Robust estimators for additive models using backfitting

Graciela Boente

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina

Alejandra Martínez

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina

and

Matias Salibián–Barrera

University of British Columbia, Canada

Abstract

Additive models provide an attractive setup to estimate regression functions in a nonparametric context. They provide a flexible and interpretable model, where each regression function depends only on a single explanatory variable and can be estimated at an optimal univariate rate. It is easy to see that most estimation procedures for these models are highly sensitive to the presence of even a small proportion of outliers in the data. In this paper, we show that a relatively simple robust version of the backfitting algorithm (consisting of using robust local polynomial smoothers) corresponds to the solution of a well-defined optimization problem. This formulation allows us to find mild conditions to show that these estimators are Fisher consistent and to study the convergence of the algorithm. Our numerical experiments show that the resulting estimators have good robustness and efficiency properties. We illustrate the use of these estimators on a real data set where the robust fit reveals the presence of influential outliers.

Keywords: Fisher–consistency; Kernel weights; Robust estimation; Smoothing

1 Introduction

Consider a general regression model, where a response variable $Y \in \mathbb{R}$ is related to a vector $\mathbf{X} = (X_1, \dots, X_d)^{\mathrm{T}} \in \mathbb{R}^d$ of explanatory variables through the following non-parametric regression model:

$$Y = g_0(\mathbf{X}) + \sigma_0 \, \varepsilon \, . \tag{1}$$

The error ε is assumed to be independent from $\mathbf{X}$ and centered at zero, while σ_0 is the error scale parameter. When ε has a finite first moment, we have the usual regression representation $E(Y|\,\mathbf{X}\,) = g_0(\mathbf{X})$. Standard estimators for g_0 can thus be derived relying on local estimates of the conditional mean, such as kernel polynomial regression estimators. It is easy to see that such procedures can be seriously affected either by a small proportion of outliers in the response variable, or when the distribution of $Y|\mathbf{X}$ has heavy tails. Note, however, that even when ε does not have a finite first moment, the function $g_0(\mathbf{X})$ can still be interpreted as a location parameter for the distribution of $Y|\mathbf{X}$. In this case, local robust estimators can be used to estimate the regression function as, for example, the local $M-$estimators proposed in Boente and Fraiman (1989) and the local medians studied in Welsh (1996).

Unfortunately both robust and non-robust non-parametric regression estimators are affected by the *curse of dimensionality*, which is caused by the fact that the expected number of observations in local neighbourhoods decreases exponentially as a function of d, the number of covariates. This results in regression estimators with a very slow convergence rate. Stone (1985) showed that additive models can avoid these problems and produce non-parametric multiple regression estimators with a univariate rate of convergence. In an additive model, the regression function is assumed to satisfy

$$g_0(\mathbf{X}) \, = \, \mu_0 + \sum_{j=1}^{d} g_{0,j}(X_j) \, , \tag{2}$$

where $\mu_0 \in \mathbb{R}$, $g_{0,j} : \mathbb{R} \to \mathbb{R}$, $1 \le j \le d$, are unknown smooth functions with $\mathbb{E}(g_{0,j}(X_j)) = 0$. Such a model retains the ease of interpretation of linear regression models, where each component $g_{0,j}$ can be thought as the effect of the j-th covariate on the centre of the conditional distribution of Y. Moreover, Linton (1997), Fan *et al.* (1998) and Mammen *et al.* (1999) obtained different oracle properties showing that each additive component can be estimated as well as when all the other ones are known.

Several algorithms to fit additive models have been proposed in the literature. In this paper, we focus on the backfitting algorithm as introduced in Friedman and Stuetzle (1981) and discussed further in Buja *et al.* (1989). The backfitting algorithm can be intuitively motivated by observing that, if (2) holds, then

$$g_{0,j}(x) \, = \, \mathbb{E}\biggl(Y - \alpha - \sum_{\ell \neq j} g_{0,\ell}(X_\ell) \biggm| X_j = x \biggr) \, . \tag{3}$$

Hence, given a sample, the backfitting algorithm iteratively estimates the components $g_{0,j}$, $1 \leq j \leq d$, using a univariate smoother of the partial residuals in (3) as functions of the j-th covariate. This algorithm is widely used due to its flexiblity (different univariate smoothers can be used), ease of implementation and intuitive motivation. Furthermore, it has been shown to work very well in simulation studies (Sperlich *et al.* 1999) and applications, although its statistical properties are difficult to study due to its iterative nature. Some results regarding its bias and conditional variance can be found in Opsomer and Ruppert (1997), Wand (1999) and Opsomer (2000).

When second moments exist, Breiman and Friedman (1985) showed that, under certain regularity conditions, the backfitting procedure finds functions $m_1(X_1), \ldots, m_d(X_d)$ minimizing $\mathbb{E}(Y - \mu_0 - \sum_{i=1}^{d} m_j(X_j))^2$ over the space of functions with $\mathbb{E}[m_j(X_j)] = 0$ and finite second moments. In other words, even if the regression function g_0 in (1) does not satisfy the additive model (2), the backfitting algorithm finds the orthogonal projection of the regression function onto the linear space of additive functions in L_2. Equivalently, backfitting finds the closest additive approximation (in the L_2 sense) to $\mathbb{E}(Y|X_1, \ldots, X_d)$. Furthemore, the backfitting algorithm is a coordinate-wise descent algorithm minimizing the squared loss functional above. The sample version of the algorithm solves a system of $nd \times nd$ normal equations and corresponds to the Gauss–Seidel algorithm for linear systems of equations.

If the smoother chosen to estimate (3) is not resistant to outliers then the estimated additive components can be seriously affected by a relatively small proportion of atypical observations. Given the local nature of non–parametric regression estimators, we will be concerned with the case where outliers are present in the response variable. Bianco and Boente (1998) considered robust estimators for additive models using kernel regression, which are a robust version of those defined in Baek and Wehrly (1993). The main drawback of this approach is that it assumes that $Y - g_{0,j}(X_j)$ is independent from X_j, which is difficult to justify or verify in practice. Outlier–resistant fits for generalized additive models have been considered recently in the literature. When the variance is a known function of the mean and the dispersion parameter is known, we refer to Alimadad and Salibián-Barrera (2012) and Wong *et al.* (2014), who consider robust fits based on backfitting and penalized splines $M-$estimators, respectively. In the case of model (1), the approach of Wong *et al.* (2014) reduces to that of Oh *et al.* (2007) which is an alternative based on penalized splines. On the other hand, Croux *et al.* (2011) provides a robust fit for generalized additive models with nuisance parameters using penalized splines, but no theoretical support is provided for their method.

In this paper, we consider an intuitively appealing way to obtain robust estimators for model (1) which combines the backfitting algorithm with robust univariate smoothers. For example, one can consider those proposed in Boente and Fraiman (1989), Härdle and Tsybakov (1988), Härdle (1990) and Oh *et al.* (2007). One of the main contributions of this paper is to show that this intuitive approach to obtain a robust backfitting algorithm is well justified. Specifically, we show that applying the backfitting algorithm using the robust non-parametric regression estimators of Boente and Fraiman (1989) corresponds to minimizing

$\mathbb{E}[\rho((Y-\mu_0-\sum_{i=1}^{d} m_j(X_j))/\sigma_0)]$ over functions $m_1(X_1),\ldots,m_d(X_d)$ with $\mathbb{E}[m_j(X_j)]=0$, where ρ is a loss function. Furthermore, this robust backfitting corresponds to a coordinate-wise descent algorithm and can be shown to converge. We also establish sufficient conditions for these robust backfitting estimators to be Fisher consistent for the true additive components when (2) holds. Our numerical experiments confirm that these estimators have very good finite-sample properties, both in terms of robustness, and efficiency with respect to the classical approach when the data do not contain outliers. These robust estimators cannot be interpreted as orthogonal projections of the regression function onto the space of additive functions of the predictors. However, the first-order conditions for the minimum of this optimization problem are closely related to the robust conditional location functional defined in Boente and Fraiman (1989).

The rest of the paper is organized as follows. In Section 2, we show that the robust backfitting algorithm mentioned above corresponds to a coordinate-descent algorithm to minimize a robust functional using a convex loss function. We also prove that the resulting estimator is Fisher consistent, which means that the solution to the population version of the problem is the object of interest (in our case, the true regression function). The convergence of this algorithm is studied in Section 2.1, while its finite-sample version using local $M-$regression smoothers is presented in Section 3. The results of our numerical experiments conducted to evaluate the performance of the proposed procedure are reported in Section 4. To study the sensitivity of the robust backfitting with respect to single outliers, Section 5 provides a numerical study of the empirical influence function. Finally, in Section 6 we illustrate the advantage of using robust backfitting on a real data set. All proofs are relegated to the Appendix.

2 The robust backfitting functional

In this section, we introduce a population-level version of the robust backfitting algorithm. By showing that the robust backfitting corresponds to a coordinate-descent algorithm to minimize a "robust functional", we are able to find sufficient conditions for the robust backfitting to be Fisher-consistent.

In what follows, we will assume that $(\mathbf{X}^{\mathrm{T}},Y)^{\mathrm{T}}$ is a random vector satisfying the additive model (2), where $Y\in\mathbb{R}$ and $\mathbf{X}=(X_1,\ldots,X_d)^{\mathrm{T}}$, that is,

$$Y=\mu_0+\sum_{j=1}^{d} g_{0,j}(X_j)+\sigma_0\,\varepsilon\,. \tag{4}$$

As it is customary, to ensure identifiability of the components of the model, we will further assume that $\mathbb{E}g_{0,j}(X_j)=0$, $1\le j\le d$. When second moments exist, it is easy to see that the backfitting estimators solve the following minimization problem

$$\min_{(\nu,m)\in\mathbb{R}\times\mathcal{H}^{ad}} \mathbb{E}\Big(Y-\nu-\sum_{j=1}^{d} m_j(X_j)\Big)^2\,, \tag{5}$$

where $\mathcal{H}^{ad}=\Big\{m(\mathbf{x})=\sum_{j=1}^d m_j(x_j),\ m_j\in\mathcal{H}_j\Big\}\subset\mathcal{H}$, $\mathcal{H}=\{r(\mathbf{x}):\mathbb{E}(r(\mathbf{X}))=0\,,\mathbb{E}(r^2(\mathbf{X}))<\infty\}$ and $\mathcal{H}_j$ is the Hilbert space of measurable functions m_j of X_j, with zero mean and finite second moment, i.e., $\mathbb{E}m_j(X_j)=0$ and $\mathbb{E}m_j^2(X_j)<\infty$. The solution to (5) is characterized by its residual $Y-\mu-g(\mathbf{X})$ being orthogonal to $\mathcal{H}^{ad}$. Since this space is spanned by $\mathcal{H}_\ell$, $1\le\ell\le d$, the solution of (5) satisfies $\mathbb{E}(\,Y-\mu-\sum_{j=1}^d g_j(X_j)\,)=0$ and $\mathbb{E}(\,Y-\mu-\sum_{j=1}^d g_j(X_j)\,|\,X_\ell\,)=0$, for $1\le\ell\le d$, from where it follows that $\mu=\mathbb{E}(Y)$ and $g_\ell(X_\ell)=\mathbb{E}(Y-\mu-\sum_{j\ne\ell}g_j(X_j)|X_\ell)$, $1\le\ell\le d$. Given a sample, the backfitting algorithm iterates the above system of equations replacing the conditional expectations with non-parametric regression estimators (e.g. local polynomial smoothers).

To reduce the effect of outliers on the regression estimates, we replace the square loss function in (5) by a function with bounded derivative such as the Huber or Tukey's–loss functions. For these losses, $\rho_c(u)=c^2\rho_1(u/c)$, where $c>0$ is a tuning constant to achieve a given efficiency. The Huber–type loss corresponds to $\rho_1=\rho_{\text{H}}$ while the Tukey's loss to ρ_{T}, where $\rho_{\text{H}}(u)=u^2/2$ if $|u|\le 1$ and $\rho_{\text{H}}(u)=|u|-1/2$ otherwise, while $\rho_{\text{T}}(u)=\min\left(3u^2-3u^4+u^6,1\right)$. Another possible choices are $\rho_1(u)=\sqrt{1+u^2}-1$ which is a smooth approximation of the Huber function and $\rho_1(u)=u\,\arctan(u)-0.5\ln(1+u^2)$ which has derivative $\rho_1'(u)=\arctan(u)$. The bounded derivative of the loss function controls the effect of outlying values in the response variable (sometimes called "vertical outliers" in the literature).

Formally, our objective function is given by

$$\Upsilon(\nu,m)=\mathbb{E}\,\rho\left(\frac{Y-\nu-\sum_{j=1}^d m_j(X_j)}{\sigma_0}\right)\,, \tag{6}$$

where $\rho:\mathbb{R}\to[0,\infty)$ is even, $\nu\in\mathbb{R}$ and the functions $m_j\in\mathcal{H}_j$, $1\le j\le d$. Let P be a distribution in $\mathbb{R}^{d+1}$ and let $(\mathbf{X}^{\text{T}},Y)^{\text{T}}\sim P$. Define the functional $(\mu(P),g(P))$ as the solution of the following optimization problem:

$$(\mu(P),g(P)) \quad = \underset{(\nu,m)\in\mathbb{R}\times\mathcal{H}^{ad}}{\operatorname{argmin}} \quad \Upsilon(\nu,m)\,, \tag{7}$$

where $g(P)(\mathbf{X})=\sum_{j=1}^d g_j(P)(X_j)\in\mathcal{H}^{ad}$.

