Models $(\mathrm{M1})_{0.2}$ and $(\mathrm{M2})_{0.2}$, and slightly improve it. Secondly, it tremendously improve those obtained in the strongly dependent case, ie. for Models $(\mathrm{M1})_{0.8}$ and $(\mathrm{M2})_{0.8}$: the performance become close to those obtained for Models $(\mathrm{M1})_{0.2}$ and $(\mathrm{M2})_{0.2}$. For example, for $(\mathrm{M2})_{0.8}$ with $n = 800$, SZ which estimates $\mathbb{V}\mathrm{ar}(\varepsilon) = 0.25$, is equal to 0.74 in Table VI and becomes 0.3 here. The quantities Sf^i also strongly decrease.
As an illustration, Fig. 8 shows the good performance of the recombined estimator for Model $(\mathrm{M2})_{0.8}$, which is hard to estimate even for a large sample size of $n = 2000$. When it is compared to Fig. 2,

the improvement is important and becomes huge outside the interval $[-2\,,\,2]$.

3. Analysis and prediction of ozone concentration

This section is dedicated to a real world problem: the analysis and prediction of ozone concentration. The partial estimators are used for data analysis purposes whereas the recombined estimators are used to build the prediction model. Details about forecasting ozone peaks in Paris area using additive models, can be found in (Bel et al., 1999; Bel et al., 2001) which precisely describe the entire project performed with the support of Airparif, the air pollution organism for Paris. This problem has been previously examined in some other cities all over the world, see for example (Hastie and Tibshirani, 1990). Let us mention that in Paris area, different models have been used to investigate the forecasting ozone peaks problem, including the fully nonparametric one and CART model for instance. The comparison between these different approaches shows that the simple nonlinear additive model described below, captures the main features of the complex underlying dynamics. In addition, it is well suited to correctly predict ozone concentration exceedances of extreme thresholds for which only few observations are available, which is the main difficulty.

3.1. The model

We focus on the following basic model which is currently used to forecast the daily maximum of ozone concentration in Paris area:

$$\begin{aligned} \mathrm{ozon}(j) \;=\; & f_1(\mathrm{temp}(j)) \;+\; f_2(\mathrm{wind}(j)) \\ & + f_3(\mathrm{mozon}(j-1)) \;+\; \mu \;+\; e(j) \end{aligned} \tag{3.1}$$

where j is the index of the day, $\mathrm{ozon}(j)$ is the maximum of ozone concentration (μg/m^3) during day j, $\mathrm{temp}(j)$ is the maximum of temperature (°C) during day j, $\mathrm{wind}(j)$ is the mean of the wind speed (m/s) during the afternoon of day j, $\mathrm{mozon}(j-1)$ is the regional spatial mean of maximum ozone concentration (μg/m^3) during day $j-1$; μ is a constant and $e(j)$ is an error term.
This model is estimated using the realizations and is then used for real-time prediction purposes following the perfect prognosis strategy (Wilks, 1995) by replacing the unobservable explanatory variables by their predictions coming from Météo France, the french meteorological company.

The bandwidths for nonparametric estimators are of the classical form $\widehat{\sigma}_X n^{-\alpha}$ with α around 0.2 and where $\widehat{\sigma}_X$ is the empirical standard deviation of the concerned explanatory variable.

3.2. Choice of the partial estimators

The regions leading to the different partial estimators are defined by interval limits chosen from the marginal distribution of each explanatory variable and according to the knowledge of the problem. These sensible limits divide each variable range in three intervals : low, medium, and large. Limits are of 16 and 24°C for the temperature, 2 and 4 m/s for the wind speed and 43 and 85 μg/m^3 for the previous day ozone.

3.3. Ozone data analysis using partial estimators

For a typical measurement station located in the thirteen arrondissement of Paris (abbreviated p13), we consider the partial estimators of the effects of the temperature, the wind and the previous day ozone, based on the summer days (about 1200 days) during the period 1992-1998 (see Figures 9 and 10). Each figure contains twelve plots. At the top left, one can find the histogram of the considered explanatory variable and in the middle, the data are displayed in the plane of the two other explanatory variables, with the delimited domains $\mathcal{D}_j$. At the top right, the nine partial estimators are superimposed. Each of the nine plots located at the bottom of the figure contains a partial estimator. The nine plots are organized as a 3×3 matrix, reflecting the nine integration regions defined following the rule specified in Paragraph 3.2.

Let us analyze the effect of the temperature (see Fig. 9). The increasing global shape is a common feature of all partial estimators. The estimated partial effects are negative for low and medium temperatures and become positive for large temperatures.

Despite the homogeneous global shape, the intensity of the estimated partial effect is highly varying, invalidating the global additivity of the modelized phenomenon. The larger the importance of the estimated partial effect of the temperature, the more polluted is the previous day. Comparing the extreme curves located at the top left and at the down right of the 3×3 square displaying the partial estimators, one can point out a large variation of the intensity of the estimated effect: from -30 to $100\,\mu$g/m^3 for the partial estimator associated to a low wind speed and a polluted previous day and on the other hand, from -20 to $20\,\mu$g/m^3 for the partial estimator associated to a large wind speed and an unpolluted previous day. Then, a global effect obtained by combining all these partial effects will underestimate the effect of

Figure 9. Partial estimators of the effect of the temperature.

the temperature for a low wind speed and overestimate it for a large wind speed.
The columns of the 3×3 square displaying the partial estimators, are more homogeneous that the rows. Then it follows that the estimated partial effects of the temperature for similar wind speed amount are reasonably compatible with the additivity property of the model.

Let us now analyze the effect of the wind speed (see Fig. 10). The inspection of the general shape of the curves immediately leads to three different behaviors, since the columns of the 3×3 square displaying the partial estimators are homogeneous, but the rows are not. For low temperatures (first column), the effect of the wind speed is very small. This is the main point. It should be noted that it is slightly increasing from -10 to 10, which is counterintuitive. This probably comes from the fact that for low temperatures, the measured ozone is the ozone naturally present in the air, whose concentration decreases with the NO_2 titration when the wind speed is small (Bel et al., 2001). For medium temperatures (column 2), the effect of the wind speed is also very small, the estimation is close to 0 for wind speeds less than 7m/s

Figure 10. Partial estimators of the effect of the wind speed.

and slightly negative otherwise. But this last element must be considered with caution whereas the number of observations in this last zone is small (see the wind speed histogram at the top left of Fig. 10). So, for temperatures less than 24 °C, the effect of the wind speed is very small and then the wind speed is not a significant explanatory variable of the model. On the other hand, for large temperatures (column 3), the effect of the wind speed is different from 0, decreasing from slightly positive values to -40, which is in accordance with the physical intuition since the wind dissipates pollution. Therefore the wind speed is a significant variable in the model as soon as the temperature is sufficiently large to allow the photo-chemical ozone production and its effect is negative. Let us note that the estimated partial effects of the wind speed for similar temperatures seem compatible with the additivity property of the model.

For the last explanatory variable, the general shape of the estimated effects of the previous day ozone is homogeneous. The estimators are close to each other, leading to a full compatibility with the additiv-

Figure 11. Global and recombined estimators.

ity property. As expected, the estimated effects increase from -20 for unpolluted previous days, to 40-60 for polluted previous days.

To summarize, the partial estimators are a useful tool to analyze ozone concentration. Indeed, they reveal the homogeneity of the shape of the effects and a kind of conditional additivity of the modelized phenomenon. In addition they suggest to model the interaction between temperature and wind speed. This is not considered here because this does not modify the actual performance evaluated as defined in the next paragraph.

3.4. Ozone prediction using a recombined estimator

For the same measurement station, we consider the global and recombined estimators of the effects of the temperature, wind speed and previous day ozone, displayed in Fig. 11.
The constant μ in (3.1) is estimated by the mean of the ozone concentrations which is equal to $73\,\mu\mathrm{g/m}^3$. Let us compare the two estimators.

The effects of the temperature for the two estimators are identical for temperatures less than 24°C. Otherwise, the recombined estimator grows quicker and the difference reaches 40 $\mu\mathrm{g/m}^3$ for a temperature of 38°C.
The effects of the previous day ozone are very similar for the two estimators. This is a consequence of the previously mentioned homogeneity of the partial estimators.
The effects of the wind speed are the same up to 4 m/s and are very small. Next, the global estimator decreases up to -20 whereas the recombined estimator is very small. Of course, the wind speed effect fitted by the recombined estimator is not satisfactory from a physical point of view. It should be easily improved by considering another recombined estimator using a constrained weighting scheme only built with the three partial estimators corresponding to large temperature

Table I. Alarms table

<table>
<tr><td colspan="2" rowspan="2"></td><th colspan="3">Estimated</th></tr>
<tr><th>Level 0</th><th>Level 1</th><th>Level 2</th></tr>
<tr><th rowspan="3">Measured</th><th>Level 0</th><td>Good alarms</td><td colspan="2">False alarms</td></tr>
<tr><th>Level 1</th><td></td><td>Good alarms</td><td></td></tr>
<tr><th>Level 2</th><td colspan="2">Missed alarms</td><td>Good alarms</td></tr>
</table>

Table II. Simplified alarms table

	Estimated 0	**Estimated 1 or 2**
Measured 0	t_{00}	t_{01}
Measured 1 or 2	t_{10}	t_{11}

(right column of Fig. 10). This point will be addressed at the end of this paragraph.

Let us first illustrate the gap between the global estimator and a crude version of the recombined estimator, by comparing the predictions given by the models, what is more interesting from the decisional viewpoint. More precisely, we will only compare the estimated values to the actual values, without replacing the temperature and wind speed measures by their predictions. In this application, the crucial objective is to correctly predict alarms defined by the maximum of ozone concentration exceedances of two thresholds: $130\,\mu\text{g/m}^3$ (level 1) and $180\,\mu\text{g/m}^3$ (level 2). The days of level 0 are those for which the maximum of ozone concentration is less than $130\,\mu\text{g/m}^3$. Criteria closely related to the decisional problem are defined starting from the contingency table I.
The rows are labelled by the measured levels and the columns by the estimated levels. The concerned days are then dispatched in the table. The levels correctly estimated are on the diagonal, otherwise we have the false alarms and the missed alarms. Then the ideal table is diagonal. Of course and fortunately for health, all but a few days are of level 0. Starting from table I, let us define three usual synthetic criteria coming from the simplified binary contingency table II.
The first criterion is the missed alarms rate given by $t_{10}/(t_{11}+t_{10})$, the second one is the false alarms rate given by $t_{01}/\,(t_{11}+t_{01})$. The last one, measuring the rate of good alarms and called the threat score, is widely used in meteorology and pollution and is defined by $t_{11}/\,(t_{11}+t_{10}+t_{01})$.

