Partial Estimators for Additive Models: Motivation, Illustration and Case Study

Nathalie Chèze

Université Paris X, Modal'X, 200 avenue de la République, 92000 Nanterre, France

Jean-Michel Poggi and Bruno Portier†

Laboratoire de Mathématique – U.M.R. C 8628, "Probabilités, Statistique et Modélisation"
Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France
and IUT Paris, Université René Descartes, 143 av. de Versailles, 75016 Paris, France

Summary. Starting from a variant of a marginal integration estimator, this paper proposes partial estimators for nonlinear additive regression models. These partial estimators are used for data analysis purposes while suitably chosen recombination of such estimators leads to improve the estimation and prediction performance for small to moderate sample size. Some simulations illustrate step-by-step the principle of the proposed estimators, which are applied to the analysis and the prediction of ozone concentration in Paris area.

Keywords: Additive models; Kernel estimation; Marginal integration.

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

1. Introduction

This paper is the first one of two companion papers dealing with partial estimators for nonlinear additive models. In short, this paper introduces the partial estimators with motivation, illustration and case study focusing on practical issues for analysis and prediction of time series using nonlinear additive models. The second one selects some related topics and deals with theoretical results, as a multivariate central limit theorem and an asymptotic test for partial additivity. These two companion papers can be independently read but are of course strongly connected, allowing the reader to focus on either practical issues or theoretical questions, as well as finite sample behaviour versus asymptotics.

Let us consider the nonlinear additive regression model

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

where Z is a real-valued dependent variable, $X^1, \ldots, X^d$ are the explanatory variables, μ is a constant (set to zero in the sequel without loss of generality) and ε is an unobservable noise.

†*Address for correspondence:* Laboratoire de Mathématique – U.M.R. C 8628, "Probabilités, Statistique et Modélisation" Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France
E-mail: bruno.portier@math.u-psud.fr

Such nonlinear additive regression models have been introduced in statistics and have been intensively used following the works of Breiman and Friedman (1985), Buja et al. (1989) and Hastie and Tibshirani (1990). Such models are particularly attractive for the following reasons. On one hand, with respect to the classical linear model $Z_n = \mu + \sum_{i=1}^{d} \theta^i X_n^i + \varepsilon_n$, 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 $Z_n = f(X_n^1, \ldots, X_n^d) + \varepsilon_n$, the separable model is more explicit and can be estimated without suffering from the so-called curse of dimensionality (Stone, 1985, 1986). Functions f^i are often estimated using the back-fitting algorithm. Hastie and Tibshirani (1990) have illustrated the efficiency of this procedure for many practical examples. This algorithm is widely used and is implemented in various scientific softwares (see for example Venables and Ripley (1994)). 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 to back-fitting, based on marginal integration, has been independently introduced by Newey (1994), Auestad and Tjostheim (1994) and 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 m (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 denote $m(x) = \sum_{i=1}^{d} f^i(x^i)$ where $x = (x^1, \ldots, x^d)$. Under the usual identifiability constraints $\mathbb{E}[f^j(X_n^j)] = 0$ for $j = 1, \ldots, d$, we have

$$f^i(x^i) = \mathbb{E}\Big[m\Big(X_n^1, \ldots, X_n^{i-1}, x^i, X_n^{i+1}, \ldots, X_n^d\Big)\Big] \tag{2}$$

$$= \int_{\mathbb{R}^{d-1}} m(x)\, p^{\underline{i}}(x^{\underline{i}})\, dx^{\underline{i}} \tag{3}$$

where $p^{\underline{i}}$ is the density of the vector $(X_n^1, \ldots, X_n^{i-1}, X_n^{i+1}, \ldots, X_n^d)$ and $x^{\underline{i}}$ denotes the vector $(x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^d)$. To estimate each function f^i, essentially two ways can be taken. The first one is to use a law of large numbers to estimate the expectation in (2) and replace m by an estimator $\widehat{m}$. Taking this way, Linton and Härdle (1996) and Linton (1997) study asymptotic properties of associated estimators. Another possibility used here, is to consider the integral in (3) and replace m and $p^{\underline{i}}$ by their respective estimators leading to an estimator that we call *global estimator* in this paper.

Of course, many variant of these estimators can be found in the literature. In short, $\widehat{m}$ can be multiplied by a compactly supported weight function or the unknown probability density function $p^{\underline{i}}$ can be replaced by a given arbitrary one. See Auestad and Tjostheim (1994), Masry and Tjostheim (1997), Cai and Masry (1997) and Fan et al. (1998) for the first approach, and Linton and Nielsen (1995), Camlong (1999) and Camlong-Viot et al. (2000) for the second one. A more detailed discussion can be found in the introduction of the companion paper dedicated to asymptotic properties.