To prove that the functional in (7) is Fisher-consistent and to derive first-order conditions for the point where it attains its minimum value, we will need the following assumptions:

E1 The random variable ε has a density function $f_0(t)$ that is even, non-increasing in $|t|$, and strictly decreasing for $|t|$ in a neighbourhood of 0.

R1 The function $\rho : \mathbb{R} \to [0, \infty)$ is continuous, non-decreasing, $\rho(0) = 0$, and $\rho(u) = \rho(-u)$. Moreover, if $0 \le u < v$ with $\rho(v) < \sup_t \rho(t)$ then $\rho(u) < \rho(v)$.

A1 Given functions $m_j \in \mathcal{H}_j$, if $\mathbb{P}(\sum_{j=1}^d m_j(X_j) = 0) = 1$ then, for all $1 \le j \le d$, we have $\mathbb{P}(m_j(X_j) = 0) = 1$

Remark 2.1. Assumption **E1** is a standard condition needed to ensure Fisher-consistency of an $M-$location functional (see, e.g. Maronna *et al.*, 2006). Assumption **R1** is satisfied by the so-called family of "rho functions" in Maronna *et al.* (2006), which include many commonly used robust loss functions, such as those mentioned above. Since the loss function ρ can be chosen by the user, this assumption is not restrictive. Finally, assumption **A1** allows us to write the functional $g(P)$ in (7) uniquely as $g(P) = \sum_{j=1}^d g_j(P)$.

Assumption **A1** appears to be the most restrictive and deserves some discussion. It is closely related to the identifiability of the additive model (4) and holds if the explanatory variables are independent from each other. Indeed, let us denote $(x, \mathbf{X}_{\underline{\alpha}})$ the vector with the $\alpha-$th coordinate equal to x and the other ones equal to X_j, $j \ne \alpha$ and by $m(\mathbf{x}) = \sum_{j=1}^d m_j(x_j)$, for $m_j \in \mathcal{H}_j$. For any fixed $1 \le \alpha \le d$, the condition $\mathbb{P}(m(\mathbf{X}) = 0) = 1$ implies that for almost every x_α, $\mathbb{P}(m(x_\alpha, \mathbf{X}_{\underline{\alpha}}) = 0 | X_\alpha = x_\alpha) = 1$. Using that the components of $\mathbf{X}$ are independent, we obtain that $\mathbb{P}(m(x_\alpha, \mathbf{X}_{\underline{\alpha}}) = 0) = 1$ which implies that $\int m(x_\alpha, \mathbf{u}_{\underline{\alpha}}) dF_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = 0$. Note that since $\mathbb{E} m_j(X_j) = 0$ for all j, $\int m(x_\alpha, \mathbf{u}_{\underline{\alpha}}) dF_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = m_\alpha(x_\alpha) + \int \sum_{j \ne \alpha} m_j(u_j) dF_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = m_\alpha(x_\alpha)$. Hence, $m_\alpha(x_\alpha) = 0$, for almost every x_α as desired. However, if the components of $\mathbf{X}$ are not independent, then $\mathbb{P}(m(x_\alpha, \mathbf{X}_{\underline{\alpha}}) = 0 | X_\alpha = x_\alpha) = 1$ does not imply $\int m(x_\alpha, \mathbf{u}_{\underline{\alpha}}) dF_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = 0$. This has already been observed by Hastie and Tibshirani (1990, page 107). The fact that $\mathcal{H}^{ad}$ is closed in $\mathcal{H}$ entails that under mild assumptions, the minimum of $\mathbb{E}(Y - m(\mathbf{X}))^2$ over $\mathcal{H}^{ad}$ exists and is unique. However, the individual functions $m_j(x_j)$ may not be uniquely determined since the dependence among the covariates may lead to more than one representation for the same surface (see also Breiman and Friedman, 1985). In fact, condition **A1** is analogous to assumption 5.1 of Breiman and Friedman (1985). It is also worth noticing that Stone (1985) gives conditions to ensure that **A1** holds. Indeed, Lemma 1 in Stone (1985) implies Proposition 2.1 below which gives weak conditions for the unique representation and hence, as shown in Theorem 2.1 below, for the Fisher–consistency of the functional $g(P)$. Its proof is omitted since it follows straightforwardly.

Proposition 2.1. *Assume that* $\mathbf{X}$ *has compact support* $\mathcal{S}$ *and that its density* $f_{\mathbf{X}}$ *is bounded in* $\mathcal{S}$ *and such that* $\inf_{\mathbf{x} \in \mathcal{S}_f} f_{\mathbf{X}}(\mathbf{x}) > 0$. *Let* $V_j = m_j(X_j)$ *be random variables such that* $\mathbb{P}(\sum_{j=1}^d V_j = 0) = 1$ *and* $\mathbb{E}(V_j) = 0$, *then* $\mathbb{P}(V_j = 0) = 1$.

The next Theorem establishes the Fisher–consistency of the functional $(\mu(P), g(P))$. In other words, it shows that the solution to the optimization problem (7) are the target quantities to be estimated under model (4).

Theorem 2.1. *Assume that the random vector $(\mathbf{X}^{\mathrm{T}}, Y)^{\mathrm{T}} \in \mathbb{R}^{d+1}$ satisfies (4) and let P stand for its distribution.*

*a) If **E1** and **R1** hold, then $\Upsilon(\nu, m)$ in (6) achieves its unique minimum over $\mathbb{R} \times \mathcal{H}^{ad}$ at $(\mu(P), g(P)) = (\mu(P), \sum_{j=1}^d g_j(P))$ when $\mu(P) = \mu_0$ and $\mathbb{P}(\sum_{j=1}^d g_j(P)(X_j) = \sum_{j=1}^d g_{0,j}(X_j)) = 1$.*

*b) If in addition **A1** holds, the unique minimum $(\mu(P), g(P)) = (\mu(P), \sum_{j=1}^d g_j(P))$ satisfies $\mu(P) = \mu_0$ and $\mathbb{P} (g_j(P)(X_j) = g_{0,j}(X_j)) = 1$ for $1 \leq j \leq d$.*

It is worth noticing that a minimizer $(\mu(P), g(P))$ of (7) always exists if ρ is a strictly convex function, even if **E1** does not hold. If in addition **A1** holds, the minimizer will have a unique representation.

For $\nu \in \mathbb{R}$, $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ and $\mathbf{m} = (m_1, \ldots, m_d)^{\mathrm{T}} \in \mathcal{H}_1 \times \mathcal{H}_2 \cdots \times \mathcal{H}_d$ let $\boldsymbol{\Gamma}(\nu, \mathbf{m}, \mathbf{x}) = (\Gamma_0(\nu, \mathbf{m}), \Gamma_1(\nu, \mathbf{m}, x_1), \ldots, \Gamma_d(\nu, \mathbf{m}, x_d))^{\mathrm{T}}$, where

$$\begin{aligned} \Gamma_0(\nu, \mathbf{m}) &= \mathbb{E}\left[\psi\left(\frac{Y - \nu - \sum_{j=1}^d m_j(X_j)}{\sigma_0}\right)\right] \\ \Gamma_\ell(\nu, \mathbf{m}, x_\ell) &= \mathbb{E}\left[\psi\left(\frac{Y - \nu - \sum_{j=1}^d m_j(X_j)}{\sigma_0}\right)\middle| X_\ell = x_\ell\right], \quad 1 \leq \ell \leq d. \end{aligned} \tag{8}$$

Our next theorem shows that it is possible to choose the solution $g(P)$ of (7) so that its additive components $g_j = g_j(P)$ satisfy first order conditions which are generalizations of those corresponding to the classical case where $\rho(u) = u^2$.

Theorem 2.2. *Let ρ be a differentiable function satisfying **R1** and such that its derivative $\rho' = \psi$ is bounded and continuous. Let $(\mathbf{X}^{\mathrm{T}}, Y)^{\mathrm{T}} \sim P$ be a random vector such that $(\mu(P), g(P))$ is a minimizer of $\Upsilon(\nu, m)$ over $\mathbb{R} \times \mathcal{H}^{ad}$ where $\mu(P) \in \mathbb{R}$, $g(P) = \sum_{j=1}^d g_j(P) \in \mathcal{H}^{ad}$, i.e., $(\mu(P), g(P))$ is the solution of (7). Then, $(\mu(P), \mathbf{g}(P))$ satisfies the system of equations $\boldsymbol{\Gamma}(\nu, \mathbf{m}, \mathbf{x}) = \mathbf{0}$ almost surely $P_{\mathbf{X}}$.*

Remark 2.2. a) It is also worth mentioning that if $(\mathbf{X}^{\mathrm{T}}, Y)^{\mathrm{T}}$ satisfies (4) with the errors satisfying **E1**, then $\boldsymbol{\Gamma}(\mu_0, \mathbf{g}_0, \mathbf{x}) = \mathbf{0}$. Moreover, if the model is heteroscedastic

$$Y = g_0(\mathbf{X}) + \sigma_0(\mathbf{X})\,\varepsilon = \mu_0 + \sum_{j=1}^d g_{0,j}(X_j) + \sigma_0(\mathbf{X})\,\varepsilon\,, \tag{9}$$

where the errors ε are symmetrically distributed and the score function ψ is odd, then $(\mu_0, \mathbf{g}_0)$

satisfies

$$\mathbb{E}\left[\psi\left(\frac{Y-\mu_0-\sum_{j=1}^d g_{0,j}(X_j)}{\sigma_0(\mathbf{X})}\right)\right] \;=\; 0\,,$$

$$\mathbb{E}\left[\frac{1}{\sigma_0(\mathbf{X})}\,\psi\left(\frac{Y-\mu_0-\sum_{j\neq\ell} g_{0,j}(X_j)-g_{0,\ell}(X_\ell)}{\sigma_0(\mathbf{X})}\right)\Bigg|\, X_\ell\right] \;=\; 0\,,\quad 1\le \ell\le d\,,$$

which provides a way to extend the robust backfitting algorithm to heteroscedastic models.

b) Assume now that missing responses can arise in the sample, that is, we have a sample $(\mathbf{X}_i^{\mathrm{T}}, Y_i, \delta_i)^{\mathrm{T}}$, $1\le i\le n$, where $\delta_i=1$ if Y_i is observed and $\delta_i=0$ if Y_i is missing, and $(\mathbf{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}$ satisfy an additive heteroscedastic model (9). Let $(\mathbf{X}^{\mathrm{T}}, Y, \delta)^{\mathrm{T}}$ be a random vector with the same distribution as $(\mathbf{X}_i^{\mathrm{T}}, Y_i, \delta_i)^{\mathrm{T}}$. Moreover, assume that responses may be missing at random (MAR), i.e., $\mathbb{P}(\delta=1|(\mathbf{X},Y))=\mathbb{P}(\delta=1|\mathbf{X})=p(\mathbf{X})$. Define $(\mu(P),g(P))=\mathrm{argmin}_{(\nu,m)\in\mathbb{R}\times\mathcal{H}^{ad}}\Upsilon_\delta(\nu,m)$ where

$$\Upsilon_\delta(\nu,m)=\mathbb{E}\,\delta\rho\left(\frac{Y-\nu-\sum_{j=1}^d m_j(X_j)}{\sigma_0(\mathbf{X})}\right)=\mathbb{E}\,p(\mathbf{X})\rho\left(\frac{Y-\nu-\sum_{j=1}^d m_j(X_j)}{\sigma_0(\mathbf{X})}\right)\,.$$

Analogous arguments to those considered in the proof of Theorem 2.1, allow to show that, if **E1** and **R1** hold, $\Upsilon_\delta(\nu,m)$ achieves its unique minimum at $(\nu,m)\in\mathbb{R}\times\mathcal{H}$ where $\nu=\mu_0$ and $\mathbb{P}(m(\mathbf{X})=\sum_{j=1}^d g_{0,j}(X_j))=1$. Besides, if in addition **A1** holds, the unique minimum satisfies that $\mu(P)=\mu_0$ and $\mathbb{P}(g_j(P)(X_j)=g_{0,j}(X_j))=1$, that is, the functional is Fisher–consistent.