Table III. Alarms table for the global and recombined estimators

		Global Estimator			*Recombined Estimator*		
		Estimated			Estimated		
		level 0	level 1	level 2	level 0	level 1	level 2
Meas.	level 0	494	1	0	479	16	0
	level 1	60	9	0	19	48	2
	level 2	7	8	0	0	11	4

Table IV. Alarms table (simplified) for the global and recombined estimators

	Global Estimator		*Recombined Estimator*	
	Estimated 0	Est. 1 or 2	Estimated 0	Est. 1 or 2
Measured 0	494	1	479	16
Meas. 1 or 2	67	17	19	65

The smaller the first two criteria, the better is the prediction model. The greater the threat score, the better is the prediction model. As it can be seen, the value t_{00} does not contribute to any of these quantities since it is considered to be easy to correctly predict the level of days of level 0, so these criteria focus on nontrivial, hard and interesting situations.

The performance evaluation is based on the days between 1994 and 1998. Tables III to V give the performance of the global and recombined estimators.

For the global estimator, the obtained contingency table is typical of an underestimation for interesting days despite the fact that the mean of absolute errors is satisfactory since it is close to the measurement error. None of the days of level 2 are detected and the table of alarms is

Table V. Performance indices for the global and recombined estimators

	Global estimator	*Recombined estimator*
Missed alarms rate	79.8%	22.6%
False alarms rate	5.6%	19.8%
Threat score	20.0%	65.0%
Mean of absolute errors	17.7	15.6

Figure 12. Global and recombined estimators.

highly non-symmetric: 67 missed alarms for 1 false alarm. The threat score of 20% is small and 7 days of level 2 are estimated as levels 0.

With the recombined estimator, the performance are tremendously improved. Four levels 2 are detected and the alarms table is near symmetric: 19 missed alarms for 16 false alarms. A very satisfactory threat score of 65% is obtained with a small mean of absolute errors of 16. Lastly, none of the days of level 2 are estimated as levels 0.

Now, let us turn back to the wind speed effect given by the recombined estimator which is not satisfactory from a physical point of view and which is the main evidence of the lack of additivity. Since the model is built for the prediction of extreme values of ozone concentration, previous analysis leads to consider a constrained weighting scheme only built with the three partial estimators corresponding to large temperature (right column of Fig. 10), leading to the new effect displayed in Fig. 12. The shape of this new recombined estimator is decreasing, which is physically consistent: the wind pushes pollution outside of Paris area.

4. Asymptotic results

For nonlinear additive models written under the form

$$Z_n \;=\; f(X_n) \;+\; g(Y_n) \;+\; \varepsilon_n \tag{4.1}$$

where $\mathbb{E}\,[f(X)] = 0$ and $\mathbb{E}\,[g(Y)] = 0$, this section presents some asymptotic properties of the marginal integration estimator $\widehat{f}_n^w(x)$ defined by

$$\widehat{f}_n^w(x) \;=\; \int w(y)\,\widehat{m}_n(x,y)\,\widehat{p}_n^{\,Y}(y)\,dy \tag{4.2}$$

where w is a compactly supported function satisfying $\mathbb{E}[w(Y)] = 1$. This estimator is directly connected to the partial estimator (2.5) considered in Paragraph 2.3, by choosing w as the normalized characteristic function of domain $\mathcal{D}$, ie. $w(y) = \mathbf{1}_{\mathcal{D}}(y) \,/\, \mathbb{P}[Y \in \mathcal{D}]$. From a theoretical point of view, introduction of the compactly supported weight function w ensures the existence of the variance of $\widehat{f}_n^w(x)$. If w is dropped, the classical limitation pointed out by Linton occurs (for an example see Linton, 1997, p. 471).

4.1. Model assumptions

In model (4.1), $(X_n, Y_n)_{n\geq 1}$ is a strictly stationary sequence of random vectors and $(\varepsilon_n)_{n\geq 1}$ is an unobservable noise, supposed to be independent of $(X_n, Y_n)_{n\geq 1}$. In addition, we assume:

Assumptions [M].

[M1] *Z has a finite moment of order $\mu > 4$;*

[M2] *the regression function m is C^δ-class with bounded partial derivatives of order δ, where δ is an integer ≥ 2;*

[M3] *the sequence of random vectors $(X_n, Y_n)_{n\geq 1}$ is β-mixing with exponential rate (see* Doukhan, 1994, *for a complete review of mixing notions);*

[M4] *p^{XY}, the probability density function of (X_n, Y_n), is C^δ-class with bounded partial derivatives of order δ;*

[M5] *$(\varepsilon_n)_{n\geq 1}$ is a sequence of independent and identically distributed random variables with zero-mean and variance σ^2.*

These assumptions are classical for this kind of models, except for the third one which allows the explanatory variables to be mixing instead of independent, as it is frequently required.

4.2. The kernel-based estimators

Let us present the kernel-based estimators involved for defining $\widehat{f}_n^w$ in (4.2). The estimator $\widehat{m}_n$ of the regression function m is given by:

$$\widehat{m}_n(x,y) \;=\; \widehat{\ell}_n(x,y)/\widehat{p}_n^{XY}(x,y) \qquad (4.3)$$

with

$$\widehat{\ell}_n(x,y) \;=\; \frac{1}{n}\sum_{k=1}^{n} K^{(1)}_{h_{1,n}}(X_k-x)\; K^{(2)}_{h_{2,n}}(Y_k-y)\; Z_k \tag{4.4}$$

$$\widehat{p}_n^{XY}(x,y) \;=\; \frac{1}{n}\sum_{k=1}^{n} K^{(1)}_{h_{1,n}}(X_k-x)\; K^{(2)}_{h_{2,n}}(Y_k-y) \tag{4.5}$$

where, for a s-dimensional kernel K, $K_h(t) = (1\,/\,h^s)\,K(t\,/\,h)$.
The numerator of $\widehat{m}_n$, denoted by $\widehat{\ell}_n$, estimates the function ℓ defined by $\ell = m \times p^{XY}$, whereas the denominator $\widehat{p}_n^{XY}$ estimates the probability density function p^{XY}.
The kernel-based estimator $\widehat{p}_n^{Y}$ of p^Y is given by:

$$\widehat{p}_n^{Y}(y) \;=\; \frac{1}{n}\sum_{k=1}^{n} K^{(3)}_{h_{3,n}}(Y_k-y)$$

The kernels $K^{(1)}:\mathbb{R}\to\mathbb{R}$ and $K^{(2)}$, $K^{(3)}:\mathbb{R}^{d-1}\to\mathbb{R}$ are subjected to the following assumptions.

Assumptions [K].

- $K^{(1)}$ *is a one-dimensional kernel of order greater than* δ *;*
- $K^{(2)}$ *is a* $(d-1)$*-dimensional kernel of order greater than* δ *;*
- $K^{(3)}$ *is a positive* $(d-1)$*-dimensional kernel of order 2,*

where a s*-dimensional kernel of order* γ *is a compactly supported symmetric Lipschitz function* K*, from* $\mathbb{R}^s$ *to* $\mathbb{R}$*, integrating to 1, such that for* $\ell = 1,\dots,\gamma-1$ *and any* $i_1,\dots,i_\ell \in \{1,\dots,s\}$*,*

$$\int\Big(\prod_{j=1}^{\ell} t_{i_j}\Big)K(t_1,\dots,t_s)\,d(t_1,\dots,t_s) \;=\; 0.$$

The bandwidths $h_{1,n}$, $h_{2,n}$ and $h_{3,n}$ are sequences of positive real numbers decreasing to 0, such that $n\,h_{1,n}h_{2,n}^{d-1}\to\infty$ and $n\,h_{3,n}^{d-1}\to\infty$. In addition, according to the context, these bandwidths will satisfy some of the following conditions:

[B1] $\log^2 n = o\left(n\,h_{1,n}h_{2,n}^{d-1}\right)$ and $\log^2 n = o\left(n\,h_{3,n}^{d-1}\right)$;
[B2] $n^{4/\mu}\log^3 n \;=\; O\Big(n\,h_{1,n}h_{2,n}^{d-1}\log(h_{1,n}^{-1}h_{2,n}^{-(d-1)}\,n^{4/\mu})\Big)$;
[B3] $n\,h_{1,n}^{1+2\delta} \;=\; o(1)$, $n\,h_{1,n}h_{2,n}^{2\delta} \;=\; o(1)$, $n\,h_{1,n}h_{3,n}^{4} \;=\; o(1)$;
[B4] $h_{2,n}^{6\delta} = o(h_{1,n})$ and $n\,h_{1,n}^{4/3}\to\infty$;
[B5] $\log^4 n = o\left(n\,h_{1,n}\,h_{2,n}^{2(d-1)}\right)$, $\log^4 n = o\left(n\,h_{2,n}^{d-1}\,h_{3,n}^{d-1}\right)$.

REMARK 4.1. In Theorems 4.1 and 4.2 stated below, if we replace [M1] by the stronger moment condition [M1bis], then the bandwidth condition [B2] could be relaxed and replaced by [B2bis], where

[M1bis] $\mathbb{E}[\exp(a\,|Z|^b)] < \infty$ *for some* $a > 0$ *and* $b > 0$.

[B2bis] $(\log n)^{3+2/b} = O\Big(n\, h_{1,n} h_{2,n}^{d-1} \log(h_{1,n}^{-1} h_{2,n}^{-(d-1)}\,(\log n)^{2/b})\Big)$.

REMARK 4.2. If $(X_n, Y_n)_{n\geq 1}$ are independent instead of mixing, then condition [B4] is cancelled. In addition, if w is of class $\mathcal{C}^\delta$ instead of $\mathcal{C}^2$, then we can take $K^{(3)} = K^{(2)}$ and $h_{3,n} = h_{2,n}$ leading to cancel the conditions $n\, h_{1,n} h_{3,n}^4 = o(1)$ of [B3] and $\log^4 n = o\left(n\, h_{2,n}^{d-1}\, h_{3,n}^{d-1}\right)$ of [B5].