The scope of this paper is to define, what we call *partial estimators* (in opposition to global estimator, previously introduced). To do that, let us remark that since $m(x)$ is additive and separable, the knowledge of $m(x)$ for a given $x^{\underline{i}}$, leads to $f^i(x^i)$ up to an additive constant, independent of $x^{\underline{i}}$. In addition, by integrating $m(x)$ over a domain $\mathcal{D}$ of $\mathbb{R}^{d-1}$ leads also to $f^i(x^i)$, but up to an additive constant, independent of x^i. Therefore, by integrating $\widehat{m}(x)$ over $\mathcal{D}$, we can define a partial estimator of f^i and by partitioning $\mathbb{R}^{d-1}$, we can build a family of partial estimators of f^i. For sure, this kind of estimators has no asymptotic value, but two aspects are interesting. Firstly, the analysis of the model additivity can be based on partial estimators used as data analysis tools. Secondly, the recomposition 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 m, leads to define *recombined estimators*. This 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 paper is organized as follows. After this introduction, section 2 deals with the global estimator. Section 3 introduces the partial estimators while section 4 defines the recombined estimators. Some simulations illustrate step-by-step the principle of the proposed estimators. Section 5 is dedicated to a real world problem: the analysis and prediction of ozone concentration in Paris area. The partial estimators are used for data analysis purposes whereas the recombined estimators are used to build the prediction model.

2. Global estimator based on marginal integration

2.1. Model

Let us consider the following additive model

$$Z_n \;=\; \sum_{i=1}^{d} f^i(X_n^i) \;+\; \varepsilon_n \;,\; n \in \mathbb{N} \tag{4}$$

where each function $f^i : \mathbb{R} \to \mathbb{R}$ satisfies the identifiability hypothesis $\mathbb{E}\left[f^i(X_n^i)\right] = 0$, $(X_n^1, \cdots, X_n^d)_{n\geq 1}$ is a strictly stationary sequence of random vectors and $(\varepsilon_n)_{n\geq 1}$ is an unobservable noise. This paper is devoted to the estimation of functions f^i. Since each function f^i plays the same role, without loss of generality, we can rewrite model (4) as

$$Z_n \;=\; f(X_n) \;+\; g(Y_n) \;+\; \varepsilon_n \;=\; m(X_n, Y_n) \;+\; \varepsilon_n \tag{5}$$

where f is one of the f^i, g the sum of the others, (X_n) one of the (X_n^i) and (Y_n) the vector of the others. Functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^{d-1} \to \mathbb{R}$ also satisfy the identifiability hypotheses, i.e. $\mathbb{E}[f(X_n)] = 0$ and $\mathbb{E}[g(Y_n)] = 0$.

To make the paper more readable, general material is formulated using form (5) and we use form (4) when actual models are concerned.

2.2. Global estimator

Let us recall the principle leading to the global estimator based on marginal integration. Let us denote by p^{XY}, p^X and p^Y the probability density function of (X, Y), X and Y, respectively. Integrating function m with respect to p^Y leads to

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

Table 1. Models

Model	$f^1(x)$	$f^2(x)$	$f^3(x)$	d
$(\mathrm{M1})_\rho$	x^2-1	$x/2$	—	2
$(\mathrm{M2})_\rho$	x^2-1	$x/2$	$\lvert x\rvert^{1/2} - E(\lvert X_1^3\rvert^{1/2})$	3

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 easily derive an estimator of $f(x)$ given by

$$\widehat{f}_n(x) \quad = \quad \int \widehat{m}_n(x,y)\,\widehat{p}_n^Y(y)\,dy \tag{6}$$

REMARK. This global estimator converges almost surely to $f(x)$, as proved in Cheze et al. (2000). The proof is based on results of almost sure convergence over dilating sets for kernel estimators (Bosq, 1996).

Using simulations, let us illustrate some well-known serious difficulties arising when such a global estimator is used.

2.3. Simulation based illustration

In this paper, simulations will illustrate step-by-step the behaviour of the different proposed estimators. The material introduced here will also be useful for the two next sections. 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)). We consider additive models of the form $Z_n = \sum_{i=1}^{d} f^i(X_n^i) + \varepsilon_n$ where d is equal to 2 or 3 (well suited for illustration purposes since the difficulties arising for large d using integration method are well known), $\varepsilon_n \sim \mathcal{N}\left(0\,,\,0.5^2\right)$ and $(X_n^1,\ldots,X_n^d) \sim \mathcal{N}\left(0\,,\,\Sigma_d\right)$ with $\Sigma_d(j,j)=1$ and $\Sigma_d(j,k)=\rho$ for $j\neq k$.
We consider two models denoted by $(\mathrm{M1})_\rho$ and $(\mathrm{M2})_\rho$ and defined in Table 1. These models are examined with $\rho=0.2$ and $\rho=0.8$, for moderate sample sizes $n=200$, 800 and 2000. The different functions f^i have been previously examined by Linton and Härdle (1996).

For each model, we simulate 10 realizations of $(X_j^1,\ldots,X_j^d,Z_j)_{1\le j\le n}$ and then obtain 10 realizations of $(\widehat{f}_n^i)_{1\le i\le d}$, denoted by $(\widehat{f}_n^{i(k)})_{1\le i\le d}$, for $1\le k\le 10$.
The kernels are direct products of Gaussian densities. 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. The integrals are approximated on a grid of 61 points within the interval $[-3\,,\,3]$, using a simple rectangle method.