On the other hand, the proof of Theorem 2.2 can be also generalized to the case of an homocedastic additive model (4) with missing responses. Effectively, when $\inf_{\mathbf{x}} p(\mathbf{x})>0$, using the MAR assumption, it is possible to show that there exists a unique measurable solution $\widetilde{g}_\ell(x)$ of $\lambda_{\ell,\delta}(x,a)=0$ where

$$\lambda_{\ell,\delta}(x,a) \;=\; \mathbb{E}\left\{p(\mathbf{X})\,\psi\left(\frac{Y-\mu(P)-\sum_{j\neq\ell} g_j(P)(X_j)-a}{\sigma_0}\right)\Bigg|\, X_\ell=x\right\}\,.$$

More precisely, let $\mathbf{\Gamma}_\delta(\nu,\mathbf{m},\mathbf{x})=(\Gamma_{0,\delta}(\nu,\mathbf{m}),\Gamma_{1,\delta}(\nu,\mathbf{m},x_1),\dots,\Gamma_{d,\delta}(\nu,\mathbf{m},x_d))^{\mathrm{T}}$ with $\mathbf{m}=(m_1,\dots,m_d)^{\mathrm{T}}$ and

$$\begin{aligned}\Gamma_{0,\delta}(\nu,\mathbf{m}) \;&=\; \mathbb{E}\left[p(\mathbf{X})\psi\left(\frac{Y-\nu-\sum_{j=1}^d m_j(X_j)}{\sigma_0}\right)\right]\\ \Gamma_{\ell,\delta}(\nu,\mathbf{m},x_\ell) \;&=\; \mathbb{E}\left[p(\mathbf{X})\,\psi\left(\frac{Y-\mu(P)-\sum_{j\neq\ell} m_j(X_j)-m_\ell(X_\ell)}{\sigma}\right)\Bigg|\, X_\ell=x_\ell\right]\quad 1\le\ell\le d\,.\end{aligned}$$

Similar arguments to those considered in the proof of Theorem 2.2, allow to show that if there exists a unique minimizer $(\mu(P),g(P))\in\mathbb{R}\times\mathcal{H}^{ad}$ of $\Upsilon_\delta(\nu,m)$, then $(\mu(P),\mathbf{g}(P))$ is a solution of $\mathbf{\Gamma}_\delta(\nu, \mathbf{m}, \mathbf{x}) = \mathbf{0}$. Note that instead of a simplified approach, a propensity score approach can also be considered taking $\delta/p(\mathbf{X})$ instead of δ. In this case, $\Upsilon_\delta(\nu, m) = \Upsilon(\nu, m)$ defined in (6) and $\Gamma_{\ell,\delta} = \Gamma_\ell$ defined in (8). The propensity approach is useful when preliminary estimates of the missing probability are available, otherwise, the simplified approach is preferred.

2.1 The population version of the robust backfitting algorithm

In this section, we derive an algorithm to solve (7) and study its convergence. For simplicity, we will assume that the vector $(\mathbf{X}^{\mathrm{T}}, Y)^{\mathrm{T}}$ is completely observed and that it satisfies (4). By Theorem 2.2, the robust functional $(\mu(P), \mathbf{g}(P))$ satisfies (8). To simplify the notation, in what follows we will put $\mu = \mu(P)$ and $g_j = g_j(P)$, $1 \leq j \leq d$ and $\sum_{s=\ell}^{m} a_s$ will be understood as 0 if $m < \ell$. The robust backfitting algorithm is given in Algorithm 1.

Algorithm 1 Population version of the robust backfitting

```
Let $\ell = 0$ and $\mathbf{g}^{(0)} = (g_1^{(0)}, \dots, g_d^{(0)})^{\mathrm{T}}$ be an initial set of additive components, for example:
  $\mathbf{g}^{(0)} = \mathbf{0}$ and $\mu^0$ an initial location parameter.
repeat
    $\ell \leftarrow \ell + 1$
    for $j = 1$ to $d$ do
        Let $R_j^{(\ell)} = Y - \mu^{(\ell-1)} - \sum_{s=1}^{j-1} \widetilde{g}_s^{(\ell)}(X_s) - \sum_{s=j+1}^{d} g_s^{(\ell-1)}(X_s)$
        Let $\widetilde{g}_j^{(\ell)}$ solve
            $$\mathbb{E}\left[\psi\left(\frac{R_j^{(\ell)} - \widetilde{g}_j^{(\ell)}(X_j)}{\sigma_0}\right) \middle| X_j = x\right] = 0 \quad \text{a.s.}$$
    end for
    for $j = 1$ to $d$ do
        $g_j^{(\ell)} = \widetilde{g}_j^{(\ell)} - \mathbb{E}[\widetilde{g}_j^{(\ell)}(X_j)]$.
    end for
    Let $\mu^{(\ell)}$ solve
        $$\mathbb{E}\left[\psi\left(\frac{Y - \mu^{(\ell)} - \sum_{j=1}^{d} g_j^{(\ell)}(X_j)}{\sigma_0}\right)\right] = 0\,.$$
until convergence
```

Our next Theorem shows that each **Step** ℓ of the algorithm above reduces the objective function $\Upsilon(\mu^{(\ell)}, g^{(\ell)})$.

Theorem 2.3. *Let ρ be a differentiable function satisfying* ***R1*** *and such that its derivative $\rho' = \psi$ is a strictly increasing, bounded and continuous function with $\lim_{t\to+\infty} \psi(t) > 0$ and $\lim_{t\to-\infty} \psi(t) < 0$. Let $\left(\mu^{(\ell)}, \mathbf{g}^{(\ell)}\right)_{\ell\geq1} = (\mu^{(\ell)}, g_1^{(\ell)}, \dots, g_d^{(\ell)})_{\ell\geq1}$ be the sequence obtained* *with Algorithm 1. Then, $\{\Upsilon(\mu^{(\ell)}, g^{(\ell)})\}_{\ell\geq 1}$ is a decreasing sequence, so that the algorithm converges.*

3 The sample version of the robust backfitting algorithm

In practice, given a random sample $(\mathbf{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}$ $1 \leq i \leq n$ from the additive model (4) we apply Algorithm 1 replacing the unknown conditional expectations with univariate robust smoothers. Different smoothers can be considered, including splines, kernel weights or even nearest neighbours with kernel weights. In what follows we describe the algorithm for kernel polynomial M-estimators.

Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function and let $K_h(t) = (1/h)K(t/h)$. The estimators of the solutions of (8) using kernel $M-$polynomial estimators of order $q \geq 0$ are given by the solution to the following system of equations:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i - \widehat{\mu} - \sum_{j=1}^{d}\widehat{g}_j(X_{i,j})}{\widehat{\sigma}_0}\right) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n}K_{h_j}(X_{i,j}-x_j)\psi\left(\frac{Y_i - \widehat{\mu} - \sum_{\ell\neq j}\widehat{g}_\ell(X_{i,\ell}) - \sum_{s=0}^{q}\beta_{s,j}Z_{i,j,s}}{\widehat{\sigma}_0}\right)\mathbf{Z}_{i,j}(x_j) &= \mathbf{0}\,,\ 1 \leq j \leq d\,, \end{aligned}$$

where $\mathbf{Z}_{i,j}(x_j) = (Z_{i,j,0}, Z_{i,j,1}, \dots, Z_{i,j,q})^{\mathrm{T}}$ with $Z_{i,j,s} = (X_{i,j} - x_j)^s$, $0 \leq s \leq d$. Then, we have $\widehat{g}_j(x_j) = \beta_{0,j}$, $1 \leq j \leq d$. The corresponding algorithm is described in detail in Algorithm 2. The same procedure can be applied to the complete sample when responses are missing.

Remark 3.1. A possible choice of the preliminary scale estimator $\widehat{\sigma}_0$ is obtained by calculating the MAD of the residuals obtained with a simple and robust nonparametric regression estimator, as local medians. In that case we have $\widehat{\sigma}_0 = \text{mad}_{1\leq i\leq n}\left\{Y_i - \widehat{Y}_i\right\}$, where $\widehat{Y}_i = \text{median}_{1\leq j\leq n}\{Y_j : |X_{j,k} - X_{i,k}| \leq h_k, \forall\, 1 \leq k \leq d\}$. The bandwidths h_k are preliminary values to be selected, or alternatively they can be chosen as the distance between $X_{i,k}$ and its r-th nearest neighbour among $\{X_{j,k}\}_{j\neq i}$.

4 Numerical studies

This section contains the results of a simulation study designed to compare the behaviour of our proposed estimator with that of the classical one. We generated $N = 500$ samples of size $n = 500$ for models with $d = 2$ and $d = 4$ components. We considered samples without outliers and also samples contaminated in different ways. We also included in our

Algorithm 2 The sample version of the robust backfitting

```
Let $\ell = 0$ and $\widehat{\mathbf{g}}^{(0)} = (\widehat{g}_1^{(0)}, \dots, \widehat{g}_d^{(0)})^{\mathrm{T}}$ be an initial set of additive components, for example: $\widehat{\mathbf{g}}^{(0)} = \mathbf{0}$, and let $\widehat{\sigma}_0$ be a robust residual scale estimator. Moreover, let $\widehat{\mu}^{(0)}$ an initial location estimator such as an $M-$location estimator of the responses.
repeat
    $\ell \leftarrow \ell + 1$
    for $j = 1$ to $d$ do
        for $i_0 = 1$ to $n$ do
            Let $x_j = X_{i_0,j}$
            for $i = 1$ to $n$ do
                Let $\mathbf{Z}_{i,j}(x_j) = (1, (X_{i,j} - x_j), (X_{i,j} - x_j)^2, \dots, (X_{i,j} - x_j)^q)^{\mathrm{T}}$ and $\widehat{R}_{i,j}^{(\ell)} = Y_i - \widehat{\mu}^{(\ell)} - \sum_{s=1}^{j-1} \widetilde{g}_s^{(\ell)}(X_{i,s}) - \sum_{s=j+1}^{d} \widehat{g}_s^{(\ell-1)}(X_{i,s})$.
            end for
            Let $\widehat{\boldsymbol{\beta}}_j(x_j) = (\widehat{\beta}_{0j}(x_j), \widehat{\beta}_{1j}(x_j), \dots, \widehat{\beta}_{qj}(x_j))^{\mathrm{T}}$ be the solution to
                $$\frac{1}{n} \sum_{i=1}^{n} K_h(X_{i,j} - x_j)\, \psi\left( \frac{\widehat{R}_{i,j}^{(\ell)} - \widehat{\boldsymbol{\beta}}_j(x_j)^{\mathrm{T}} \mathbf{Z}_{i,j}(x_j)}{\widehat{\sigma}_0} \right) \mathbf{Z}_{i,j}(x_j) \;=\; \mathbf{0}\,.$$
            Let $\widetilde{g}_j^{(\ell)}(x_j) = \widehat{\beta}_{0j}(x_j)$.
        end for
    end for
    for $j = 1$ to $d$ do
        $\widehat{g}_j^{(\ell)} = \widetilde{g}_j^{(\ell)} - \sum_{i=1}^{n} \widetilde{g}_j^{(\ell)}(X_{i,j})/n$.
    end for
    Let $\widehat{\mu}^{(\ell)}$ solve
        $$\frac{1}{n} \sum_{i=1}^{n} \psi\left( \frac{Y_i - \widehat{\mu}^{(\ell)} - \sum_{j=1}^{d} \widehat{g}_j^{(\ell)}(X_{i,j})}{\widehat{\sigma}_0} \right) \;=\; 0\,.$$
until convergence
```

experiment cases where the response variable may be missing, as described in Remark 2.2. All computations were carried out using an `R` implementation of our algorithm, publicly available on–line at `http://www.stat.ubc.ca/~matias/soft.html`.

To generate missing responses, we first generated observations $(\mathbf{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}$ satisfying the additive model $Y = g_0(\mathbf{X}) + u = \mu_0 + \sum_{j=1}^{d} g_{0,j}(X_j) + u$, where $u = \sigma_0\, \varepsilon$. Then, we generate $\{\delta_i\}_{i=1}^{n}$ independent Bernoulli random variables such that $\mathbb{P}\left(\delta_i = 1 | Y_i, \mathbf{X}_i\right) = \mathbb{P}\left(\delta_i = 1 | \mathbf{X}_i\right) = p\left(\mathbf{X}_i\right)$ where we used a different function p for each value of d. When $d = 2$ we used $p(\mathbf{x}) = p_2(\mathbf{x}) = 0.4 + 0.5(\cos(x_1 + 0.2))^2$ which yields around 31.5% of missing responses. For $d = 4$, we set $p(\mathbf{x}) = p_4(\mathbf{x}) = 0.4 + 0.5(\cos(x_1 * x_3 + 0.2))^2$, which produces approximately 33% of missing Y_i's. In addition, we also consider the case where all responses are observed, i.e., $p(\mathbf{x}) \equiv 1$.

We compared the following estimators:

- The classical backfitting estimator adapted to missing responses, denoted $\widehat{g}_{\text{BC}}$.
- A robust backfitting estimator $\widehat{g}_{\text{BR,H}}$ using Huber's loss function with tuning constant $c = 1.345$. This loss function is such that $\rho_c'(u) = \psi_c(u) = \min(c, \max(-c, u))$.
- A robust backfitting estimator $\widehat{g}_{\text{BR,T}}$ using Tukey's bisquare loss function with tuning constant $c = 4.685$. This loss function satisfies $\rho_c'(u) = \psi_c(u) = u\ (1 - (u/c)^2)^2\ \mathbb{I}_{[-c,c]}(u)$.