4.3. Asymptotic results

The following theorem gives an almost sure convergence result for $\widehat{f}_n^w$, allowing us to derive almost sure convergence properties for partial and recombined estimators.

THEOREM 4.1. *Assume that [M], [K], [B1] and [B2] hold. Then,*

$$\widehat{f}_n^w(x) \xrightarrow[n\to\infty]{a.s.} f(x) \ + \ C_w$$

where $C_w = \mathbb{E}\,[w(Y)\,g(Y)]$.

REMARK 4.3. If we choose $w(y) = \mathbf{1}_{\mathcal{D}}(y)\,/\,\mathbb{P}\,[Y\in\mathcal{D}]$, then $\widehat{f}_n^w(x)$ reduces to a partial estimator as previously mentioned and in addition, if w is such that $\mathbb{P}\,[Y\in\mathcal{D}]$ is close to 1, then C_w is close to 0 since $\mathbb{E}\,[g(Y)] = 0$.
Of course, constant C_w is unknown but can be estimated following the idea used in Paragraph 2.3 to propose unbiased versions of partial estimators.

Now, let us present a central limit theorem for $\widehat{f}_n^w(x)$.

THEOREM 4.2. *Assume that [M], [K] and [B1] to [B5] hold. Assume also that* w *is* C^2*-class. Then,*

$$G_n(x) = \sqrt{n\, h_{1,n}}\,\Big(\widehat{f}_n^w(x) - f(x) - C_w\Big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(0\,,\,v(x)\Big) = G(x)$$

where $C_w = \mathbb{E}\,[w(Y)\,g(Y)]$ *and* $v(x) = \sigma^2 \|K^{(1)}\|_2^2 \displaystyle\int \frac{(w(y)\,p^Y(y))^2}{p^{XY}(x,y)}\,dy.$
Besides, for $x_1, \cdots, x_q$, q *distinct points of* $\mathbb{R}$,

$$\Big(G_n(x_1), \cdots, G_n(x_q)\Big) \xrightarrow[n\to\infty]{\mathcal{L}} \Big(G(x_1), \cdots, G(x_q)\Big)$$

where $G(x_1), \cdots, G(x_q)$ *are independent.*

PROOFS. The proofs of Theorem 4.1 and Theorem 4.2 are respectively given by part 1 and by parts 2 and 3 of Theorem B.1 (see Appendix B), setting $S_n(x) = \widehat{f}_n^w(x)$ and taking function w such that $\mathbb{E}[w(Y)] = 1$. Function $S(x)$ is then equal to $f(x) + C_w$.

REMARK 4.4. The central limit theorem for $\widehat{f}_n^w(x)$ holds with the one-dimensional nonparametric rate which is the expected result in the context of estimation for nonlinear additive models. But as usual in the marginal integration context, bandwidth conditions depend on the regularity δ and the dimension d.

REMARK 4.5. For additive models, an estimator of f using marginal integration is said to be efficient (following Linton, 1997) if its variance $v(x)$ attains the optimal variance $v^*(x) = \sigma^2 \|K^{(1)}\|_2^2 / p^X(x)$. For their estimator, (Fan et al., 1998) mention that one can choose a weight function w achieving, at least theoretically, the optimal variance $v^*(x)$. In our case, this optimal weight function is defined by $w^*(y) = p^{XY}(x,y) / p^X(x)\, p^Y(y)$.
Unfortunately, this weighting method is not appropriate from a practical point of view since w^* is unknown and must be estimated by $\widehat{w}_n^*(y) = \widehat{p}_n^{XY}(x,y) / \widehat{p}_n^X(x)\, \widehat{p}_n^Y(y)$. When, we replace w by $\widehat{w}_n^*$ in (4.2), we obtain the following estimator of f

$$\frac{1}{n\, h_{1,n}\, \widehat{p}_n^X(x)} \sum_{k=1}^{n} K^{(1)}\Big((X_k - x)/h_{1,n}\Big)\, Z_k$$

which is obviously not efficient.

5. Towards a test for partial additivity

The multivariate central limit theorem established for $\widehat{f}_n^w$ allows us to easily derive a test for partial additivity. More precisely, let us consider general regression models of the form

$$Z_n = m(X_n, Y_n) + \varepsilon_n \tag{5.1}$$

where $Z \in \mathbb{R}$ is the explained variable, $X \in \mathbb{R}$ is a given explanatory variable and $Y \in \mathbb{R}^{d-1}$ is the vector of the other $(d-1)$ explanatory variables. We are interested in the subclass of nonlinear partially additive models of the form

$$Z_n = f(X_n) + g(Y_n) + \varepsilon_n \tag{5.2}$$

where function g is not supposed to be additive and separable.

To test this kind of partial additivity of model (5.1), we introduce the following hypotheses about the regression function m:

$$\mathcal{H}_0 \;:\; \ll m \text{ is partially additive of the form } m = f + g \gg$$
$$\mathcal{H}_1 \;:\; \ll m \text{ is any function of } \mathbb{R}^d \text{ to } \mathbb{R} \gg$$

The null hypothesis $\mathcal{H}_0$ means that the nonlinear effects of the given explanatory variable X on one hand and the vector of explanatory variables Y on the other hand, are additive and separable. This leads to the special form of the regression function m, ie. $m(x, y) = f(x) + g(y)$.

Such a testing problem is different from testing for the global additivity of model (5.1) which has been initially addressed by (Hastie and Tibshirani, 1990), and then by (Auestad and Tjøstheim, 1994; Barry, 1993; Chen et al., 1995; Eubank et al., 1995) and more recently by (Amato and Antoniadis, 2001; Camlong-Viot, 2001).

In this section, we derive the principle of the test statistic and we prove an asymptotic result for building the test. It is a first step towards an entirely applicable procedure which is out of the scope of the paper.

5.1. Principle

The test procedure is based on simple remarks about the property of partial estimators for additive models. Assume that $\mathcal{H}_0$ is true and let us consider two different partial estimators $\widehat{f}_n^{w_1}(x)$ and $\widehat{f}_n^{w_2}(x)$ given by (4.2), where w_1 and w_2 are two different weight functions integrating to 1. Since these two partial estimators converge to the same limit $f(x)$, but up to an additive constant, then the difference estimates a constant. Indeed, we have

$$\widehat{f}_n^{w_1}(x) \;-\; \widehat{f}_n^{w_2}(x) \;=\; \int (w_1(y) - w_2(y))\widehat{m}_n(x, y)\,\widehat{p}_n^{\,Y}(y)\,dy$$

and by Theorem 4.1, we have

$$\widehat{f}_n^{w_1}(x) \;-\; \widehat{f}_n^{w_2}(x) \;\xrightarrow[n\to\infty]{a.s.}\; C_{w_1,w_2}$$

where C_{w_1,w_2} is independent of x. Thus, for any weight function w integrating to 0, the estimator

$$S_n(x) \;=\; \int w(y)\widehat{m}_n(x, y)\,\widehat{p}_n^{\,Y}(y)\,dy \tag{5.3}$$

almost surely converges to $S(x)$ equal to

$$S(x) \;=\; \mathbb{E}\left[w(Y)\, m(x, Y)\right] \tag{5.4}$$

Under $\mathcal{H}_0$, $S(x)$ is equal to a constant which is independent of x. Therefore, for two distinct points x_1 and x_2, we have

$$S_n(x_1) - S_n(x_2) \xrightarrow[n\to\infty]{a.s.} S(x_1) - S(x_2) = 0 \tag{5.5}$$

Of course, this result does not hold under $\mathcal{H}_1$. This is the starting idea of our test procedure that we combine with the previous multivariate central limit theorem.

5.2. Testing for partial additivity

Let us consider $w : \mathbb{R}^{d-1} \to \mathbb{R}$ with compact support and such that $\mathbb{E}\,[w(Y)] = 0$. Let $x_1, \ldots, x_{2q}$ be $2q$ distinct design points of $\mathbb{R}$. Starting from (5.5), we introduce the following test statistic $T_q(n)$ which is a suitably normalized sum of squares of quantities estimating 0 under $\mathcal{H}_0$:

$$T_q(n) \;=\; nh_{1,n} \sum_{j=1}^{q} \frac{1}{\widehat{v}_n(x_j) \,+\, \widehat{v}_n(x_{j+q})} \,\Big(S_n(x_j) - S_n(x_{j+q})\Big)^2$$

where $\widehat{v}_n(x) \;=\; \sigma^2 \;\|K^{(1)}\|_2^2 \displaystyle\int \frac{w^2(y)\,(\widehat{p}_n^{\,Y}(y))^2}{\widehat{p}_n^{\,XY}(x,y)}\,dy.$
Then, for the test statistic $T_q(n)$, we have the following theorem.

THEOREM 5.1. *Assume that [M], [K] and [B1] to [B5] hold. Assume also that w is C^2-class. Then,*

1) under $\mathcal{H}_0$, $T_q(n) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(q)$.

2) under $\mathcal{H}_1$, $(n\,h_{1,n})^{-1}\,T_q(n) \xrightarrow[n\to\infty]{a.s.} \ell_q$ where constant $\ell_q > 0$ if there exists $j \in \{1, \ldots, q\}$ such that $\mathbb{E}\,[w(Y)\,(m(x_j, Y) - m(x_{j+q}, Y))] \neq 0$.

Proof The proof is given in Appendix C.

These asymptotic results make it possible to construct a test for partial additivity of m: Part 1 gives the null distribution and Part 2 guarantees that the asymptotic power of the test is equal to 1 since the test statistic almost surely explodes.

Appendix A.

This appendix presents some technical results about uniform almost sure convergence of the kernel-based estimators introduced in Paragraph 4.2.