To visualize bias and variance of the estimators, Fig. 1 and Fig. 2 present four curves for each function f^i: in dotted 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 $\widehat{f}_n^{i(k)}$, for $1\le k\le 10$.

Let us examine in Fig. 1 and Fig. 2 two extremes situations: a model with two weakly dependent explanatory variables, for a small sample size (Model $(\mathrm{M1})_{0.2}$ with $n=200$) and a model with three strongly dependent explanatory variables, for a large sample size (Model $(\mathrm{M2})_{0.8}$ with $n=2000$).

For Model $(\mathrm{M2})_{0.8}$ the results are of bad quality even if n is quite large (for $n=2000$, see Fig. 1), especially outside the interval $[-2\,,\,2]$. For Model $(\mathrm{M1})_{0.2}$, the estimation is

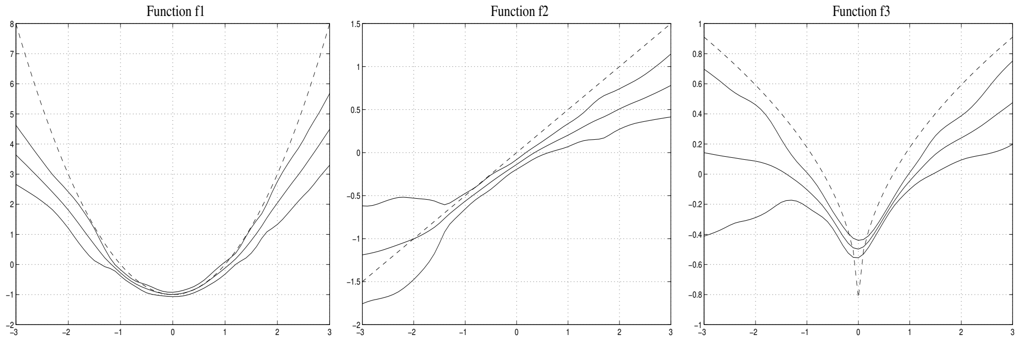

Fig. 1. Global estimators for Model $(\mathrm{M2})_{0.8}$ with $n = 2000$.

quite satisfactory as soon as $n = 200$ (see Fig. 2). 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 performance depend heavily on the dependence of the explanatory variables and the number of them (see Nielsen and Linton (1998), p. 221).

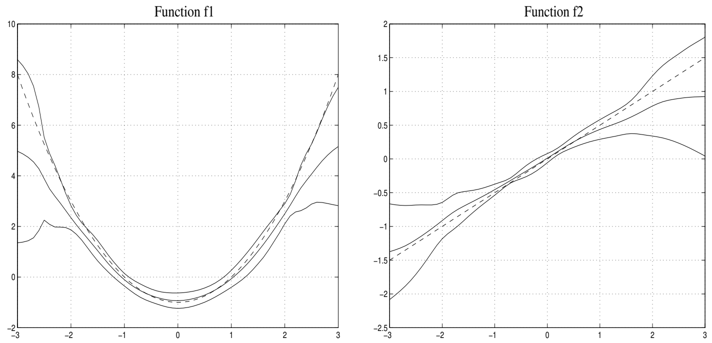

Fig. 2. Global estimators for Model $(\mathrm{M1})_{0.2}$ with $n = 200$.

In addition to evaluate estimation and prediction errors, we compute two quadratic criteria. The estimation error criterion for f^i is restricted to $[-2\,,\,2]$ and is 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 $J^{i(k)} = \{j \in \mathbb{N}\,, 1 \leq j \leq n \,/\, X_j^{i(k)} \in [-2\,,\,2]\}$. 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. The prediction error criterion is defined by $SZ = \dfrac{1}{10}\displaystyle\sum_{k=1}^{10}\frac{1}{\left|\cap_i J^{i(k)}\right|}\sum_{j\in\cap_i J^{i(k)}}\left(\widehat{Z}_j^{(k)} - Z_j^{(k)}\right)^2$ where $\widehat{Z}_j^{(k)}$ is the prediction of $Z_j^{(k)}$; SZ is an estimate of $\mathbb{E}\left(\varepsilon^2\right)$ and quantifies the quality of the prediction of Z. The closer to $\mathbb{E}\left(\varepsilon^2\right) = 0.5^2$, the better is the prediction.

From Table 7 of Appendix, 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 $(\mathrm{M1})_{0.8}$ and $(\mathrm{M2})_{0.8}$ are clearly less satisfactory (for the same sample size) than the results for models $(\mathrm{M1})_{0.2}$ and $(\mathrm{M2})_{0.2}$ 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.