The univariate smoothers were computed using the Epanechnikov kernel $K(u) = 0.75\,(1 - u^2)\mathbb{I}_{[-1,1]}(u)$. We used both local constants and local linear polynomials which correspond to $q = 0$ and $q = 1$ in Algorithm 2 denoted in all Tables and Figures with a subscript of 0 and 1, respectively.

The performance of each estimator $\widehat{g}_j$ of $g_{0,j}$, $1 \leq j \leq d$, was measured through the following approximated integrated squared error (ISE):

$$\text{ISE}(\widehat{g}_j) = \frac{1}{\sum_{i=1}^n \delta_i} \sum_{i=1}^n \left(g_{0,j}\left(X_{ij}\right) - \widehat{g}_j\left(X_{ij}\right)\right)^2 \delta_i\,.$$

where X_{ij} is the jth component of $\mathbf{X}_i$ and $\delta_i = 0$ if the i-th response was missing and $\delta_i = 1$ otherwise. An approximation of the mean integrated squared error (MISE) was obtained by averaging the ISE above over all replications. A similar measure was used to compare the estimators of the regression function $g_0 = \mu_0 + \sum_{j=1}^d g_{0,j}$.

4.1 Monte Carlo study with $d = 2$ additive components

In this case, the covariates were generated from a uniform distribution on the unit square, $\mathbf{X}_i = (X_{i,1}, X_{i,2})^{\text{T}} \sim U([0,1]^2)$, the error scale was $\sigma_0 = 0.5$ and the overall location $\mu_0 = 0$. The additive components were chosen to be

$$g_{0,1}(x_1) \;=\; 24\,(x_1 - 0.5)^2 - 2\,, \qquad\qquad g_{0,2}(x_2) \;=\; 2\pi\sin(\pi x_2) - 4\,. \tag{10}$$

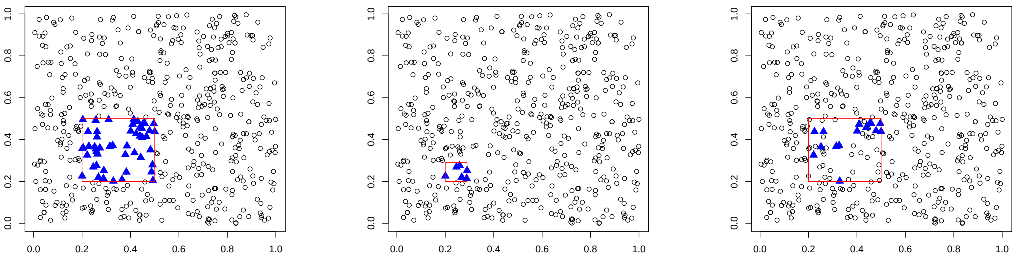
Optimal bandwidths with respect to the MISE can be computed assuming that the other components in the model are known (see, for instance, Härdle *et al.*, 2004). Since the explanatory variables are uniformly distributed, the optimal bandwidths for the local constant ($q = 0$) and local linear ($q = 1$) estimators are the same and very close to our choice of $\mathbf{h} = (0.10, 0.10)$.

For the errors, we considered the following settings:

- C_0: $u_i \sim N(0, \sigma_0^2)$.

- C_1: $u_i \sim (1-0.15)\, N(0,\sigma_0^2) + 0.15\, N(15, 0.01)$.
- C_2: $u_i \sim N(15, 0.01)$ for all i's such that $\mathbf{X}_i \in \mathcal{D}_{0.3}$, where $\mathcal{D}_\eta = [0.2, 0.2+\eta]^2$.
- C_3: $u_i \sim N(10, 0.01)$ for all i's such that $\mathbf{X}_i \in \mathcal{D}_{0.09}$, where $\mathcal{D}_\eta$ is as above.
- C_4: $u_i \sim (1-0.30)\, N(0,\sigma_0^2) + 0.30\, N(15, 0.01)$ for all i's such that $\mathbf{X}_i \in \mathcal{D}_{0.3}$.

Case C_0 corresponds to samples without outliers and they will illustrate the loss of efficiency incurred by using a robust estimator when it may not be needed. The contamination setting C_1 corresponds to a *gross-error model* where all observations have the same chance of being contaminated. On the other hand, case C_2 is highly pathological in the sense that all observations with covariates in the square $[0.2, 0.5] \times [0.2, 0.5]$ are severely affected, while C_3 is similar but in the region $[0.2, 0.29] \times [0.2, 0.29]$. The difference between C_2 and C_3 is that the bandwidths we used ($h = 0.10$) are wider than the contaminated region in C_3. Finally, case C_4 is a gross-error model with a higher probability of observing an outlier, but these are restricted to the square $[0.2, 05] \times [0.2, 0.5]$. Figure 1 illustrate these contamination scenarios on one randomly generated sample. The panels correspond to settings C_2, C_3 and C_4, with solid triangles indicating contaminated cases.



(a) C_2 (b) C_3 (c) C_4

Figure 1: Scatter plots of covariates $(x_1, x_2)^{\mathrm{T}}$ with solid triangles indicating observations with contaminated response variables, for contamination settings C_2, C_3 and C_4. The square regions indicate the sets $\mathcal{D}_\eta$ for each scenario.

Tables 1 to 3 report the obtained values of the MISE when estimating the regression function $g_0 = \mu_0 + g_{0,1} + g_{0,2}$ and each additive component $g_{0,1}$ and $g_{0,2}$, respectively. Estimates obtained using local constant ($q = 0$) and local linear smoothers ($q = 1$) are indicated with a subscript "0" and "1", respectively. To complement the reported results on the MISE, given in Tables 1 to 3, we report in Tables 4 to 6 the ratio between the mean integrated squared error (MISE) obtained for each considered contamination and for the clean data. To identify the ratio, it is denoted as $\text{MISE}(C_j)/\text{MISE}(C_0)$.

Consider first the situation where there are no missing responses. As expected, when the data do not contain outliers the robust estimators are slightly less efficient than the

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$
C_0	0.0704	0.0703	0.0718	0.0077	0.0079	0.0079	0.0741	0.0756	0.0790	0.0110	0.0113	0.0113
C_1	5.8437	0.1306	0.0730	5.9026	0.0620	0.0091	6.1657	0.1457	0.0800	6.3125	0.0897	0.0272
C_2	8.5823	0.2470	0.0753	8.5218	0.1471	0.0086	10.0841	0.3550	0.0837	10.0100	0.3040	0.0396
C_3	0.1560	0.0722	0.0719	0.0930	0.0090	0.0080	0.1882	0.0786	0.0790	0.1224	0.0127	0.0113
C_4	0.9159	0.0764	0.0718	0.8523	0.0122	0.0080	1.1017	0.0840	0.0782	1.0307	0.0169	0.0115

Table 1: MISE of the estimators of the regression function $g_0 = \mu_0 + g_{0,1} + g_{0,2}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$
C_0	0.0445	0.0445	0.0454	0.0096	0.0097	0.0097	0.0493	0.0502	0.0521	0.0138	0.0140	0.0140
C_1	0.3638	0.0503	0.0459	0.3903	0.0152	0.0103	0.5176	0.0592	0.0523	0.5904	0.0334	0.0271
C_2	3.5490	0.1348	0.0473	3.3751	0.0665	0.0101	3.7935	0.1641	0.0547	3.5914	0.0942	0.0176
C_3	0.0935	0.0468	0.0454	0.0490	0.0103	0.0097	0.1084	0.0532	0.0521	0.0599	0.0147	0.0140
C_4	0.4351	0.0514	0.0453	0.3547	0.0119	0.0098	0.4910	0.0587	0.0517	0.3991	0.0166	0.0141

Table 2: MISE of the estimators of the additive component $g_{0,1}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$
C_0	0.0405	0.0406	0.0415	0.0106	0.0107	0.0107	0.0465	0.0476	0.0497	0.0157	0.0159	0.0159
C_1	0.3559	0.0494	0.0418	0.3907	0.0160	0.0113	0.4945	0.0599	0.0498	0.5731	0.0273	0.0184
C_2	3.1891	0.0932	0.0434	3.3314	0.0633	0.0110	4.0168	0.1572	0.0520	4.1969	0.1642	0.0385
C_3	0.0683	0.0403	0.0416	0.0480	0.0111	0.0107	0.0893	0.0476	0.0496	0.0695	0.0165	0.0159
C_4	0.3237	0.0391	0.0415	0.3415	0.0119	0.0108	0.4195	0.0465	0.0492	0.4433	0.0177	0.0160

Table 3: MISE of the estimators of the additive component $g_{0,2}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	82.9580	1.8577	1.0161	766.6964	7.8275	1.1499	83.1557	1.9285	1.0124	575.4987	7.9530	2.4127
$\text{MISE}(C_2)/\text{MISE}(C_0)$	121.8359	3.5141	1.0492	1106.9072	18.5640	1.0907	136.0025	4.6984	1.0594	912.5855	26.9607	3.5103
$\text{MISE}(C_3)/\text{MISE}(C_0)$	2.2147	1.0277	1.0015	12.0853	1.1315	1.0051	2.5383	1.0407	0.9996	11.1610	1.1286	1.0060
$\text{MISE}(C_4)/\text{MISE}(C_0)$	13.0019	1.0869	0.9997	110.7086	1.5411	1.0138	14.8589	1.1118	0.9898	93.9705	1.4974	1.0160

Table 4: Ratio between the MISE of the estimators of the additive component $g_0 = \mu_0 + \sum_{j=1}^{2} g_{0,j}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	8.1696	1.1298	1.0113	40.5953	1.5601	1.0552	10.5007	1.1803	1.0044	42.6524	2.3876	1.9396
$\text{MISE}(C_2)/\text{MISE}(C_0)$	79.6889	3.0283	1.0427	351.0462	6.8407	1.0395	76.9559	3.2722	1.0495	259.4529	6.7405	1.2555
$\text{MISE}(C_3)/\text{MISE}(C_0)$	2.0994	1.0500	1.0020	5.0949	1.0586	1.0024	2.1986	1.0609	1.0008	4.3251	1.0529	1.0031
$\text{MISE}(C_4)/\text{MISE}(C_0)$	9.7694	1.1541	0.9993	36.8950	1.2287	1.0063	9.9606	1.1701	0.9934	28.8339	1.1895	1.0073

Table 5: Ratio between the MISE of the estimators of the additive component $g_{0,1}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	8.7994	1.2173	1.0062	36.7398	1.4951	1.0528	10.6429	1.2567	1.0022	36.4372	1.7173	1.1599
$\text{MISE}(C_2)/\text{MISE}(C_0)$	78.8384	2.2953	1.0454	313.3083	5.8990	1.0285	86.4525	3.2998	1.0461	266.8154	10.3392	2.4245
$\text{MISE}(C_3)/\text{MISE}(C_0)$	1.6873	0.9925	1.0010	4.5158	1.0365	1.0020	1.9229	0.9997	0.9985	4.4195	1.0386	1.0027
$\text{MISE}(C_4)/\text{MISE}(C_0)$	8.0022	0.9615	1.0000	32.1152	1.1092	1.0054	9.0289	0.9769	0.9896	28.1834	1.1154	1.0063

Table 6: Ratio between the MISE of the estimators of the additive component $g_{0,1}$ under the considered contaminations and under clean data for the two missing mechanisms.

least squares one, although the differences in the estimated MISE's are well within the Monte Carlo margin of error. For the contamination cases C_1 and C_2, when using either local constants or local linear smoothers, the MISE of the classical estimator for g_0 is notably larger than those of the robust estimators (more than 40 times larger). This difference is smaller when estimating each component $g_{0,1}$ and $g_{0,2}$, but remains fairly large nonetheless. In general, when the data contain outliers, we note that the robust estimators give noticeably better regression estimators (both for g_0 and its components) than the classical one. The estimator based on Tukey's bisquare function is very stable across all the contamination cases considered here, while for the Huber one, there's a slight increase in MISE for the estimated additive components under the contamination setting C_1 and a larger one under C_2. The advantage of the estimators based on Tukey's loss function over those based on the Huber one becomes more apparent when one inspects the ratio between the MISE's obtained with and without outliers given in Tables 4 to 6.

It is worth noting that, for clean data, local linear smoothers achieve much smaller MISE values than local constants. This may be explained by the well-known bias-reducing property of local polynomials, particularly near the boundary. This behaviour can also be observed across contamination settings for the robust estimators.

Interestingly, the results of our experiments with missing responses yield the same overall conclusions. The estimators based on the subset of complete data remain Fisher-consistent, but, as one would expect, they all yield larger MISE's due to the smaller sample sizes used to compute them. Moreover, the estimators based on Tukey's loss function are still preferable, although they are slightly less efficient than those based on the Huber loss when $q = 0$. When comparing the behaviour of the robust local constant and local linear estimators one notices that, as above, local linear estimators have smaller MISE's. However, the ratios of MISE, reported in Tables 4 to 6, show that under C_2 the MISE's of the local linear estimators of g_0 with missing responses are more than 4 times larger than those obtained with local constants. This effect is mainly due to the poor estimation of $g_{0,2}$ and of the location parameter μ_0 as suggested by the reported results.