LEMMA A.1. *Assume that [M] and [K] hold. Then, for any compact subset $\mathcal{D}$ of $\mathbb{R}^{d-1}$, we have*

$$\sup_{y\in\mathcal{D}}\left|\widehat{p}_n^{XY}(x,y)-p^{XY}(x,y)\right| \overset{a.s.}{=} O\Big(\frac{\log n}{\sqrt{n\,h_{1,n}h_{2,n}^{d-1}}}+h_{1,n}^{\delta}+h_{2,n}^{\delta}\Big) \tag{A.1}$$

$$\sup_{y\in\mathcal{D}}\left|\widehat{p}_n^{Y}(y)\,-\,p^{Y}(y)\right| \overset{a.s.}{=} O\Big(\log n/\sqrt{n\,h_{3,n}^{d-1}}+h_{3,n}^2\Big) \tag{A.2}$$

Besides, if $(\log n)^2 = o\left(n\,h_{1,n}h_{2,n}^{d-1}\right)$, *then*

$$\int_{\mathcal{D}}\widehat{p}_n^{Y}(y)\,dy \ \xrightarrow[n\to\infty]{a.s.}\ \mathbb{P}\left[Y\in\mathcal{D}\right] \tag{A.3}$$

$$\int_{\mathcal{D}}\widehat{p}_n^{XY}(x,y)\,dy \ \xrightarrow[n\to\infty]{a.s.}\ \int_{\mathcal{D}}p^{XY}(x,y)\,dy \tag{A.4}$$

and

$$\liminf_{n\to\infty}\inf_{y\in\mathcal{D}}\left|\widehat{p}_n^{XY}(x,y)\right| \ >\ 0,\ a.s. \tag{A.5}$$

LEMMA A.2. *Assume that [M] and [K] hold. Assume also that [B2] holds (or [B2bis] if [M1bis] is assumed instead of [M1]). Then, for any compact subset $\mathcal{D}$ of $\mathbb{R}^{d-1}$, we have*

$$\sup_{y\in\mathcal{D}}\left|\widehat{\ell}_n(x,y)-\ell(x,y)\right| \overset{a.s.}{=} O\Big(\frac{\log n}{\sqrt{n\,h_{1,n}h_{2,n}^{d-1}}}+h_{1,n}^{\delta}+h_{2,n}^{\delta}\Big) \tag{A.6}$$

and

$$\sup_{y\in\mathcal{D}}\left|\widehat{m}_n(x,y)-m(x,y)\right| \overset{a.s.}{=} O\Big(\frac{\log n}{\sqrt{n\,h_{1,n}h_{2,n}^{d-1}}}+h_{1,n}^{\delta}+h_{2,n}^{\delta}\Big) \tag{A.7}$$

The sequel of the appendix is concerned with the proofs of Lemma A.1 and Lemma A.2.

PROOF OF LEMMA A.1. As p^{XY} is C^{δ}-class with bounded derivatives of order δ, we obtain using a Taylor's expansion of order δ and assumptions on kernels $K^{(1)}$ and $K^{(2)}$ that

$$\sup_{y\in\mathbb{R}^{d-1}}\left|\mathbb{E}\,\widehat{p}_n^{XY}(x,y)\,-\,p^{XY}(x,y)\right| = O\Big(h_{1,n}^{\delta}+h_{2,n}^{\delta}\Big) \tag{A.8}$$

In the same manner, using a Taylor's expansion of order 2, we easily show that

$$\sup_{y\in\mathbb{R}^{d-1}}\left|\mathbb{E}\,\widehat{p}_n^{Y}(y)\,-\,p^{Y}(y)\right| \,=\, O\Big(h_{3,n}^2\Big) \tag{A.9}$$

In addition, results of (Liebscher, 1996) can be applied in our context and give

$$\sup_{y\in\mathcal{D}} \left|\widehat{p}_n^{XY}(x,y) - \mathbb{E}\,\widehat{p}_n^{XY}(x,y)\right| \overset{a.s.}{=} O\Big(\log n/\sqrt{n\,h_{1,n}h_{2,n}^{d-1}}\Big) \quad \text{(A.10)}$$

$$\sup_{y\in\mathcal{D}} \left|\widehat{p}_n^{Y}(y) - \mathbb{E}\,\widehat{p}_n^{Y}(y)\right| \overset{a.s.}{=} O\Big(\log n/\sqrt{n\,h_{3,n}^{d-1}}\Big) \quad \text{(A.11)}$$

Then, combining (A.8) with (A.10), and (A.9) with (A.11) give (A.1) and (A.2), respectively. Results (A.3) and (A.4) are directly deduced from (A.1) and (A.2), respectively. Finally, (A.5) comes from (A.1) and the following inequality:

$$\inf_{y\in\mathcal{D}} \left|\widehat{p}_n^{XY}(x,y)\right| > \inf_{y\in\mathcal{D}} p^{XY}(x,y) - \sup_{y\in\mathcal{D}} \left|\widehat{p}_n^{XY}(x,y) - p^{XY}(x,y)\right|$$

Proof of Lemma A.2. Firstly, let us recall that by Collomb's decomposition (Collomb, 1984),

$$\widehat{m}_n - m = \left(\widehat{p}_n^{XY}\right)^{-1} \Big[\Big(\widehat{\ell}_n - \mathbb{E}\,\widehat{\ell}_n\Big) - m\left(\widehat{p}_n^{XY} - \mathbb{E}\,\widehat{p}_n^{XY}\right) \quad \text{(A.12)}$$

$$+ \Big(\mathbb{E}\,\widehat{\ell}_n - m\,\mathbb{E}\,\widehat{p}_n^{XY}\Big)\Big] \quad \text{(A.13)}$$

As p^{XY} and m are C^{δ}-class with bounded derivatives of order δ, we obtain using a Taylor's expansion of order δ and assumptions on kernels $K^{(1)}$ and $K^{(2)}$ that

$$\sup_{y\in\mathbb{R}^{d-1}} \left|\mathbb{E}\,\widehat{\ell}_n(x,y) - m(x,y)\,\mathbb{E}\,\widehat{p}_n^{XY}(x,y)\right| = O\Big(h_{1,n}^{\delta} + h_{2,n}^{\delta}\Big) \quad \text{(A.14)}$$

$$\sup_{y\in\mathbb{R}^{d-1}} \left|\mathbb{E}\,\widehat{\ell}_n(x,y) - \ell(x,y)\right| = O\Big(h_{1,n}^{\delta} + h_{2,n}^{\delta}\Big) \quad \text{(A.15)}$$

Now, adapting the proof of Proposition 2 of (Ango Nze and Portier, 1994) to the multi-dimensional case, we derive that

$$\sup_{y\in\mathcal{D}} \left|\widehat{\ell}_n(x,y) - \mathbb{E}\,\widehat{\ell}_n(x,y)\right| \overset{a.s.}{=} O\Big(\log n/\sqrt{n\,h_{1,n}h_{2,n}^{d-1}}\Big). \quad \text{(A.16)}$$

This result holds under the following bandwidth condition

$$u_n^2\,(\log n)^3 = O\Big(n\,h_{1,n}h_{2,n}^{d-1}\log\Big((h_{1,n}h_{2,n}^{d-1})^{-1}u_n^2\Big)\Big) \quad \text{(A.17)}$$

with $u_n = \mathrm{cte}\,n^{2/\mu}$ (leading to bandwidth condition [B2]) if assumption [M1] holds, and with $u_n = \mathrm{cte}\,(\log n)^{1/b}$ (leading to bandwidth condition [B2bis]) if assumption [M1bis] holds, where "cte" denotes any strictly positive constant.

Finally, combining (A.15) with (A.16) gives (A.6), and (A.1), (A.5), (A.14) and (A.16) together with (A.13) gives (A.7).

Appendix B.

In this appendix, we establish an almost sure convergence result and a multivariate central limit theorem for the estimator $S_n(x)$ defined by

$$S_n(x) \;=\; \int w(y)\, \widehat{m}_n(x,y)\, \widehat{p}_n^{\,Y}(y)\, dy \tag{B.1}$$

where $w : \mathbb{R}^{d-1} \to \mathbb{R}$ may be any bounded compactly supported function. It estimates

$$S(x) \;=\; \mathbb{E}\left[w(Y)m(x,Y)\right] \tag{B.2}$$

If in addition w is such that $\mathbb{E}\left[w(Y)\right] = 1$ (this condition plays no role in the study of $S_n(x)$), then $S_n(x)$ and $S(x)$ reduce to the estimator $\widehat{f}_n^{\,w}(x)$ and its limit $f(x)+C_w$, respectively. Therefore, Theorem 4.1 and 4.2 are consequences of the following theorem.

THEOREM B.1. *Assume that [M] and [K] hold. Assume also that the bandwidth $h_{1,n}$, but also bandwidths $h_{2,n}$ and $h_{3,n}$ (used for $\widehat{m}_n$ and $\widehat{p}_n^{\,Y}$ respectively) satisfy [B1] and [B2] (or [B2bis] if [M1] is replaced by [M1bis]).*

1. *Then, we have* $S_n(x) \xrightarrow[n\to\infty]{a.s.} S(x)$.

2. *In addition, if w is C^2-class and if [B3], [B4] and [B5] hold, then*

$$G_n(x) = \sqrt{n\, h_{1,n}}\,\Big(S_n(x) \;-\; S(x)\Big) \;\xrightarrow[n\to\infty]{\mathcal{L}}\; \mathcal{N}\Big(0\,,\, v(x)\Big) = G(x)$$

where $v(x) = \sigma^2 \|K^{(1)}\|_2^2 \displaystyle\int \frac{(w(y)\, p^Y(y))^2}{p^{XY}(x,y)}\, dy$.

3. *Besides, for $x_1, \cdots, x_q$, q distinct points of $\mathbb{R}$,*

$$\Big(G_n(x_1), \cdots, G_n(x_q)\Big) \;\xrightarrow[n\to\infty]{\mathcal{L}}\; \Big(G(x_1), \cdots, G(x_q)\Big)$$

where $G(x_1), \cdots, G(x_q)$ are independent.

PROOF. To establish Part 1, let us rewrite $S_n(x) - S(x)$ under the form

$$\begin{aligned} S_n(x) \;-\; S(x) \;=\; & \int w(y)\,(\widehat{m}_n(x,y) - m(x,y))\, \widehat{p}_n^{\,Y}(y)\, dy \\ & + \int w(y)\, m(x,y)\, \Big(\widehat{p}_n^{\,Y}(y) - p^Y(y)\Big)\, dy \end{aligned}$$

As w is compactly supported, we deduce that $\int |w(y)\, m(x,y)|\, dy < \infty$ and $\int \left|w(y)\, \widehat{p}_n^Y(y)\right| dy < \infty$, and we infer from Lemmas A.1 and A.2 and the bandwidth conditions [B1] and [B2] that $S_n(x) \xrightarrow[n\to\infty]{a.s.} S(x)$.