3. Partial estimators

3.1. Principle

Let us present the idea leading to 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 $\mathbb{P}[Y \in \mathcal{D}]$, leads to $f(x)$ up to an additive constant :

$$\int_{\mathcal{D}} m(x,y)\, p^Y(y)\, dy \,/\, \mathbb{P}[Y \in \mathcal{D}] \;=\; f(x) \;+\; C(Y,g,\mathcal{D}) \tag{7}$$

Since the constant does not depend on x, it follows that for any given real x_0,

$$C(Y,g,\mathcal{D}) \;=\; \Big(\int_{\mathcal{D}} m(x_0,y)\, p^Y(y)\, dy\Big)\Big/\Big(\int_{\mathcal{D}} p^Y(y)\, dy\Big) \;-\; f(x_0) \tag{8}$$

So, we can use this relation combined with (7) to derive an "unbiased" partial estimator of function f, given by

$$\widehat{f}_n^{\mathcal{D}}(x) \;=\; \Big(\int_{\mathcal{D}} \Big(\widehat{m}_n(x,y) - \widehat{m}_n(x_0,y)\Big)\, \widehat{p}_n^Y(y)\, dy\Big)\Big/\Big(\int_{\mathcal{D}} \widehat{p}_n^Y(y)\, dy\Big) \;+\; \widehat{f}_n(x_0) \tag{9}$$

where $\widehat{f}_n(x_0) \;=\; \displaystyle\int \widehat{m}_n(x_0,y)\, \widehat{p}_n^Y(y)\, dy\,.$

Here we prefer to define, by partitioning $\mathbb{R}^{d-1}$, not only one but a family of partial estimators of f, in order to recombine them from a non-asymptotic perspective (see next section). Let us be a little bit more precise. Let $\mathcal{D}_1, \dots, \mathcal{D}_q$ be compact sets partitioning $\mathbb{R}^{d-1}$ such that for any j, $\mathbb{P}[Y \in \mathcal{D}_j] \neq 0$. Then, integrating over each $\mathcal{D}_j$ instead of $\mathcal{D}$, we can build q partial estimators of $f(x)$, denoted by $\widehat{f}_n^{\mathcal{D}_j}(x)$ and defined by (9), replacing $\mathcal{D}$ by $\mathcal{D}_j$. The terminology of partial estimators is justified by the reconstruction property :

$$\sum_{j=1}^{q} \Big(\int_{\mathcal{D}_j} \widehat{p}_n^Y(y)\, dy\Big)\, \widehat{f}_n^{\mathcal{D}_j}(x) \;=\; \widehat{f}_n(x) \tag{10}$$

Remark 1. Another way to obtain "unbiased" partial estimators is to estimate the left hand side of (7) for each point of a grid and then to estimate the integral leading 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.

Remark 2. In the companion paper (see paragraph 2.3), the asymptotic distribution of the following estimator

$$\widehat{f}_n^{w}(x) \;=\; \int w(y)\, \widehat{m}_n(x,y)\, \widehat{p}_n^Y(y)\, dy \tag{11}$$

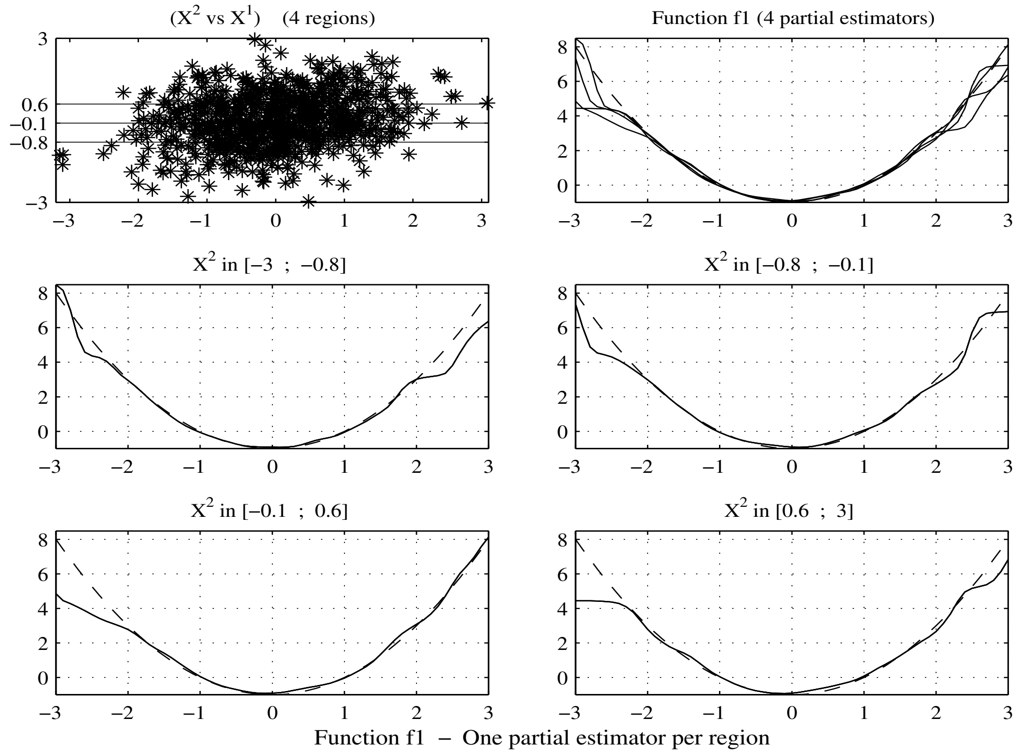

Fig. 3. Partial estimators of f^1 for model (M1)$_{0.2}$ with $n = 800$.

is given where w is a compactly supported weight function satisfying $\mathbb{E}[w(Y)] = 1$. This estimator is directly connected to the partial estimator considered previously, by choosing w as the characteristic function of domain $\mathcal{D}$, well normalized, ie. $w(y) = \mathbf{1}_{\mathcal{D}}(y) \,/\, \mathbb{P}\,[Y \in \mathcal{D}]$.