We also looked at the median ISE across the simulation replications to determine whether the differences in the observed averaged ISE's are representative of most samples or they were influenced by poor fits obtained in a few bad samples. Tables 7 to 9 report the median over replications of the ISE, denoted MEDISE for the estimators of the regression function g_0 and each additive component when $d = 2$. Besides, Tables 10 to 12 report the ratio between the median integrated squared error (MEDISE) obtained for each considered contamination and for the clean data. To identify the ratio, it is denoted as MEDISE(C_j)/MEDISE(C_0). The median ISE results show that the robust estimators behave similarly when $p(\mathbf{x}) \equiv 1$ and when responses may be missing. Furthermore, the performances of Tukey's local linear and local constant smoothers are equally good.

Based on the above results we recommend using local linear smoothers computed with Tukey's loss function, although in some situations local constants may give better results,

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{\mathrm{BC},0}$	$\widehat{g}_{\mathrm{BR,H},0}$	$\widehat{g}_{\mathrm{BR,T},0}$	$\widehat{g}_{\mathrm{BC},1}$	$\widehat{g}_{\mathrm{BR,H},1}$	$\widehat{g}_{\mathrm{BR,T},1}$	$\widehat{g}_{\mathrm{BC},0}$	$\widehat{g}_{\mathrm{BR,H},0}$	$\widehat{g}_{\mathrm{BR,T},0}$	$\widehat{g}_{\mathrm{BC},1}$	$\widehat{g}_{\mathrm{BR,H},1}$	$\widehat{g}_{\mathrm{BR,T},1}$
C_0	0.0700	0.0700	0.0714	0.0073	0.0074	0.0074	0.0733	0.0752	0.0784	0.0105	0.0108	0.0107
C_1	5.7681	0.1256	0.0713	5.8235	0.0587	0.0086	6.0524	0.1379	0.0793	6.1475	0.0676	0.0126
C_2	8.4718	0.2370	0.0747	8.4056	0.1333	0.0082	9.7880	0.2982	0.0825	9.7619	0.1794	0.0118
C_3	0.1375	0.0717	0.0710	0.0753	0.0083	0.0075	0.1535	0.0783	0.0781	0.0892	0.0122	0.0108
C_4	0.8481	0.0762	0.0712	0.7852	0.0113	0.0075	0.9826	0.0827	0.0776	0.9236	0.0157	0.0110

Table 7: MEDISE of the estimators of the regression function g_0 under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{1,\mathrm{BC},0}$	$\widehat{g}_{1,\mathrm{BR,H},0}$	$\widehat{g}_{1,\mathrm{BR,T},0}$	$\widehat{g}_{1,\mathrm{BC},1}$	$\widehat{g}_{1,\mathrm{BR,H},1}$	$\widehat{g}_{1,\mathrm{BR,T},1}$	$\widehat{g}_{1,\mathrm{BC},0}$	$\widehat{g}_{1,\mathrm{BR,H},0}$	$\widehat{g}_{1,\mathrm{BR,T},0}$	$\widehat{g}_{1,\mathrm{BC},1}$	$\widehat{g}_{1,\mathrm{BR,H},1}$	$\widehat{g}_{1,\mathrm{BR,T},1}$
C_0	0.0430	0.0433	0.0440	0.0068	0.0069	0.0070	0.0465	0.0478	0.0501	0.0099	0.0099	0.0099
C_1	0.3296	0.0489	0.0443	0.3597	0.0130	0.0073	0.4574	0.0557	0.0498	0.5377	0.0194	0.0109
C_2	3.5188	0.1292	0.0453	3.3350	0.0613	0.0073	3.7000	0.1526	0.0523	3.4773	0.0750	0.0108
C_3	0.0865	0.0452	0.0441	0.0406	0.0078	0.0070	0.0981	0.0510	0.0502	0.0498	0.0109	0.0099
C_4	0.4121	0.0500	0.0436	0.3325	0.0096	0.0070	0.4404	0.0567	0.0496	0.3520	0.0131	0.0100

Table 8: MEDISE of the estimators of the additive component $g_{0,1}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{2,\mathrm{BC},0}$	$\widehat{g}_{2,\mathrm{BR,H},0}$	$\widehat{g}_{2,\mathrm{BR,T},0}$	$\widehat{g}_{2,\mathrm{BC},1}$	$\widehat{g}_{2,\mathrm{BR,H},1}$	$\widehat{g}_{2,\mathrm{BR,T},1}$	$\widehat{g}_{2,\mathrm{BC},0}$	$\widehat{g}_{2,\mathrm{BR,H},0}$	$\widehat{g}_{2,\mathrm{BR,T},0}$	$\widehat{g}_{2,\mathrm{BC},1}$	$\widehat{g}_{2,\mathrm{BR,H},1}$	$\widehat{g}_{2,\mathrm{BR,T},1}$
C_0	0.0382	0.0383	0.0391	0.0071	0.0072	0.0072	0.0430	0.0443	0.0462	0.0102	0.0103	0.0103
C_1	0.3180	0.0465	0.0398	0.3517	0.0130	0.0078	0.4503	0.0563	0.0463	0.5036	0.0193	0.0114
C_2	3.1744	0.0874	0.0409	3.3084	0.0575	0.0077	3.9785	0.1258	0.0489	4.1319	0.0884	0.0113
C_3	0.0620	0.0381	0.0387	0.0409	0.0077	0.0072	0.0767	0.0437	0.0463	0.0588	0.0111	0.0104
C_4	0.2993	0.0369	0.0391	0.3171	0.0086	0.0073	0.3722	0.0434	0.0461	0.3955	0.0122	0.0105

Table 9: MEDISE of the estimators of the additive component $g_{0,2}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{\mathrm{BC},0}$	$\widehat{g}_{\mathrm{BR,H},0}$	$\widehat{g}_{\mathrm{BR,T},0}$	$\widehat{g}_{\mathrm{BC},1}$	$\widehat{g}_{\mathrm{BR,H},1}$	$\widehat{g}_{\mathrm{BR,T},1}$	$\widehat{g}_{\mathrm{BC},0}$	$\widehat{g}_{\mathrm{BR,H},0}$	$\widehat{g}_{\mathrm{BR,T},0}$	$\widehat{g}_{\mathrm{BC},1}$	$\widehat{g}_{\mathrm{BR,H},1}$	$\widehat{g}_{\mathrm{BR,T},1}$
$\frac{\text{MEDISE}(C_1)}{\text{MEDISE}(C_0)}$	82.3814	1.7957	0.9991	802.7378	7.9218	1.1601	82.6104	1.8354	1.0113	584.4446	6.2633	1.1736
$\frac{\text{MEDISE}(C_2)}{\text{MEDISE}(C_0)}$	120.9971	3.3876	1.0457	1158.6577	17.9962	1.1054	133.5979	3.9676	1.0525	928.0673	16.6135	1.0978
$\frac{\text{MEDISE}(C_3)}{\text{MEDISE}(C_0)}$	1.9638	1.0254	0.9951	10.3761	1.1264	1.0095	2.0946	1.0416	0.9961	8.4822	1.1289	1.0078
$\frac{\text{MEDISE}(C_4)}{\text{MEDISE}(C_0)}$	12.1124	1.0899	0.9971	108.2387	1.5199	1.0123	13.4123	1.1003	0.9906	87.8085	1.4519	1.0276

Table 10: Ratio between the MEDISE of the estimators of the additive component g_0 under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{1,\mathrm{BC},0}$	$\widehat{g}_{1,\mathrm{BR,H},0}$	$\widehat{g}_{1,\mathrm{BR,T},0}$	$\widehat{g}_{1,\mathrm{BC},1}$	$\widehat{g}_{1,\mathrm{BR,H},1}$	$\widehat{g}_{1,\mathrm{BR,T},1}$	$\widehat{g}_{1,\mathrm{BC},0}$	$\widehat{g}_{1,\mathrm{BR,H},0}$	$\widehat{g}_{1,\mathrm{BR,T},0}$	$\widehat{g}_{1,\mathrm{BC},1}$	$\widehat{g}_{1,\mathrm{BR,H},1}$	$\widehat{g}_{1,\mathrm{BR,T},1}$
$\frac{\text{MEDISE}(C_1)}{\text{MEDISE}(C_0)}$	7.6679	1.1314	1.0061	52.8814	1.8807	1.0462	9.8420	1.1645	0.9930	54.2267	1.9519	1.0976
$\frac{\text{MEDISE}(C_2)}{\text{MEDISE}(C_0)}$	81.8740	2.9874	1.0294	490.3515	8.8710	1.0445	79.6110	3.1894	1.0435	350.6805	7.5627	1.0876
$\frac{\text{MEDISE}(C_3)}{\text{MEDISE}(C_0)}$	2.0131	1.0458	1.0019	5.9736	1.1294	1.0016	2.1097	1.0655	1.0019	5.0231	1.0977	0.9964
$\frac{\text{MEDISE}(C_4)}{\text{MEDISE}(C_0)}$	9.5897	1.1565	0.9911	48.8814	1.3842	1.0081	9.4753	1.1843	0.9894	35.5014	1.3224	1.0083

Table 11: Ratio between the MEDISE of the estimators of the additive component $g_{0,1}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_2(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 + 0.2)$					
	$\widehat{g}_{2,\mathrm{BC},0}$	$\widehat{g}_{2,\mathrm{BR,H},0}$	$\widehat{g}_{2,\mathrm{BR,T},0}$	$\widehat{g}_{2,\mathrm{BC},1}$	$\widehat{g}_{2,\mathrm{BR,H},1}$	$\widehat{g}_{2,\mathrm{BR,T},1}$	$\widehat{g}_{2,\mathrm{BC},0}$	$\widehat{g}_{2,\mathrm{BR,H},0}$	$\widehat{g}_{2,\mathrm{BR,T},0}$	$\widehat{g}_{2,\mathrm{BC},1}$	$\widehat{g}_{2,\mathrm{BR,H},1}$	$\widehat{g}_{2,\mathrm{BR,T},1}$
$\frac{\text{MEDISE}(C_1)}{\text{MEDISE}(C_0)}$	8.3252	1.2142	1.0177	49.2713	1.8064	1.0717	10.4835	1.2699	1.0020	49.1453	1.8770	1.1082
$\frac{\text{MEDISE}(C_2)}{\text{MEDISE}(C_0)}$	83.1079	2.2814	1.0453	463.4817	8.0080	1.0595	92.6170	2.8382	1.0571	403.2129	8.6146	1.0931
$\frac{\text{MEDISE}(C_3)}{\text{MEDISE}(C_0)}$	1.6226	0.9934	0.9911	5.7296	1.0781	0.9914	1.7847	0.9856	1.0022	5.7381	1.0852	1.0044
$\frac{\text{MEDISE}(C_4)}{\text{MEDISE}(C_0)}$	7.8349	0.9617	0.9990	44.4295	1.1974	1.0120	8.6646	0.9798	0.9963	38.5909	1.1879	1.0214

Table 12: Ratio between the MEDISE of the estimators of the additive component $g_{0,2}$ under the considered contaminations and under clean data for the two missing mechanisms.

in particular when responses are missing and some neighbourhoods contain too few observations.

4.2 Monte Carlo study with $d = 4$ additive components

For this model we generated covariates $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}) \sim U([-3,3]^4)$, independent errors $\varepsilon_i \sim N(0,1)$ and $\sigma_0 = 0.15$. We used the following additive components: $g_{0,1}(x_1) = x_1^3/12$, $g_{0,2}(x_2) = \sin(-x_2)$, $g_{0,3}(x_3) = x_3^2/2 - 1.5$, $g_{0,4}(x_4) = e^{x_4}/4 - (e^3 - e^{-3})/24$. As in dimension $d = 2$, optimal bandwidths with respect to the MISE were computed assuming that the other components in the model are known, resulting in $\mathbf{h}_{\text{opt}}^{\text{MISE}} = (0.36, 0.38, 0.34, 0.29)$. However, it was difficult to obtain a reliable estimate for the residual scale σ_0 using these bandwidths (see Remark 3.1), since many 4-dimensional neighbourhoods did not contain sufficient observations. To solve this problem we used $\mathbf{h}_\sigma = (0.93, 0.93, 0.93, 0.93)$ to estimate σ_0 (we expect an average of 5 points in each 4-dimensional neighbourhood using $\mathbf{h}_\sigma$). We then applied the backfitting algorithm with the optimal bandwidths $\mathbf{h}_{\text{opt}}^{\text{MISE}}$.

In addition to the settings C_0 and C_1 described above, we modified the contamination setting C_2 so that $u_i \sim N(15, 0.01)$ for all i such that $\mathbf{X}_{i,j} \in [-1.5, 1.5]$ for all $1 \leq j \leq 4$.

Tables 13 to 17 report the MISE for the different estimators, contamination settings and missing mechanisms. Ratios of MISE's for clean and contaminated settings and median ISE's are reported in tables included Tables 18 to 32 .