Now, let us prove Part 2. Starting from the definition of $S_n(x)$, we can rewrite $S_n(x) - S(x)$ under the form

$$S_n(x) \;-\; S(x) \;=\; A_n(x) \;+\; B_n(x) \;+\; C_n(x) \;+\; D_n(x)$$

where

$$\begin{aligned}
A_n(x) \;&=\; \int w(y)\, m(x,y) \Big(\widehat{p}_n^Y(y) - p^Y(y)\Big)\, dy \\
B_n(x) \;&=\; \int \frac{w(y)}{p^{XY}(x,y)} \Big(\widehat{m}_n(x,y) - m(x,y)\Big) \\
&\qquad\qquad \Big(p^{XY}(x,y) - \widehat{p}_n^{XY}(x,y)\Big)\widehat{p}_n^Y(y)\, dy \\
C_n(x) \;&=\; \int \frac{w(y)}{p^{XY}(x,y)} \Big(\widehat{\ell}_n(x,y) - m(x,y)\, \widehat{p}_n^{XY}(x,y)\Big) \\
&\qquad\qquad \Big(\widehat{p}_n^Y(y) - p^Y(y)\Big)\, dy \\
D_n(x) \;&=\; \int \frac{w(y)}{p^{XY}(x,y)} \Big(\widehat{\ell}_n(x,y) - m(x,y)\, \widehat{p}_n^{XY}(x,y)\Big)\, p^Y(y)\, dy
\end{aligned}$$

Now, let us study the convergence of each term and show that,

$$\sqrt{n\, h_{1,n}} \Big(A_n(x) + B_n(x) + C_n(x)\Big) \;\xrightarrow[n\to\infty]{a.s.}\; 0 \tag{B.3}$$

$$\sqrt{n\, h_{1,n}}\, D_n(x) \;\xrightarrow[n\to\infty]{\mathcal{L}}\; \mathcal{N}\Big(0\,,\, v(x)\Big) \tag{B.4}$$

Convergence study of $A_n(x)$. Starting from the definition of $\widehat{p}_n^Y$, we can rewrite $A_n(x)$ under the form $A_{1,n}(x) + A_{2,n}(x)$ where

$$\begin{aligned}
A_{1,n}(x) \;&=\; \frac{1}{n}\sum_{k=1}^{n} \int \Big(w(Y_k + h_{3,n}u)\, m(x, Y_k + h_{3,n}u) \\
&\qquad\qquad\qquad -\; w(Y_k)\, m(x, Y_k)\Big)\, K^{(3)}(u)\, du \\
A_{2,n}(x) \;&=\; \frac{1}{n}\sum_{k=1}^{n} w(Y_k)\, m(x, Y_k) \;-\; \mathbb{E}\left[w(Y)\, m(x,Y)\right]
\end{aligned}$$

Using a Taylor's expansion of order 2 of $w(\,.\,)m(x,\,.\,)$ and the kernel assumption on $K^{(3)}$, we derive that

$$|A_{1,n}(x)| \;\stackrel{a.s.}{=}\; O\Big(h_{3,n}^2\Big) \tag{B.5}$$

In addition, since $\mathbb{E}\Big[\Big(w(Y)\,m(x,Y)\Big)^2\Big] < \infty$ and (Y_n) is β-mixing with exponential rate, Theorem 2 of (Rio, 1996) gives

$$|A_{2,n}(x)| \overset{a.s.}{=} O\left(\left(\frac{\text{loglog } n}{n}\right)^{1/2}\right) \tag{B.6}$$

Therefore, as $n\,h_{1,n}\,h_{3,n}^4 = o(1)$ and $h_{1,n}\,\text{loglog } n = o(1)$ by the bandwidth conditions [B1] and [B3], then (B.5) together with (B.6) imply that $\sqrt{n\,h_{1,n}}\,A_n(x) \xrightarrow[n\to\infty]{a.s.} 0$.

Convergence study of $B_n(x)$ *and* $C_n(x)$. Let us rewrite $C_n(x)$ under the form $C_{1,n}(x) + C_{2,n}(x)$ where

$$\begin{aligned} C_{1,n}(x) &= \int \frac{w(y)}{p^{XY}(x,y)} \Big(\widehat{\ell}_n(x,y) - \ell(x,y)\Big)\Big(\widehat{p}_n^{\,Y}(y) - p^Y(y)\Big)\,dy \\ C_{2,n}(x) &= \int \frac{w(y)\,m(x,y)}{p^{XY}(x,y)} \Big(p^{XY}(x,y) - \widehat{p}_n^{\,XY}(x,y)\Big) \\ &\qquad\qquad \times \Big(\widehat{p}_n^{\,Y}(y) - p^Y(y)\Big)\,dy \end{aligned}$$

As $\displaystyle\int \frac{|w(y)|}{p^{XY}(x,y)}\,dy$ and $\displaystyle\int \frac{|w(y)\,m(x,y)|}{p^{XY}(x,y)}\,dy$ are bounded, we easily deduce that

$$\begin{aligned} |B_n(x)| &\overset{a.s.}{=} O\Big(\sup_{y\in\mathcal{D}} |\widehat{m}_n(x,y) - m(x,y)| \sup_{y\in\mathcal{D}} \Big|p^{XY}(x,y) - \widehat{p}_n^{\,XY}(x,y)\Big|\Big) \\ |C_{1,n}(x)| &\overset{a.s.}{=} O\Big(\sup_{y\in\mathcal{D}} \Big|\widehat{\ell}_n(x,y) - \ell(x,y)\Big| \sup_{y\in\mathcal{D}} \Big|\widehat{p}_n^{\,Y}(y) - p^Y(y)\Big|\Big) \\ |C_{2,n}(x)| &\overset{a.s.}{=} O\Big(\sup_{y\in\mathcal{D}} \Big|p^{XY}(x,y) - \widehat{p}_n^{\,XY}(x,y)\Big| \sup_{y\in\mathcal{D}} \Big|\widehat{p}_n^{\,Y}(y) - p^Y(y)\Big|\Big) \end{aligned}$$

Finally, combining results of Lemma A.1 and Lemma A.2, we deduce that

$$\sqrt{n\,h_{1,n}}\,\Big(B_n(x) + C_n(x)\Big) \xrightarrow[n\to\infty]{a.s.} 0 \tag{B.7}$$

as soon as

$$\sqrt{n\,h_{1,n}}\,\Big(\frac{\log n}{\sqrt{n\,h_{1,n}h_{2,n}^{d-1}}} + h_{1,n}^{\delta} + h_{2,n}^{\delta}\Big)^2 = o(1) \tag{B.8}$$

$$\sqrt{n\,h_{1,n}}\,\Big(\frac{\log n}{\sqrt{n\,h_{1,n}h_{2,n}^{d-1}}} + h_{1,n}^{\delta} + h_{2,n}^{\delta}\Big)\Big(\frac{\log n}{\sqrt{n\,h_{3,n}^{d-1}}} + h_{3,n}^2\Big) = o(1) \tag{B.9}$$

From the bandwidth conditions [B1] and [B3], conditions (B.8) and (B.9) reduce to the bandwidth condition [B5].

Convergence study of $D_n(x)$. Let us rewrite $D_n(x)$ as $D_{1,n}(x) + D_{2,n}(x)$ where

$$\begin{aligned} D_{1,n}(x) &= \frac{1}{n}\sum_{k=1}^{n} U_{k,n}(x) \\ D_{2,n}(x) &= \frac{1}{n}\sum_{k=1}^{n} K^{(1)}_{h_{1,n}}(X_k - x)\,\varepsilon_k \int K^{(2)}_{h_{2,n}}(Y_k - y)\,H(x,y)\,dy \end{aligned}$$

with $H(x,y) = w(y)\,p^Y(y)\,/\,p^{XY}(x,y)$ and

$$\begin{aligned} U_{k,n}(x) &= K^{(1)}_{h_{1,n}}(X_k - x)\int K^{(2)}_{h_{2,n}}(Y_k - y) \\ &\qquad \times \Big(m(X_k, Y_k) - m(x,y)\Big) H(x,y)dy \end{aligned}$$

Let us remark that for any x, $\|H(x,\,.\,)\|_\infty$ is bounded since w is compactly supported. The asymptotic normality of $D_n(x)$ will be given by $D_{2,n}(x)$. First of all, let us show that $\sqrt{n\,h_{1,n}}\,D_{1,n}(x) \xrightarrow[n\to\infty]{P} 0$. Starting from

$$\begin{aligned} \mathbb{E}\left[U_{1,n}(x)\right] &= \iiint K^{(1)}(u)\,K^{(2)}(v) H(x,y)\,p^{XY}(x + u\,h_{1,n}, y + v\,h_{2,n}) \\ &\qquad \times \Big(m(x + u\,h_{1,n}, y + v\,h_{2,n}) - m(x,y)\Big)\,dy\,du\,dv \end{aligned}$$

and using a Taylor's expansion of order δ for m and p^{XY} as well as the kernel assumptions on $K^{(1)}$ and $K^{(2)}$, we obtain that

$$\left|\mathbb{E}\left[U_{1,n}(x)\right]\right| = O\Big(h_{1,n}^{\delta} + h_{2,n}^{\delta}\Big). \tag{B.10}$$

Furthermore,

$$\begin{aligned} n\,h_{1,n}\mathbb{V}\mathrm{ar}\Big(D_{1,n}(x)\Big) &= h_{1,n}\,\mathbb{V}\mathrm{ar}\Big(U_{1,n}(x)\Big) \\ &\qquad + \frac{2\,h_{1,n}}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{C}\mathrm{ov}\Big(U_{i,n}(x), U_{j,n}(x)\Big) \end{aligned}$$

Let us study the term of variance. From (B.10) and the bandwidth condition [B3], we deduce that

$$h_{1,n}\,\mathbb{V}\mathrm{ar}\Big(U_{1,n}(x)\Big) = h_{1,n}\,\mathbb{E}\left[U_{1,n}^2(x)\right] + o\,(1) \tag{B.11}$$