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 $(\mathcal{D}_j)_{1 \le j \le 4}$ 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 (8) 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 the next section, dedicated to air pollution analysis. In addition, partial estimators could also be used to build a test for partial additivity (see the companion paper Cheze et al. (2001)).

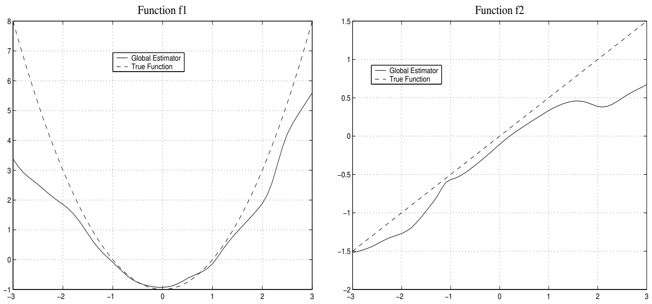

Fig. 4. Global estimators for model (M1)$_{0.8}$ with $n = 800$.

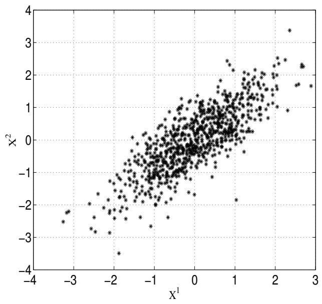

Fig. 5. Model (M1)$_{0.8}$ with $n = 800$. Explanatory variables X^2 versus X^1.

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*.

4. Recombined estimators

4.1. Principle

We present a simple situation from which the inspection of the partial estimators suggests a new estimator, called recombined estimator, and 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 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 $[-3\,,\,-0.6]$ 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 $[-2\,,\,0.4]$.

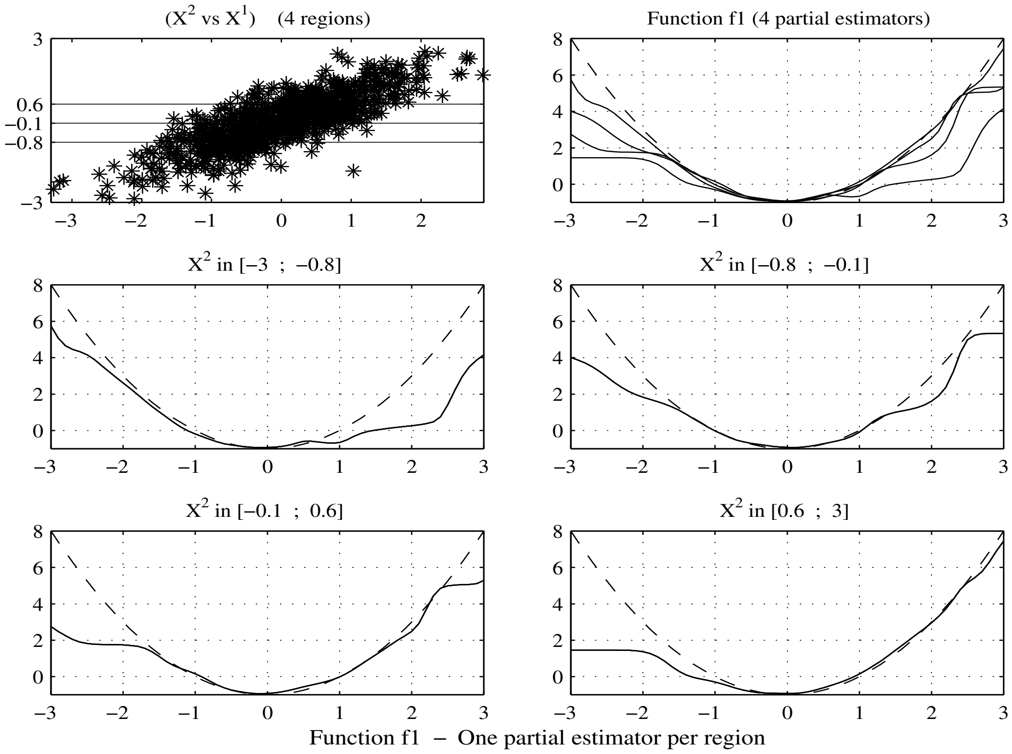

Fig. 6. Partial estimators of f^1. Model $(\text{M1})_{0.8}$ with $n = 800$.

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. To do that, we propose to choose weights depending on the joint density of the variables X^1 and X^2 and on the location of the estimation point.