As observed in Section 4.1, our experiments with and without missing responses yield similar conclusions regarding the advantage of the robust procedure over the classical backfitting and the benefits of the Tukey's loss function over the Huber one. As in dimension $d = 2$, when responses are missing, all the MISE's are slightly inflated, as it is to be expected. It is also not surprising that when the data do not contain outliers (C_0), the robust estimators have a slightly larger MISE than their classical counterparts. However, when outliers are present, both robust estimators provide a substantially better performance than the classical one, given similar results to those for clean data. Also as noted for the model with $d = 2$, the MISE of the estimators based on Tukey's bisquare score function are more stable across the different contamination settings than those using Huber's score function. However, one difference with the results in dimension $d = 2$ should be pointed out. Under the gross error model C_1, the local linear smoother performs worse than the local constant when missing responses arise. This difference is highlighted when one looks at the ratio of the corresponding MISE's. However, median ISE's reported in Tables 23 to 27 are more stable, which indicates that the difference in MISE's may be due to a few samples. In fact, due to the relatively large number of missing observations that may be present in this setting (around 33%) around 6% of the replications produce robust fits with Tukey's loss function with atypically large values of MISE, which are due to sparsely populated neighbourhoods. Based on these observations, we also recommend using Tukey's local linear estimators.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{\mathrm{BC},0}$	$\widehat{g}_{\mathrm{BR,H},0}$	$\widehat{g}_{\mathrm{BR,T},0}$	$\widehat{g}_{\mathrm{BC},1}$	$\widehat{g}_{\mathrm{BR,H},1}$	$\widehat{g}_{\mathrm{BR,T},1}$	$\widehat{g}_{\mathrm{BC},0}$	$\widehat{g}_{\mathrm{BR,H},0}$	$\widehat{g}_{\mathrm{BR,T},0}$	$\widehat{g}_{\mathrm{BC},1}$	$\widehat{g}_{\mathrm{BR,H},1}$	$\widehat{g}_{\mathrm{BR,T},1}$
C_0	0.0123	0.0125	0.0125	0.0023	0.0023	0.0023	0.0135	0.0142	0.0150	0.0033	0.0033	0.0033
C_1	7.3345	0.0525	0.0134	7.6095	0.0497	0.0046	8.4183	0.0699	0.0192	8.8581	0.1563	0.0932
C_2	4.8441	0.0360	0.0128	4.8224	0.0221	0.0025	5.9889	0.0392	0.0148	6.0121	0.0258	0.0038

Table 13: MISE of the estimators of the regression function $g_0 = \mu_0 + \sum_{j=1}^{4} g_{0,j}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{1,\mathrm{BC},0}$	$\widehat{g}_{1,\mathrm{BR,H},0}$	$\widehat{g}_{1,\mathrm{BR,T},0}$	$\widehat{g}_{1,\mathrm{BC},1}$	$\widehat{g}_{1,\mathrm{BR,H},1}$	$\widehat{g}_{1,\mathrm{BR,T},1}$	$\widehat{g}_{1,\mathrm{BC},0}$	$\widehat{g}_{1,\mathrm{BR,H},0}$	$\widehat{g}_{1,\mathrm{BR,T},0}$	$\widehat{g}_{1,\mathrm{BC},1}$	$\widehat{g}_{1,\mathrm{BR,H},1}$	$\widehat{g}_{1,\mathrm{BR,T},1}$
C_0	0.0046	0.0047	0.0047	0.0020	0.0020	0.0020	0.0059	0.0060	0.0061	0.0030	0.0030	0.0030
C_1	0.5547	0.0085	0.0049	0.6356	0.0066	0.0021	0.8355	0.0107	0.0063	0.9703	0.0250	0.0178
C_2	0.9691	0.0089	0.0047	0.9897	0.0060	0.0020	1.1446	0.0105	0.0062	1.1960	0.0073	0.0030

Table 14: MISE of the estimators of the additive component $g_{0,1}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{2,\mathrm{BC},0}$	$\widehat{g}_{2,\mathrm{BR,H},0}$	$\widehat{g}_{2,\mathrm{BR,T},0}$	$\widehat{g}_{2,\mathrm{BC},1}$	$\widehat{g}_{2,\mathrm{BR,H},1}$	$\widehat{g}_{2,\mathrm{BR,T},1}$	$\widehat{g}_{2,\mathrm{BC},0}$	$\widehat{g}_{2,\mathrm{BR,H},0}$	$\widehat{g}_{2,\mathrm{BR,T},0}$	$\widehat{g}_{2,\mathrm{BC},1}$	$\widehat{g}_{2,\mathrm{BR,H},1}$	$\widehat{g}_{2,\mathrm{BR,T},1}$
C_0	0.0025	0.0025	0.0025	0.0016	0.0016	0.0016	0.0035	0.0035	0.0036	0.0024	0.0024	0.0024
C_1	0.5332	0.0062	0.0027	0.6189	0.0081	0.0026	0.7802	0.0101	0.0038	0.8982	0.0226	0.0136
C_2	0.9298	0.0066	0.0026	0.9482	0.0055	0.0016	1.1950	0.0083	0.0037	1.2339	0.0067	0.0024

Table 15: MISE of the estimators of the additive component $g_{0,2}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{3,\mathrm{BC},0}$	$\widehat{g}_{3,\mathrm{BR,H},0}$	$\widehat{g}_{3,\mathrm{BR,T},0}$	$\widehat{g}_{3,\mathrm{BC},1}$	$\widehat{g}_{3,\mathrm{BR,H},1}$	$\widehat{g}_{3,\mathrm{BR,T},1}$	$\widehat{g}_{3,\mathrm{BC},0}$	$\widehat{g}_{3,\mathrm{BR,H},0}$	$\widehat{g}_{3,\mathrm{BR,T},0}$	$\widehat{g}_{3,\mathrm{BC},1}$	$\widehat{g}_{3,\mathrm{BR,H},1}$	$\widehat{g}_{3,\mathrm{BR,T},1}$
C_0	0.0091	0.0092	0.0092	0.0042	0.0042	0.0042	0.0136	0.0141	0.0146	0.0082	0.0082	0.0082
C_1	0.5991	0.0137	0.0095	0.6741	0.0106	0.0052	0.9195	0.0273	0.0157	1.0679	0.0449	0.0337
C_2	1.0137	0.0155	0.0094	1.0007	0.0085	0.0042	1.2200	0.0207	0.0144	1.2256	0.0126	0.0082

Table 16: MISE of the estimators of the additive component $g_{0,3}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{4,\mathrm{BC},0}$	$\widehat{g}_{4,\mathrm{BR,H},0}$	$\widehat{g}_{4,\mathrm{BR,T},0}$	$\widehat{g}_{4,\mathrm{BC},1}$	$\widehat{g}_{4,\mathrm{BR,H},1}$	$\widehat{g}_{4,\mathrm{BR,T},1}$	$\widehat{g}_{4,\mathrm{BC},0}$	$\widehat{g}_{4,\mathrm{BR,H},0}$	$\widehat{g}_{4,\mathrm{BR,T},0}$	$\widehat{g}_{4,\mathrm{BC},1}$	$\widehat{g}_{4,\mathrm{BR,H},1}$	$\widehat{g}_{4,\mathrm{BR,T},1}$
C_0	0.0070	0.0070	0.0071	0.0036	0.0036	0.0036	0.0095	0.0098	0.0103	0.0058	0.0058	0.0058
C_1	0.6818	0.0118	0.0072	0.7558	0.0097	0.0037	1.0890	0.0235	0.0127	1.2310	0.0583	0.0429
C_2	1.0582	0.0125	0.0071	1.0592	0.0078	0.0036	1.3678	0.0159	0.0101	1.3877	0.0117	0.0061

Table 17: MISE of the estimators of the additive component $g_{0,4}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	594.6856	4.1927	1.0660	3263.3298	21.3232	1.9603	622.6124	4.9188	1.2865	2677.7184	47.1836	28.0937
$\text{MISE}(C_2)/\text{MISE}(C_0)$	392.7639	2.8789	1.0204	2068.0873	9.4793	1.0625	442.9360	2.7631	0.9876	1817.4052	7.8002	1.1449

Table 18: Ratio between the MISE of the estimators of the additive component $g_0 = \mu_0 + \sum_{j=1}^{4} g_{0,j}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	119.9119	1.8225	1.0448	318.2129	3.3018	1.0502	142.1759	1.7651	1.0330	327.3855	8.4212	5.9965
$\text{MISE}(C_2)/\text{MISE}(C_0)$	209.4957	1.9119	1.0185	495.5000	3.0240	1.0185	194.7711	1.7403	1.0134	403.5116	2.4681	1.0207

Table 19: Ratio between the MISE of the estimators of the additive component $g_{0,1}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	213.8436	2.4642	1.0692	385.1917	5.0106	1.6478	225.1335	2.8729	1.0599	379.1973	9.5277	5.7400
$\text{MISE}(C_2)/\text{MISE}(C_0)$	372.9278	2.6408	1.0309	590.1075	3.4041	1.0208	344.8200	2.3533	1.0324	520.8911	2.8161	1.0259

Table 20: Ratio between the MISE of the estimators of the additive component $g_{0,2}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{3,\text{BC},0}$	$\widehat{g}_{3,\text{BR,H},0}$	$\widehat{g}_{3,\text{BR,T},0}$	$\widehat{g}_{3,\text{BC},1}$	$\widehat{g}_{3,\text{BR,H},1}$	$\widehat{g}_{3,\text{BR,T},1}$	$\widehat{g}_{3,\text{BC},0}$	$\widehat{g}_{3,\text{BR,H},0}$	$\widehat{g}_{3,\text{BR,T},0}$	$\widehat{g}_{3,\text{BC},1}$	$\widehat{g}_{3,\text{BR,H},1}$	$\widehat{g}_{3,\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	66.0132	1.4880	1.0318	160.3085	2.5264	1.2368	67.6382	1.9308	1.0775	130.9593	5.5051	4.1322
$\text{MISE}(C_2)/\text{MISE}(C_0)$	111.7021	1.6804	1.0150	237.9769	2.0109	1.0087	89.7438	1.4627	0.9885	150.2898	1.5440	1.0080

Table 21: Ratio between the MISE of the estimators of the additive component $g_{0,3}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{4,\text{BC},0}$	$\widehat{g}_{4,\text{BR,H},0}$	$\widehat{g}_{4,\text{BR,T},0}$	$\widehat{g}_{4,\text{BC},1}$	$\widehat{g}_{4,\text{BR,H},1}$	$\widehat{g}_{4,\text{BR,T},1}$	$\widehat{g}_{4,\text{BC},0}$	$\widehat{g}_{4,\text{BR,H},0}$	$\widehat{g}_{4,\text{BR,T},0}$	$\widehat{g}_{4,\text{BC},1}$	$\widehat{g}_{4,\text{BR,H},1}$	$\widehat{g}_{4,\text{BR,T},1}$
$\text{MISE}(C_1)/\text{MISE}(C_0)$	97.8712	1.6842	1.0244	211.8380	2.7222	1.0340	114.1378	2.4048	1.2307	212.2685	10.0466	7.3858
$\text{MISE}(C_2)/\text{MISE}(C_0)$	151.9140	1.7760	1.0115	296.8837	2.1873	1.0131	143.3664	1.6293	0.9839	239.2791	2.0083	1.0541

Table 22: Ratio between the MISE of the estimators of the additive component $g_{0,4}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$
C_0	0.0123	0.0125	0.0125	0.0023	0.0023	0.0023	0.0135	0.0142	0.0148	0.0033	0.0033	0.0033
C_1	7.2318	0.0497	0.0133	7.5283	0.0421	0.0027	8.3171	0.0481	0.0155	8.7854	0.0435	0.0040
C_2	4.7574	0.0344	0.0128	4.7560	0.0203	0.0024	5.9049	0.0365	0.0147	5.9358	0.0215	0.0035

Table 23: Medise of the estimators of the regression function g_0 under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$
C_0	0.0040	0.0040	0.0040	0.0013	0.0013	0.0013	0.0050	0.0051	0.0052	0.0019	0.0019	0.0019
C_1	0.5139	0.0077	0.0042	0.5909	0.0059	0.0014	0.7747	0.0099	0.0055	0.9060	0.0075	0.0021
C_2	0.9560	0.0084	0.0041	0.9677	0.0056	0.0013	1.1128	0.0096	0.0053	1.1591	0.0061	0.0020

Table 24: Medise of the estimators of the additive component $g_{0,1}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$
C_0	0.0020	0.0020	0.0020	0.0011	0.0011	0.0011	0.0028	0.0028	0.0029	0.0016	0.0016	0.0016
C_1	0.4980	0.0056	0.0022	0.5879	0.0054	0.0012	0.7289	0.0071	0.0031	0.8413	0.0064	0.0018
C_2	0.9108	0.0062	0.0021	0.9260	0.0050	0.0011	1.1789	0.0077	0.0031	1.2041	0.0059	0.0016