In addition, after change of variable, we obtain that

$$\begin{aligned}\mathbb{E}\left[U_{1,n}^2(x)\right] &= h_{1,n}^{-1} \iiiint \left(K^{(1)}(u)\right)^2 K^{(2)}(v)\, K^{(2)}(t)\, H(x,y) \\ &\quad H(x, y+h_{2,n}(t+v))p^{XY}(x+u\,h_{1,n}, y+v\,h_{2,n}) \\ &\quad \Big(m(x+u\,h_{1,n}, y+v\,h_{2,n}) - m(x, y+h_{2,n}(t+v))\Big) \\ &\quad \Big(m(x+u\,h_{1,n}, y+v\,h_{2,n}) - m(x,y)\Big)\, du\, dv\, dy\, dt\end{aligned}$$

and Lebesgue's theorem implies that $h_{1,n}\,\mathbb{E}\left[U_{1,n}^2(x)\right] \underset{n\to\infty}{\longrightarrow} 0$. Then, combining this result with (B.11) leads to

$$h_{1,n}\mathbb{V}\mathrm{ar}\Big(U_{1,n}(x)\Big) \underset{n\to\infty}{\longrightarrow} 0 \tag{B.12}$$

Now, applying a well-known bound for covariance of α-mixing random variables (Davydov, 1970), we obtain

$$\Big|\mathbb{C}\mathrm{ov}\Big(U_{i,n}(x), U_{j,n}(x)\Big)\Big| \leq 10\, \alpha_{j-i}^{1-2/a} \Big(\mathbb{E}\,|U_{i,n}(x)|^a\, \mathbb{E}\,|U_{j,n}(x)|^a\Big)^{1/a}$$

with $2 < a < \infty$. Thus, taking $a = 3$ and using the exponential rate of the mixing coefficients, we derive that

$$\left|\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{C}\mathrm{ov}\Big(U_{i,n}(x), U_{j,n}(x)\Big)\right| = O\Big(n \left(\mathbb{E}\,|U_{1,n}(x)|^3\right)^{2/3}\Big) \tag{B.13}$$

Let us give a bound for $\mathbb{E}\,|U_{1,n}(x)|^3$. Let us denote

$$\begin{aligned}L_n(u,v,x) &= \int K_{h_{2,n}}^{(2)}(v-y)\Big(m(u,v) - m(x,y)\Big)\, H(x,y)\, dy \\ &= \Big(m(u,v) - m(x,v)\Big) \int K^{(2)}(t)\, H(x, v+t\,h_{2,n})\, dt \\ &+ \int K^{(2)}(t)\Big(m(x,v) - m(x, v+t\,h_{2,n})\Big) H(x, v+t\,h_{2,n})\, dt\end{aligned}$$

Using once again Taylor's expansions and the kernel assumptions on $K^{(2)}$, we derive that

$$|L_n(u,v,x)| = O\Big(h_{2,n}^{\delta} + |m(u,v) - m(x,v)|\,(1 + h_{2,n}^2)\Big)$$

Then, as $U_{1,n}(x) = K_{h_{1,n}}^{(1)}(X_1 - x)\, L_n(X_1, Y_1, x)$, it follows easily that

$$\mathbb{E}\,|U_{1,n}(x)|^3 = O\Big(h_{1,n}^{-2} \int \left|K^{(1)}(t)\right|^3 \left(h_{1,n}^3\,|t|^3 + h_{2,n}^{3\delta}\right) p^X(x + h_{1,n}t)\, dt\Big)$$

Hence, we obtain $\mathbb{E}\,|U_{1,n}(x)|^3 = O\Big(h_{1,n} \,+\, h_{2,n}^{3\delta} h_{1,n}^{-2}\Big)$ and therefore

$$\frac{2\,h_{1,n}}{n}\left|\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{C}\mathrm{ov}\Big(U_{i,n}(x), U_{j,n}(x)\Big)\right| = O\Big(h_{1,n}^{5/3} \,+\, \frac{h_{2,n}^{2\delta}}{h_{1,n}^{1/3}}\Big) \quad \text{(B.14)}$$

From the bandwidth condition [B4], then $h_{2,n}^{2\delta} h_{1,n}^{-1/3} = o(1)$. Therefore, combining (B.12) and (B.14), we derive that

$$n\,h_{1,n}\,\mathbb{V}\mathrm{ar}\Big(D_{1,n}(x)\Big) \underset{n\to\infty}{\longrightarrow} 0 \quad \text{(B.15)}$$

which implies that $\sqrt{n\,h_{1,n}}\,D_{1,n}(x) \xrightarrow[n\to\infty]{P} 0$.

Convergence study of $D_{2,n}(x)$. Let us rewrite $\sqrt{n\,h_{1,n}}\,D_{2,n}(x)$ under the form $\sum_{k=1}^{n} \xi_{n,k}(x)$ with $\xi_{n,k}(x) = G_n(x, X_k, Y_k)\,\varepsilon_k$ and

$$G_n(x, X_k, Y_k) \;=\; \sqrt{\frac{h_{1,n}}{n}}\,K_{h_{1,n}}^{(1)}(X_k - x)\int K_{h_{2,n}}^{(2)}(Y_k - y)\,H(x,y)\,dy$$

Let us denote by $\mathcal{F}_k^n$ the σ-field generated by $(\varepsilon_j, X_j, Y_j)_{j=1,\dots,k}$. For any x, $\Big\{\sum_{j=1}^{k} \xi_{n,j}(x), \mathcal{F}_k^n, 1 \le k \le n,\ n \ge 1\Big\}$ is a zero-mean, square integrable martingale array. Then, to establish that

$$\sqrt{n\,h_{1,n}}\,D_{2,n}(x) \;=\; \sum_{k=1}^{n} \xi_{n,k}(x) \;\xrightarrow[n\to\infty]{\mathcal{L}}\; \mathcal{N}\Big(0\,,\,v(x)\Big) \quad \text{(B.16)}$$

we have only to check the assumptions of Corollary 3.1 of (Hall and Heyde, 1981, p.58), i.e.

$$\sum_{k=1}^{n}\mathbb{E}\Big[\xi_{n,k}^2(x)/\mathcal{F}_{k-1}^n\Big] \;\xrightarrow[n\to\infty]{P}\; v(x) \quad \text{(B.17)}$$

$$\forall\,\eta > 0,\ \rho_n(\eta) \;\xrightarrow[n\to\infty]{P}\; 0 \quad \text{(B.18)}$$

where $\rho_n(\eta) = \sum_{k=1}^{n}\mathbb{E}\left[|\xi_{n,k}(x)|^2\,\mathbf{1}_{\{|\xi_{n,k}(x)|\,\ge\,\eta\}}/\mathcal{F}_{k-1}^n\right]$.

To establish (B.17), let us rewrite $\displaystyle\sum_{k=1}^{n} \mathbb{E}\Big[\xi_{n,k}^2(x)/\mathcal{F}_{k-1}^n\Big]$ under the form $J_{1,n}(x) + J_{2,n}(x)$ where

$$
\begin{aligned}
J_{1,n}(x) &= \sum_{k=1}^{n} \mathbb{E}\Big[\xi_{n,k}^2(x)\Big] \\
&= \sigma^2 \iiiint \Big(K^{(1)}(u)\Big)^2 K^{(2)}(v)\, K^{(2)}(t)\, H(x, h_{2,n}(t+v)+y) \\
&\qquad \times\, H(x,y)\, p^{XY}(x+u\, h_{1,n}, y+v\, h_{2,n})\, du\, dv\, dy\, dt \\
J_{2,n}(x) &= \sum_{k=1}^{n} \Big(\mathbb{E}\Big[\xi_{n,k}^2(x)/\mathcal{F}_{k-1}^n\Big] - \mathbb{E}\Big[\xi_{n,k}^2(x)\Big]\Big) \\
&= \sigma^2 \sum_{k=1}^{n} \widetilde{G}_n(x, X_k, Y_k)
\end{aligned}
$$

with $\widetilde{G}_n(x, X_k, Y_k) = \mathbb{E}\Big[G_n^2(x, X_k, Y_k)/\mathcal{F}_{k-1}^n\Big] \;-\; \mathbb{E}\Big[G_n^2(x, X_k, Y_k)\Big]$.
By virtue of Lebesgue's theorem, we prove that

$$
J_{1,n}(x) \underset{n\to\infty}{\longrightarrow} \sigma^2 \|K^{(1)}\|_2^2 \int H^2(x,y)\, p^{XY}(x,y)\, dy \;=\; v(x) \tag{B.19}
$$

Let us now study the convergence of $J_{2,n}(x)$. We have

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\Big(J_{2,n}(x)\Big) &= \sigma^4 \sum_{k=1}^{n} \mathbb{V}\mathrm{ar}\Big(\widetilde{G}_n(x, X_k, Y_k)\Big) \\
&\quad +\; 2\,\sigma^4 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \mathbb{C}\mathrm{ov}\Big(\widetilde{G}_n(x, X_j, Y_j); \widetilde{G}_n(x, X_k, Y_k)\Big)
\end{aligned}
$$

Starting from the inequality

$$
\mathbb{V}\mathrm{ar}\Big(\widetilde{G}_n(x, X_k, Y_k)\Big) \le \mathbb{E}\Big[\Big(\mathbb{E}\Big[G_n^2(x, X_k, Y_k)/\mathcal{F}_{k-1}^n\Big]\Big)^2\Big]
$$

and using Jensen's inequality, we deduce that

$$
\mathbb{V}\mathrm{ar}\Big(\widetilde{G}_n(x, X_k, Y_k)\Big) \;\le\; \mathbb{E}\left[G_n^4(x, X_k, Y_k)\right] \tag{B.20}
$$

Now, let us remark that as w is compactly supported then $\|H(x,\,.\,)\|_\infty$ is bounded. Hence, we can find a finite constant C_x such that

$$
\left| \int K_{h_{2,n}}^{(2)}(Y_k - u)\, H(x,u)\, du \right| \;\le\; C_x \tag{B.21}
$$

and derive that

$$
\mathbb{E}\left[G_n^4(x, X_1, Y_1)\right] \;\le\; \frac{C_x^4\, h_{1,n}^2}{n^2}\, \mathbb{E}\left[\Big(K_{h_{1,n}}^{(1)}(X_1 - x)\Big)^4\right] \tag{B.22}
$$