Using the notations of Paragraph 3, we start from q partial estimators of f (where f stands for any given f^i) denoted by $\widehat{f_n}^{\mathcal{D}_j}$ and defined by (9). Then, we define the following recombined estimator

$$\widehat{f}_n(x) \;=\; \sum_{j=1}^{q} \alpha_j(x)\, \widehat{f_n}^{\mathcal{D}_j}(x) \quad \text{with} \quad \sum_{j=1}^{q} \alpha_j(x) = 1 \tag{12}$$

where $\alpha_j(x)$ is an estimate of $\mathbb{P}\left[Y \in \mathcal{D}_j | X = x\right]$. It is defined by

$$\alpha_j(x) = \Big(\int_{\mathcal{D}_j} \widehat{p}_n^{XY}(x,y)\, dy\Big) \Big/ \widehat{p}_n^{X}(x) \;\text{ if }\; \widehat{p}_n^{X}(x) \neq 0 \text{ and } \alpha_j(x) = 0 \;\text{ otherwise,}$$

where $\widehat{p}_n^{XY}$ and $\widehat{p}_n^{X}$ are usual kernel-based estimators of p^{XY} and p^{X}, respectively. Since the weight $\alpha_j(x)$ estimates the conditional probability that Y belongs to $\mathcal{D}_j$ when $X = x$, the greater $\alpha_j(x)$, the better is (locally) the estimation of $\widehat{m}_n$ and then, the better is the estimation of $\widehat{f_n}^{\mathcal{D}_j}(x)$.

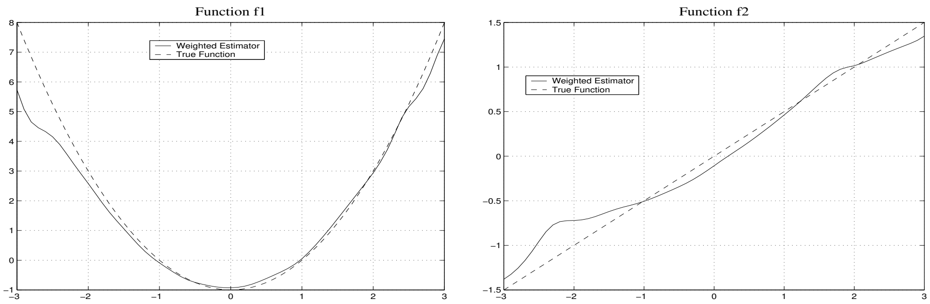

Fig. 7. Recombined estimators for model (M1)$_{0.8}$ with $n = 800$.

Fig. 7, to be compared with Fig. 4, displays the results obtained using this strategy. Functions f^1 and f^2 are now very well estimated in this hard context of high dependence and moderate sample size.

4.2. Illustration

This paragraph is dedicated to the performance study of the recombined estimator. Table 8 of Appendix contains the results obtained using the recombined estimator defined previously, for the models examined for the global estimator. It has to be compared with Table 7 of Appendix.

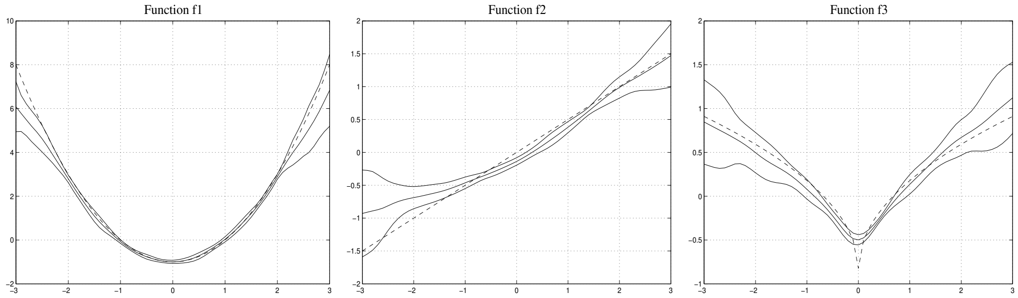

Fig. 8. Recombined estimator. Model (M2)$_{0.8}$ with $n = 2000$.

Firstly, the recombined estimator does not alter the good performance obtained by the global estimator in the weakly dependent case, ie. for Models $(\text{M1})_{0.2}$ and $(\text{M2})_{0.2}$, and slightly improve it. Secondly, it tremendously improve those obtained in the strongly dependent case, ie. for Models $(\text{M1})_{0.8}$ and $(\text{M2})_{0.8}$: the performance become close to those obtained for Models $(\text{M1})_{0.2}$ and $(\text{M2})_{0.2}$. For example, for $(\text{M2})_{0.8}$ with $n = 800$, SZ which estimates the noise variance $\sigma^2 = 0.25$, is equal to 0.74 in Table 7 and becomes 0.3 here. The quantities Sf^i also decrease strongly.
As an illustration, Fig. 8 shows the good performance of the recombined estimator for Model $(\text{M2})_{0.8}$, which is hard to estimate even for a large sample size of $n = 2000$. When it is compared to Fig. 1, the improvement is important and becomes huge outside the interval $[-2\,,\,2]$.

5. 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, 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 a few observations are available, which is the main difficulty.