Table 25: Medise of the estimators of the additive component $g_{0,2}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{3,\text{BC},0}$	$\widehat{g}_{3,\text{BR,H},0}$	$\widehat{g}_{3,\text{BR,T},0}$	$\widehat{g}_{3,\text{BC},1}$	$\widehat{g}_{3,\text{BR,H},1}$	$\widehat{g}_{3,\text{BR,T},1}$	$\widehat{g}_{3,\text{BC},0}$	$\widehat{g}_{3,\text{BR,H},0}$	$\widehat{g}_{3,\text{BR,T},0}$	$\widehat{g}_{3,\text{BC},1}$	$\widehat{g}_{3,\text{BR,H},1}$	$\widehat{g}_{3,\text{BR,T},1}$
C_0	0.0076	0.0077	0.0077	0.0023	0.0023	0.0023	0.0102	0.0108	0.0111	0.0044	0.0044	0.0044
C_1	0.5686	0.0124	0.0080	0.6297	0.0075	0.0024	0.8675	0.0167	0.0110	1.0010	0.0115	0.0046
C_2	1.0046	0.0141	0.0079	0.9907	0.0069	0.0023	1.1809	0.0177	0.0110	1.1953	0.0096	0.0045

Table 26: Medise of the estimators of the additive component $g_{0,3}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{4,\text{BC},0}$	$\widehat{g}_{4,\text{BR,H},0}$	$\widehat{g}_{4,\text{BR,T},0}$	$\widehat{g}_{4,\text{BC},1}$	$\widehat{g}_{4,\text{BR,H},1}$	$\widehat{g}_{4,\text{BR,T},1}$	$\widehat{g}_{4,\text{BC},0}$	$\widehat{g}_{4,\text{BR,H},0}$	$\widehat{g}_{4,\text{BR,T},0}$	$\widehat{g}_{4,\text{BC},1}$	$\widehat{g}_{4,\text{BR,H},1}$	$\widehat{g}_{4,\text{BR,T},1}$
C_0	0.0056	0.0057	0.0058	0.0022	0.0022	0.0022	0.0073	0.0076	0.0080	0.0034	0.0034	0.0034
C_1	0.6510	0.0106	0.0061	0.7153	0.0078	0.0023	1.0298	0.0141	0.0080	1.1628	0.0112	0.0037
C_2	1.0484	0.0115	0.0059	1.0447	0.0066	0.0022	1.3321	0.0140	0.0079	1.3391	0.0084	0.0035

Table 27: Medise of the estimators of the additive component $g_{0,4}$ under different contaminations and missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$	$\widehat{g}_{\text{BC},0}$	$\widehat{g}_{\text{BR,H},0}$	$\widehat{g}_{\text{BR,T},0}$	$\widehat{g}_{\text{BC},1}$	$\widehat{g}_{\text{BR,H},1}$	$\widehat{g}_{\text{BR,T},1}$
$\frac{\textsc{Medise}(C_1)}{\textsc{Medise}(C_0)}$	587.9280	3.9901	1.0632	3284.7120	18.3599	1.1840	616.5264	3.3837	1.0457	2694.8344	13.3422	1.2335
$\frac{\textsc{Medise}(C_2)}{\textsc{Medise}(C_0)}$	386.7698	2.7623	1.0253	2075.0990	8.8790	1.0682	437.7162	2.5680	0.9951	1820.7418	6.6007	1.0631

Table 28: Ratio between the Medise of the estimators of the additive component g_0 under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$	$\widehat{g}_{1,\text{BC},0}$	$\widehat{g}_{1,\text{BR,H},0}$	$\widehat{g}_{1,\text{BR,T},0}$	$\widehat{g}_{1,\text{BC},1}$	$\widehat{g}_{1,\text{BR,H},1}$	$\widehat{g}_{1,\text{BR,T},1}$
$\frac{\textsc{Medise}(C_1)}{\textsc{Medise}(C_0)}$	129.7275	1.9127	1.0513	460.2723	4.5320	1.0668	156.4803	1.9257	1.0649	466.3972	3.8856	1.0932
$\frac{\textsc{Medise}(C_2)}{\textsc{Medise}(C_0)}$	241.3087	2.0909	1.0125	753.7706	4.3051	1.0116	224.7727	1.8676	1.0267	596.6858	3.1300	1.0223

Table 29: Ratio between the Medise of the estimators of the additive component $g_{0,1}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$	$\widehat{g}_{2,\text{BC},0}$	$\widehat{g}_{2,\text{BR,H},0}$	$\widehat{g}_{2,\text{BR,T},0}$	$\widehat{g}_{2,\text{BC},1}$	$\widehat{g}_{2,\text{BR,H},1}$	$\widehat{g}_{2,\text{BR,T},1}$
$\frac{\textsc{Medise}(C_1)}{\textsc{Medise}(C_0)}$	253.9908	2.8292	1.0822	557.0175	5.1005	1.0940	258.5182	2.4780	1.0665	524.3816	4.0351	1.1120
$\frac{\textsc{Medise}(C_2)}{\textsc{Medise}(C_0)}$	464.5554	3.1424	1.0488	877.3765	4.7631	1.0244	418.1190	2.7026	1.0618	750.5317	3.6802	1.0270

Table 30: Ratio between the Medise of the estimators of the additive component $g_{0,2}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{3,\text{BC},0}$	$\widehat{g}_{3,\text{BR,H},0}$	$\widehat{g}_{3,\text{BR,T},0}$	$\widehat{g}_{3,\text{BC},1}$	$\widehat{g}_{3,\text{BR,H},1}$	$\widehat{g}_{3,\text{BR,T},1}$	$\widehat{g}_{3,\text{BC},0}$	$\widehat{g}_{3,\text{BR,H},0}$	$\widehat{g}_{3,\text{BR,T},0}$	$\widehat{g}_{3,\text{BC},1}$	$\widehat{g}_{3,\text{BR,H},1}$	$\widehat{g}_{3,\text{BR,T},1}$
$\frac{\textsc{Medise}(C_1)}{\textsc{Medise}(C_0)}$	75.1423	1.6025	1.0372	273.7164	3.2475	1.0414	85.4013	1.5531	0.9944	226.6848	2.6012	1.0510
$\frac{\textsc{Medise}(C_2)}{\textsc{Medise}(C_0)})$	132.7657	1.8189	1.0282	430.6451	3.0177	1.0127	116.2487	1.6494	0.9913	270.6875	2.1641	1.0084

Table 31: Ratio between the Medise of the estimators of the additive component $g_{0,3}$ under the considered contaminations and under clean data for the two missing mechanisms.

	$p(\mathbf{x}) \equiv 1$						$p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1\, x_3 + 0.2)$					
	$\widehat{g}_{4,\text{BC},0}$	$\widehat{g}_{4,\text{BR,H},0}$	$\widehat{g}_{4,\text{BR,T},0}$	$\widehat{g}_{4,\text{BC},1}$	$\widehat{g}_{4,\text{BR,H},1}$	$\widehat{g}_{4,\text{BR,T},1}$	$\widehat{g}_{4,\text{BC},0}$	$\widehat{g}_{4,\text{BR,H},0}$	$\widehat{g}_{4,\text{BR,T},0}$	$\widehat{g}_{4,\text{BC},1}$	$\widehat{g}_{4,\text{BR,H},1}$	$\widehat{g}_{4,\text{BR,T},1}$
$\frac{\textsc{Medise}(C_1)}{\textsc{Medise}(C_0)}$	115.3211	1.8521	1.0528	331.9275	3.6263	1.0442	140.1222	1.8468	0.9912	344.1223	3.3286	1.0869
$\frac{\textsc{Medise}(C_2)}{\textsc{Medise}(C_0)}$	185.7104	2.0115	1.0223	484.7585	3.0544	1.0200	181.2518	1.8305	0.9832	396.2957	2.4830	1.0267

Table 32: Ratio between the Medise of the estimators of the additive component $g_{0,2}$ under the considered contaminations and under clean data for the two missing mechanisms.

5 Empirical Influence

A well-known measure of robustness of an estimator is given by its influence function (see Hampel *et al.* 1986). The influence function measures resistance of an estimator against infinitesimal proportions of outliers and helps study the local robustness and asymptotic efficiency of an estimator. The finite-sample version of the influence function, called the empirical influence function (Tukey, 1977), is a useful measure of sensitivity quantifying the effect of a single outlier on the estimator computed on a given sample. Although influence functions have been widely studied for many parametric models, much less attention has been paid to nonparametric estimators. To measure the influence of a contaminating point on the estimators, we follow the approach of Manchester (1996), who proposed a graphical method to display the sensitivity of a scatter plot smoother that is related to the finite–sample influence function introduced by Tukey (1977).

Given a data set $\{(\mathbf{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}\}_{1\le i\le n}$ satisfying the additive model $Y = \mu_0 + \sum_{j=1}^{d} g_{0,j}(X_j) + \sigma_0\,\varepsilon$, let $\widehat{g}_{n,j}(\tau)$ be the estimator of the $j-$th component based on this data set evaluated at the point $\tau \in \mathbb{R}$. Assume that $\mathbf{z}_0 = (\mathbf{x}_0^{\mathrm{T}}, y_0)^{\mathrm{T}}$ represents a contaminating point and let $\widehat{g}_{n,j}^{(\mathbf{z}_0)}(\tau)$ be the estimator based on the augmented data set $\{(\mathbf{X}_1^{\mathrm{T}}, Y_1)^{\mathrm{T}}, \dots (\mathbf{X}_n^{\mathrm{T}}, Y_n)^{\mathrm{T}}, \mathbf{z}_0\}$ evaluated at the point τ. For a fixed value of τ, we define the empirical influence function of $\widehat{g}_{n,j}(\tau)$ at $\mathbf{z}_0$ as the surface

$$\mathrm{EIF}_{j,\tau}(\mathbf{z}_0) \;=\; (n+1)\,\left[\widehat{g}_{n,j}^{(\mathbf{z}_0)}(\tau) - \widehat{g}_{n,j}(\tau)\right]\,, \tag{11}$$

as $\mathbf{z}_0$ varies in $\mathbb{R}^d \times \mathbb{R}$. To explore the sensitivity of the backfitting estimators to the presence of outliers using the empirical influence function (11), we generated a data set of size $n = 500$ following an additive model with location $\mu_0 = 0$, additive components $g_{0,1}(x_1) = 24\,(x_1 - 0.5)^2 - 2$ and $g_{0,2}(x_2) = 2\pi \sin(\pi x_2) - 4$ and covariates $\mathbf{X}_i = (X_{i,1}, X_{i,2})^{\mathrm{T}} \sim U([0,1]\times[0,1])$. The data and the regression function are shown in Figure 2.

We used an Epanechnikov kernel with bandwidths $h_1 = h_2 = 0.10$, local constant smoothers ($q = 0$) and the same tuning constants as in our simulation study. We computed $\mathrm{EIF}_{j,\tau}(\mathbf{z}_0)$ for $\tau = 0.20, 0.40, 0.60$ and 0.80 and a grid of points $\mathbf{z}_0 = ((x_1, 0.5)^{\mathrm{T}}, y)^{\mathrm{T}}$, where x_1 ranges over 30 equidistant points in the interval $[0.15, 0.85]$ and y takes 50 equally spaced points in $[-20, 20]$.

The results for each estimator and for $\tau = 0.2$ and 0.4 are displayed in Figure 3, while the results for $\tau = 0.6$ and 0.8 are given in Figure 4.

These plots illustrate the expected lack of robustness of the classical backfitting estimator, for which the empirical influence function takes very large values. Note the EIF attain the largest absolute value when x_1 is close to τ, and estimators based on Tukey's bisquare loss function have a slightly larger $|\,\mathrm{EIF}\,|$ than those based on Huber's loss. The redescending structure of the score function can also be observed in the plot, showing that very large values of the responses have less effect on the estimator based on the Tukey loss function than in that based on the Huber loss, as noted also in the simulation study. It is important to