Therefore, we infer from (B.20) together with (B.22) that

$$\mathbb{V}\mathrm{ar}\Big(\widetilde{G}_n(x, X_k, Y_k)\Big) \;=\; O\left((n^2\, h_{1,n})^{-1}\right) \tag{B.23}$$

The term of covariance remains to be studied. Applying the Davidov's inequality and using Jensen's inequality once again, we derive that

$$\begin{aligned}&\Big|\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} \mathbb{C}\mathrm{ov}\Big(\widetilde{G}_n(x, X_j, Y_j); \widetilde{G}_n(x, X_k, Y_k)\Big)\Big| \\ &\qquad\qquad = \; O\Big(n\Big(\mathbb{E}\left[G_n^6(x, X_1, Y_1)\right]\Big)^{2/3}\Big)\end{aligned} \tag{B.24}$$

By proceeding as for (B.22), we obtain that

$$\mathbb{E}\left[G_n^6(x, X_1, Y_1)\right] \;=\; O\left((n^3\, h_{1,n})^{-2}\right) \tag{B.25}$$

and together with (B.24), we deduce that

$$\Big|\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} \mathbb{C}\mathrm{ov}\Big(\widetilde{G}_n(x, X_j, Y_j); \widetilde{G}_n(x, X_k, Y_k)\Big)\Big| = O\left((n\, h_{1,n}^{4/3})^{-1}\right)$$

This result together with (B.23) ensures that $J_{2,n}(x) \xrightarrow[n\to\infty]{P} 0$ as soon as $n\, h_{1,n}^{4/3} \xrightarrow[n\to\infty]{} \infty$, which holds by the bandwidth condition [B4]. Finally, combining the convergence results of $J_{1,n}(x)$ and $J_{2,n}(x)$ gives (B.17). The Lindeberg's condition (B.18) remains to be proved. Let us denote $\Phi(t) = \mathbb{E}\left[\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1| \geq t\}}\right]$. For any $\eta > 0$, we have

$$\rho_n(\eta) \;\leq\; \Phi\left(\frac{\sqrt{n\, h_{1,n}}\,\eta}{\|K^{(1)}\|_\infty\, \|H(x,\,.\,)\|_\infty}\right)\, \sigma^{-2}\sum_{k=1}^{n}\mathbb{E}\Big[\xi_{n,k}^2(x)\,/\,\mathcal{F}_{k-1}^n\Big]$$

As ε has a finite moment of order > 2, then $\lim\limits_{t\to\infty}\Phi(t) = 0$. Therefore, combining this result with (B.17), we derive that for any $\eta > 0$, $\rho_n(\eta) \xrightarrow[n\to\infty]{P} 0$, and Lindeberg's condition is fulfilled, which achieves the proof of Theorem B.1, part 2.

Now, let us establish the joint asymptotic normality. Taking the previous results into account, it suffices to prove that for q distinct points of $\mathbb{R}$, denoted $x_1, \ldots, x_q$, the vector $\sqrt{n\, h_{1,n}}\Big(D_{2,n}(x_1), \ldots, D_{2,n}(x_q)\Big)$ converges in distribution to a centered Gaussian vector with independent components. We easily verify this by remarking that for $x \neq y$,

$$\sum_{k=1}^{n}\mathbb{E}\Big[\xi_{n,k}(x)\,\xi_{n,k}(y)/\mathcal{F}_{k-1}^n\Big] \;\xrightarrow[n\to\infty]{P}\; 0 \tag{B.26}$$

Indeed, we have

$$\begin{aligned}
&\mathbb{E}\Big|\sum_{k=1}^{n}\mathbb{E}\Big[\xi_{n,k}(x)\,\xi_{n,k}(y)/\mathcal{F}_{k-1}^{n}\Big]\Big| \\
&\quad\le\; C_x^2\,\sigma^2\,h_{1,n}\,\mathbb{E}\left[\Big|K^{(1)}_{h_{1,n}}(X_1-x)K^{(1)}_{h_{1,n}}(X_1-y)\Big|\right] \\
&\quad\le\; C_x^2\,\sigma^2\int\Big|K^{(1)}(z)\,K^{(1)}(z+(x-y)/h_{1,n})\Big|\;p^X(x+z\,h_{1,n})\,dz
\end{aligned}$$

and since $K^{(1)}$ is compactly supported, we deduce by the Lebesgue's theorem that

$$\mathbb{E}\Big|\sum_{k=1}^{n}\mathbb{E}\Big[\xi_{n,k}(x)\,\xi_{n,k}(y)/\mathcal{F}_{k-1}^{n}\Big]\Big| \;=\; o(1)$$

This last argument establishes (B.26) and closes the proof of Theorem B.1.

Appendix C.

This appendix is concerned with the proof of Theorem 5.1. From Theorem B.1 used with a compactly supported weight function w satisfying $\mathbb{E}\,[w(Y)]=0$, we easily derive that for two distinct points x and t,

$$\sqrt{n\,h_{1,n}}\,\Big(S_n(x)\;-\;S_n(t)\;+\;\mathbb{E}\Big[w(Y)(m(x,Y)-m(t,Y))\Big]\Big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(0\,,\,v(x)+v(t)\Big) \qquad \text{(C.1)}$$

Therefore, as soon as $\widehat{v}_n(y)\xrightarrow[n\to\infty]{a.s.} v(y)$ for any $y\in\mathbb{R}$, we infer from (C.1) that

$$\frac{n\,h_{1,n}}{\widehat{v}_n(x)+\widehat{v}_n(t)}\Big(S_n(x)\;-\;S_n(t)\;+\;\mathbb{E}\Big[w(Y)(m(x,Y)-m(t,Y))\Big]\Big)^2 \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(1)$$

and as $\mathbb{E}\Big[w(Y)\Big(m(x,Y)-m(t,Y)\Big)\Big]=0$ under $\mathcal{H}_0$, Part 1 is easily derived. Part 2 directly comes from Part 1 of Theorem B.1.

To close the proof, let us establish the convergence of $\widehat{v}_n(x)$ which reduces to show that

$$\int\frac{w^2(y)\,(\widehat{p}_n^{\,Y}(y))^2}{\widehat{p}_n^{\,XY}(x,y)}\,dy \xrightarrow[n\to\infty]{a.s.} \int\frac{w^2(y)\,(p^Y(y))^2}{p^{XY}(x,y)}\,dy \qquad \text{(C.2)}$$

Result (C.2) is easily obtained from Lemma A.1 together with:

$$
\Big|\int w^2(y)\Big(\frac{(\widehat{p}_n^{\,Y}(y))^2}{\widehat{p}_n^{\,XY}(x,y)} - \frac{(p^Y(y))^2}{p^{XY}(x,y)}\Big)\,dy\Big| \;=\;
O\left(\frac{\sup_{y\in\mathcal{D}}\left|\widehat{p}_n^{\,Y}(y) - p^Y(y)\right|}{\inf_{y\in\mathcal{D}}\left|\widehat{p}_n^{\,XY}(x,y)\right|}\right) + O\left(\frac{\sup_{y\in\mathcal{D}}\left|\widehat{p}_n^{\,XY}(x,y) - p^{XY}(x,y)\right|}{\inf_{y\in\mathcal{D}}\left|\widehat{p}_n^{\,XY}(x,y)\right|}\right).
$$

Appendix D.

This appendix presents an almost sure convergence result for a slightly modified version of the global estimator introduced in Paragraph 2.2. The proof is based on results of almost sure convergence over dilating sets for kernel estimators (Bosq, 1996).

Notations, definitions and assumptions used in this appendix can be found in the paragraphs 4.1 and 4.2.

THEOREM D.1. *Assume that [M1bis], [M2] to [M5] and [K] hold with $\delta = 2$. Let the bandwidths $h_{1,n}$, $h_{2,n}$ and $h_{3,n}$ be of the form $h_{1,n} = h_{2,n}$ with $h_{2,n} = \mathrm{cte}\,((\log n)^{2-1/b}/n)^{1/(d+4)}$, and $h_{3,n} = \mathrm{cte}\ (\log n/n)^{1/(d+3)}$, where* cte *denotes any strictly positive constant.*
Let $(c_n)_{n\geq 1}$ be a sequence of positive real numbers, increasing to ∞ and such that $c_n = O(n^\nu)$ with $\nu > 0$.
Then, if $p_n = \inf_{\|y\|\leq c_n} p^{XY}(x,y) > n^{-2/(d+4)}$, we have:

$$
\widehat{f}_{n,c_n}(x) \;=\; \int_{-c_n}^{c_n} \widehat{m}_n(x,y)\,\widehat{p}_n^{\,Y}(y)\,dy \;\xrightarrow[n\to\infty]{a.s.}\; f(x).
$$

PROOF. Let us rewrite $\widehat{f}_{n,c_n} - f(x)$ under the form

$$
\widehat{f}_{n,c_n}(x) - f(x) \;=\; A_n(x) + B_n(x) - \mathbb{E}\left[m(x,Y)\mathbf{1}_{\{\|Y\|\geq c_n\}}\right] \tag{D.1}
$$

where

$$
A_n(x) \;=\; \int_{-c_n}^{c_n} \Big(\widehat{m}_n(x,y) - m(x,y)\Big)\,\widehat{p}_n^{\,Y}(y)\,dy \tag{D.2}
$$

$$
B_n(x) \;=\; \int_{-c_n}^{c_n} m(x,y)\,\Big(\widehat{p}_n^{\,Y}(y) - p^Y(y)\Big)\,dy \tag{D.3}
$$

Since $\mathbb{E}\left[|Z|^2\right] < \infty$ and $c_n \to \infty$, then by a Markov's inequality

$$
\mathbb{E}\left[m(x,Y)\mathbf{1}_{\{\|Y\|\geq c_n\}}\right] \;\underset{n\to\infty}{\longrightarrow}\; 0 \tag{D.4}
$$

Now, using Theorem 2.2 of Bosq (1996, p. 49), we obtain that

$$\sup_{\|y\|\le c_n} \left|\widehat{p}_n^{\,Y}(y) - p^Y(y)\right| \overset{a.s.}{=} O\Big(\log n\Big((\log n)/n\Big)^{2/(d+3)}\Big) \tag{D.5}$$

It follows that

$$|B_n(x)| \xrightarrow[n\to\infty]{a.s.} 0 \tag{D.6}$$

as soon as $\log n\Big((\log n)/n\Big)^{2/(d+3)} \times \int_{-c_n}^{c_n} |m(x,y)|\, dy \underset{n\to\infty}{\longrightarrow} 0$. This condition is obviously fulfilled since $\int_{-c_n}^{c_n} |m(x,y)|\, dy \le p_n^{-1}\,\mathbb{E}\,[|Z|]$ and $p_n > n^{-2/(d+4)}$. In addition, using Theorem 3.3 of Bosq (1996, p. 74), we have

$$\sup_{\|y\|\le c_n} |\widehat{m}_n(x,y) - m(x,y)| \overset{a.s.}{=} o(1) \tag{D.7}$$

leading to

$$|A_n(x)| \xrightarrow[n\to\infty]{a.s.} 0 \tag{D.8}$$

Finally, combining (D.4), (D.6) and (D.8) together with (D.1) closes the proof of Theorem D.1.