5.1. The model

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

$$\text{ozon}(j) \;=\; f(\text{temp}(j)) \;+\; g(\text{wind}(j)) \;+\; h(\text{mozon}(j-1)) \;+\; \mu \;+\; e(j) \qquad (13)$$

where j is the index of the day, ozon(j) is the maximum of ozone concentration (μg/m^3) during day j, temp(j) is the maximum of temperature (°C) during day j, wind(j) is the mean of the wind speed (m/s) during the afternoon of day j, 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.

5.2. 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 regions of integration. These regions are defined by interval limits chosen from the marginal distribution of each explanatory variable and according to the knowledge of the problem. These 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.

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 have a tendancy to underestimate the effect of the temperature for a low wind speed and to 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 leads immediately to three different behaviors, since the columns

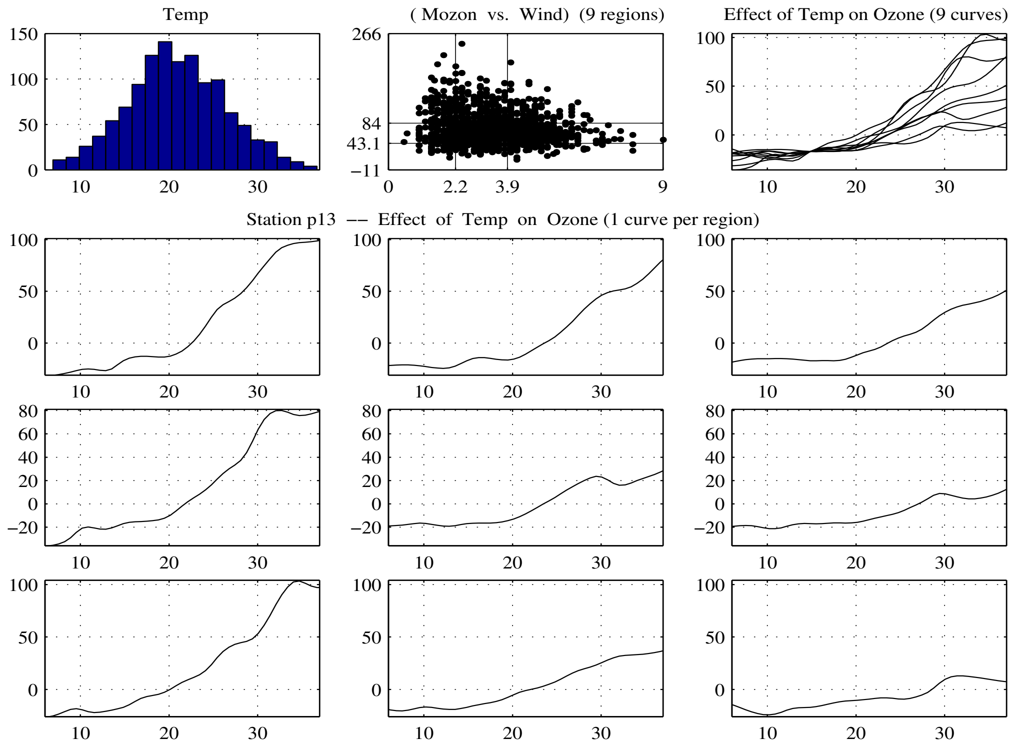

Fig. 9. Partial estimators of the effect of the temperature.

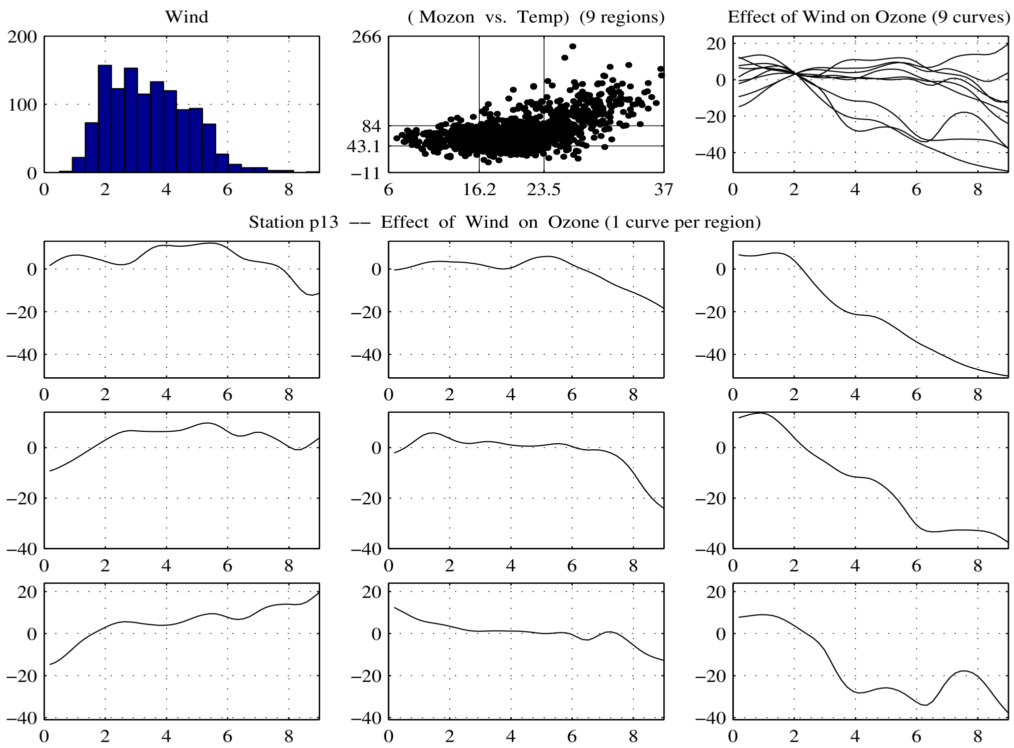

Fig. 10. Partial estimators of the effect of the wind speed.