REMARK D.1. When the explanatory variables are Gaussian, we can easily specify the sequence (c_n). Indeed, if $(X,Y) \sim \mathcal{N}(0\,,\Gamma)$ with Γ invertible, we have

$$\inf_{\|y\|\le c_n} p^{XY}(x,y) \;\ge\; \text{cte}\,\exp\Big(-\frac{c_n^2}{2\,\lambda_{\min}(\Gamma)}\Big)$$

Therefore, for $c_n = A(\log\log n)^{1/2}$ with $A < \infty$, or $c_n = A(\log n)^{1/2}$ with A such that $A^2 < 4\,\lambda_{\min}(\Gamma)\,/\,(d+4)$, we have $\widehat{f}_{n,c_n}(x) \xrightarrow[n\to\infty]{a.s.} f(x)$.

Appendix E.

This appendix is concerned with the performance of the global estimator defined by (2.3) and the recombined estimator defined by (2.11) (see Table VI and Table VII respectively). The models considered in both cases are defined in Paragraph 2.1.2.

Table VI. Global estimator performance

Model	n	SZ	Sf^1	Sf^2	Sf^3
$(\mathrm{M1})_{0.2}$	200	0.2529	0.0412	0.0066	
	800	0.2465	0.0073	0.0026	
	2000	0.2506	0.0036	0.0015	
$(\mathrm{M1})_{0.8}$	200	0.4436	0.1372	0.0569	
	800	0.3369	0.0532	0.0298	
	2000	0.3325	0.0401	0.0266	
$(\mathrm{M2})_{0.2}$	200	0.3270	0.0577	0.0115	0.0274
	800	0.2855	0.0128	0.0059	0.0123
	2000	0.2716	0.0078	0.0037	0.0077
$(\mathrm{M2})_{0.8}$	200	1.0174	0.2627	0.0820	0.1075
	800	0.7373	0.1444	0.0491	0.0677
	2000	0.5950	0.0902	0.0425	0.0516

To evaluate estimation and prediction errors, we compute two quadratic criteria. The first one is the estimation error criterion calculated for each f^i on $[-2\,,\,2]$, and defined by:

$$Sf^i = \frac{1}{10}\sum_{k=1}^{10}\frac{1}{\left|J^{i(k)}\right|}\sum_{j\in J^{i(k)}}\left(\widehat{f}_n^{i(k)}(X_j^{i(k)}) - f^i(X_j^{i(k)})\right)^2$$

where $J^{i(k)} = \{j \in \mathbb{N}\,, 1 \le j \le n \,/\, X_j^{i(k)} \in [-2\,,\,2]\}$. Quantities Sf^i allow to appreciate the quality of the estimation of f^i over $[-2\,,\,2]$: the smaller Sf^i, the better is the estimation of f^i.
The second one is the prediction error criterion, defined by:

$$SZ = \frac{1}{10}\sum_{k=1}^{10}\frac{1}{\left|\cap_i J^{i(k)}\right|}\sum_{j\in\cap_i J^{i(k)}}\left(\widehat{Z}_j^{(k)} - Z_j^{(k)}\right)^2$$

where $\widehat{Z}_j^{(k)}$ is the prediction of $Z_j^{(k)}$. It estimates the noise variance $\mathbb{V}\mathrm{ar}(\varepsilon)$ and quantifies the quality of the prediction of Z: the closer to $\mathbb{V}\mathrm{ar}(\varepsilon) = 0.5^2$, the better is the prediction.

References

Amato, U. and Antoniadis, A. (2001). Adaptive wavelet series estimation in separable nonparametric regression models. *To appear in Statistics and Computing.*

Table VII. Recombined estimator performance

Model	n	SZ	Sf^1	Sf^2	Sf^3
$(M1)_{0.2}$	200	0.2460	0.0375	0.0066	
	800	0.2452	0.0067	0.0025	
	2000	0.2498	0.0032	0.0013	
$(M1)_{0.8}$	200	0.2594	0.0410	0.0375	
	800	0.2540	0.0150	0.0162	
	2000	0.2562	0.0067	0.0086	
$(M2)_{0.2}$	200	0.2962	0.0449	0.0100	0.0234
	800	0.2760	0.0100	0.0051	0.0107
	2000	0.2653	0.0058	0.0031	0.0068
$(M2)_{0.8}$	200	0.3229	0.0449	0.0505	0.0273
	800	0.3014	0.0109	0.0236	0.0158
	2000	0.2824	0.0067	0.0151	0.0096

Ango Nze, P. and Portier, B. (1994). Estimation of the density and the regression functions of an absolutely regular stationary process. *Pub. Ins. Stat. Univ. Paris*, **38**, 2, 59-87.

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

Barry, D. (1993). Testing for additivity of a regression function. *Ann. of Stat.*, **21**, 235-254.

Bel, L., Bonneau, V., Chèze, N., Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.-M. and Portier, B. (1999). Etudes de pollution : modèles de prévision à 3 échéances pour le dioxyde d'azote et l'ozone. *Technical report for Airparif*, 215 pages, July 1999.

Bel, L., Bellanger, L., Bonneau, V., Ciuperca, G., Dacunha-Castelle, D., Deniau, C., Ghattas, B., Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.-M. and Tomassone, R. (1999). Eléments de comparaison de prévisions statistiques des pics d'ozone. *Revue de Stat. Appl.*, vol. XLVII (3), 7-25.

Bel, L., Oppenheim, G., Poggi, J.-M., Portier B. and Vautard, R. (2001). Statistical forecasting of ozone peaks over Paris area with a nonlinear additive model. Impact of imported ozone. *Proc. of the* 10^{th} *Int. Symp. on Applied Stoch. Models and Data Analysis*, June 12-15, 2001, Compiègne, 166-171.

Bosq, D. (1996). *Nonparametric statistics for stochastic processes, estimation and prediction*, Lecture Notes in Statistics, 110, Springer Verlag.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.*, **80**, 580-619.

Buja A., Hastie T. J. and Tibshirani, R. J. (1989). Linear smoothers and additive models. *Ann. of Stat.*, **17**, 453-510.

Cai, Z. and Masry, E. (1997). Nonparametric estimation of additive nonlinear ARX time series:local linear fitting and projections. *Preprint*, 1-30.

Camlong, C. (1999). Convergence presque sûre de l'estimateur à noyau d'une fonction de régression additive sous une hypothèse de mélange. *C. R Acad. Sci. Paris*, **t 329**, Ser. I, 75-78.

Camlong-Viot, C., Sarda, P. and Vieu, P. (2000). Additive time series: the kernel integration method. *Math. Meth. Stat*, **9**, 358-375.

Camlong-Viot, C. (2001). Vers un test d'additivité en régression non-paramétrique sous des conditions de mélange. *C. R Acad. Sci. Paris*, **t 333**, Ser. I, 877-880.

Chen, R., Lui, J. S. and Tsay, R. S. (1995). Additivity tests for nonlinear autoregression. *Biometrika*, **82**, 369-383.

Collomb, G. (1984). Propriétés de convergence presque complète du prédicteur à noyau. *Zeit. Wahrsch. verw. Gebiete*, **66**, 441-460.

Davydov, Y.A. (1970). The invariance principle for stationary processes. *Theo. Prob. Appl.*, **15**, 487-498.

Doukhan, P. (1994). *Mixing. Properties and examples.* Lecture Notes in Statistics, **72**, Springer Verlag.

Eubank, R. L., Hart, J. D., Simpson, D. G. and Stefanski, L. A. (1995). Testing for additivity in nonparametric regression. *Ann. of Stat.*, **23**(6), 1896-1920.

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. of Stat.*, **26**(3), 943-971.

Hall, P. and Heyde, C. C. (1981). *Martingale limit theory and its applications.* Academic Press.

Härdle, W. and Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, **47**, 43-57.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stoch. Proc. Appl.*, **65**, 69-80.

Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93-100.

Linton, O. B. and Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika*, **83**, 529-540.

Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, **84**, 469-473.

Mammen, E., Linton, O. B. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. of Stat.*, **27**, 1443-1490.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric theory*, **13**, 214-252.

Newey, W. K. (1994). Kernel estimation of partial means. *Econometric Theory*, **10**, 233-253.

Nielsen, J. P. and Linton, O. B. (1998). An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *J. R. Statist. Soc. B*, **60**, 217-222.

Opsomer, J. D. and Ruppert, D. (1997). On the existence and asymptotic properties of back-fitting estimators. *Ann. of Stat.*, **25**, 186-211.

Rio, E. (1996). The functional law of the iterated logarithm for stationary strongly mixing sequences. *Ann. of Prob.*, **23**, 1188-1203.

Sperlich, S., Linton, O.B. and Härdle, W. (1999). Integration and backfitting methods in additive models - finite sample properties and comparison. *Test.*, **8**, 419-458.

Stone, C. J. (1985). Additive regresssion and other nonparametric models. *Ann. of Stat.*, **13**, 698-705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. of Stat.*, **14**, 592-606.

Venables, W. N. and Ripley, B. (1994). *Modern Applied statistics with S-Plus*, Springer Verlag, New-York.

Wilks, D. R. (1995). *Statistical methods in the atmospheric sciences: an introduction*, Academic Press.

Address for Correspondence: Bruno Portier
Laboratoire de Mathématique – U.M.R. C 8628,
"Probabilités, Statistique et Modélisation"
Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France