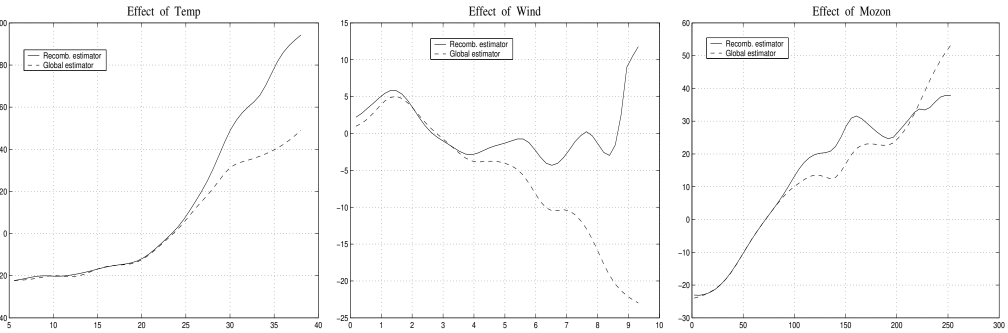

Fig. 11. Global and recombined estimators.

of the 3×3 square displaying the partial estimators are homogeneous, but the rows, not at all. 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 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 additivity 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, 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.

5.3. Ozone prediction using a recombined estimator

For the same 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 (13) is estimated by the mean of the ozone concentrations which is equal to $73\,\mu g/m^3$. Let us compare the two estimators.

Table 2. Alarms table

		Estimated		
		Level 0	**Level 1**	**Level 2**
Measured	**Level 0**	Good alarms	False alarms	
	Level 1		Good alarms	
	Level 2	Missed alarms		Good alarms

Table 3. Simplified alarms table

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

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 μg/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 (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 μg/m^3 (level 1) and 180 μg/m^3 (level 2). The days of level 0 are those for which the maximum of ozone concentration is less than 130 μg/m^3. Criteria closely related to the decisional problem are defined starting from the contingency table 2.
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 2, let us define three usual synthetic criteria coming from the simplified binary contingency table 3.
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 $t_{11}/\,(t_{11} + t_{10} + t_{01})$.
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.

Table 4. 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
Measured	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 5. Alarms table (simplified) for the global and recombined estimators

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

The performance evaluation is based on the days between 1994 and 1998. Tables 4 to 6 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 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.

6. Conclusion

Starting from a variant of a global integration estimator, this paper proposes partial and recombined estimators for nonlinear additive regression models.
Nevertheless, two non-asymptotic features seem, to us, to be promising. The first one deals

Table 6. 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

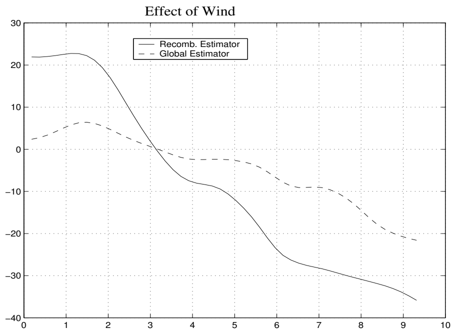

Fig. 12. Global and recombined estimators.

Table 7. Global estimator performance

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

with the use of partial estimators for data analysis purposes. The analysis of ozone concentration in Paris area using additive models shows that, even if the underlying phenomenon is not fully additive, partial estimators do capture some kind of conditional additivity.

The second one is the use of recombined estimators to improve the estimation and prediction performance for finite sample size, which seem to overcome global integration estimators, especially for strongly dependent explanatory variables. This is illustrated by the simulations carried out along the paper as well as in handling real world pollution data.

In the companion paper (Cheze et al. (2001)), some theoretical results are provided for partial estimators, ensuring desirable asymptotic properties and an asymptotic test for partial additivity is derived.

Appendix

This appendix is concerned with the performance of the global estimator defined by (6) and the recombined estimator defined by (12) (see Table 7 and Table 8 respectively). The models considered in both cases are defined in Paragraph 2.3.

Table 8. Recombined estimator performance

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

References

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

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.

Chèze, N., Poggi, J.-M. and Portier, B. (2001). Partial estimators for additive models: multivariate CLT and test for partial additivity. *Preprint Orsay, 2001-75, 17 pages.*

Chèze, N., Poggi, J.-M. and Portier, B. (2000). Partial and Weighted Integration Estimators for Nonlinear Additive Models. *Preprint Orsay 2000-53*, 26 pages.

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

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.

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.

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.