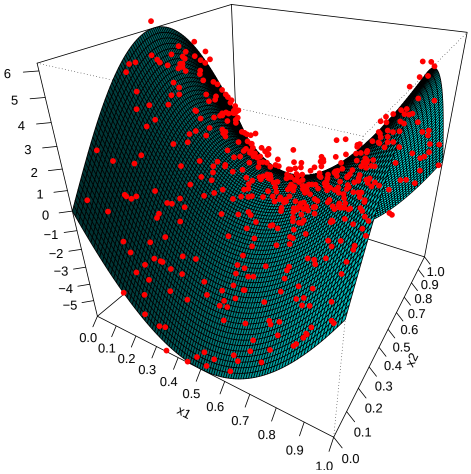
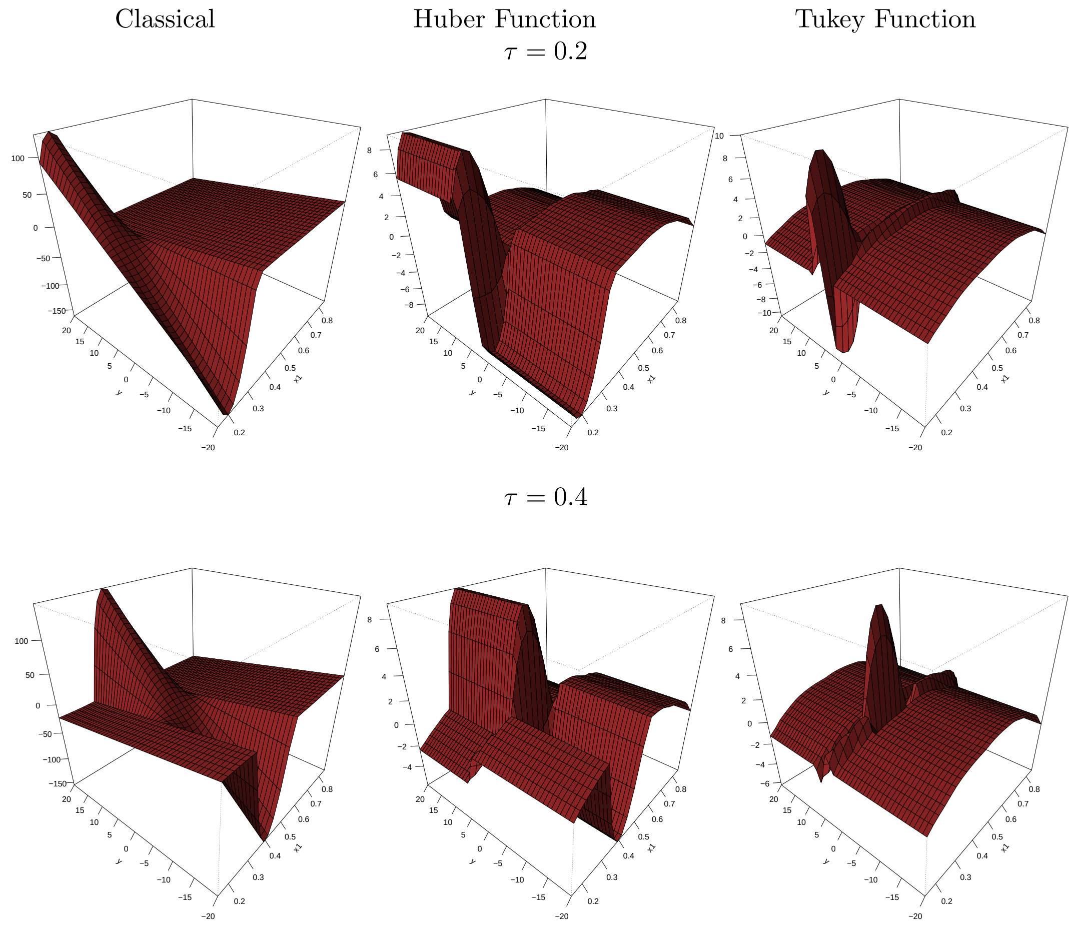


Figure 2: Data used for the influence function study, and the corresponding regression function g_0.

note that, when the nonparametric regression model does not take into account an additive structure and when using a kernel with compact support to compute a kernel regression estimator only outliers near the value at which the regression function estimator is evaluated may impact the regression estimator. However, the situation is different for the backfitting method, which involves the estimation of the location parameter and an iterative algorithm involving all the residuals.

Since the absolute value of $\text{EIF}_{1,\tau}(\mathbf{x}, y)$ attains its maximum value near τ, Figure 5 shows the surfaces $\text{EIF}_{1,x_1}((x_1, 0.5), y)$, which represent the worst possible bias of these estimators in this setting. The plots of $|\text{EIF}_{1,x_1}((x_1, 0.5), y)|$ are given in Figure 6. As expected, the bias of the classical estimators follows the size of the contaminated responses. On the other hand, the empirical functions of the robust estimators are bounded, and the most influential points correspond to x_1 near 0.2 and 0.8, which reflects the expected boundary effect. Due to the redescending nature of the Tukey score function, the absolute value of the empirical function for larger values of y ($|y| > 5$, say) remains very low, near its minimum absolute value of 0.019.



Classical

Huber Function

Tukey Function

$\tau = 0.2$

$\tau = 0.4$

Figure 3: Empirical influence for the classical and robust estimators, $\text{EIF}_{1,\tau}(\mathbf{x}, y)$ when $\tau = 0.2$ and 0.4 and $\mathbf{x} = (x_1, 0.5)$.

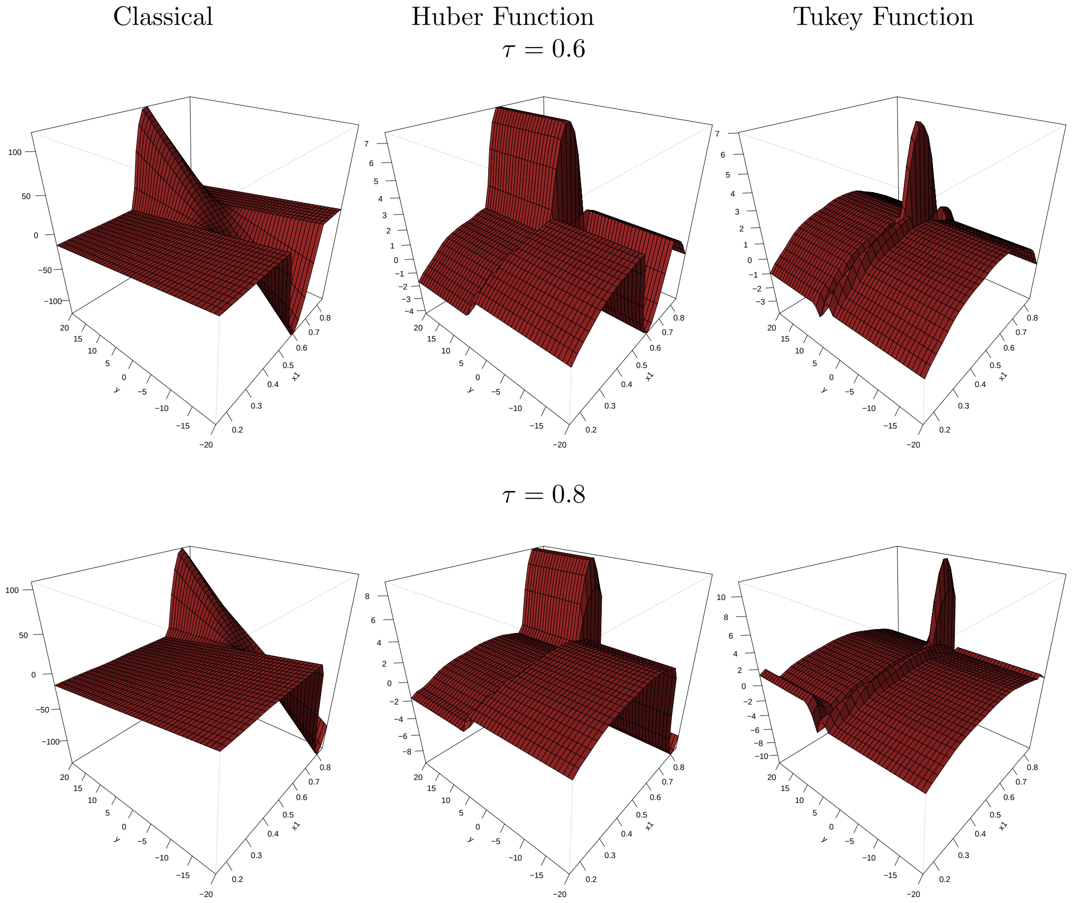
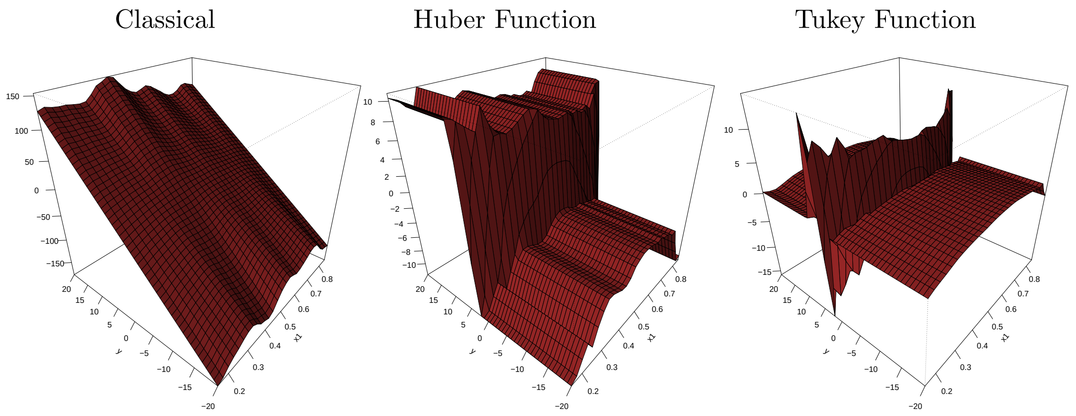


Figure 4: Empirical influence for the classical and robust estimators, $\mathrm{EIF}_{1,\tau}(\mathbf{x}, y)$ when $\tau = 0.6$ and 0.8 and $\mathbf{x} = (x_1, 0.5)$

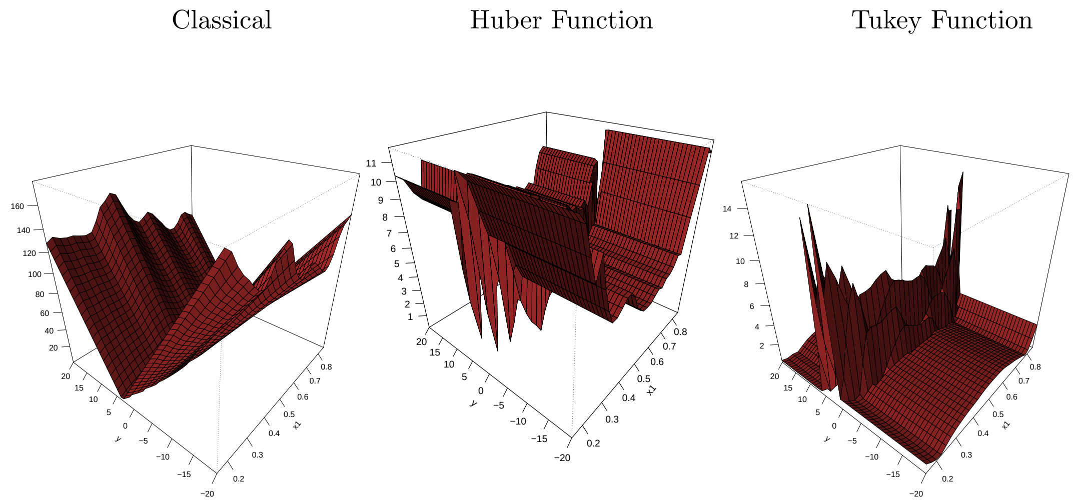


Classical

Huber Function

Tukey Function

Figure 5: Empirical influence $\mathrm{EIF}_{1,x_1}((x_1, 0.5), y)$ for the classical and robust estimators.



Classical

Huber Function

Tukey Function

Figure 6: Absolute value of the empirical influence, $|\mathrm{EIF}_{1,x_1}((x_1, 0.5), y)|$ for the classical and robust estimators.

6 Real data example

In this section, we compare the performance of the robust backfitting described in this paper with the classical one on a real data set. We considered the `airquality` data set available in `R`. The data set corresponds to 153 daily air quality measurements in the New York region between May and September, 1973 (see Chambers *et al.*, 1983). The interest is in explaining mean Ozone concentration ("O_3", measured in ppb) as a function of 3 potential explanatory variables: temperature ("Temp", in degrees Fahrenheit), wind speed ("Wind", in mph) and solar radiance measured in the frequency band 4000-7700 ("Solar.R", in Langleys). In our analysis, we only considered the 111 cases that do not contain missing observations. Dengyi and Kawagochi (1986) and Lacour *et al.* (2006) report a positive correlation between ozone concentration and temperature in the Antarctica during Spring and also, in France during the 2003 heat wave. Cleveland (1985) finds that the relationship between ozone concentration and wind speed is non-linear, with higher wind speeds associated to lower Ozone concentrations. Simple visual exploration of the data indicates that the relationship between ozone and the other variables does not appear to be linear, so we propose to fit an additive model of the form

$$O_3 = \mu_0 + g_{0,1}(\text{Temp}) + g_{0,2}(\text{Wind}) + g_{0,3}(\text{Solar.R}) + u\,,$$

where the errors $u = \sigma_0\,\varepsilon$ are assumed to be independent, homoscedastic and with location parameter 0.

Based on the results obtained in Section 4, we used local linear backfitting estimators with the classical squared loss function and also with Tukey's bisquare loss (with tuning constant $c = 4.685$) to provide a robust alternative. Bandwidths were selected using a 3-dimensional grid search. For the bandwidth h_j of the j-th covariate, $1 \le j \le 3$, we considered 6 possible values (equal to multiples of its estimated standard deviation): $G_j = \{\hat{\sigma}_j/2, \hat{\sigma}_j, 1.5\,\hat{\sigma}_j, 2\,\hat{\sigma}_j, 2.5\,\hat{\sigma}_j, 3\,\hat{\sigma}_j\}$, where $\hat{\sigma}_j = \text{sd}(X_j)$. Our 3-dimensional grid is the product of these sets: $\mathcal{G} = G_1 \times G_2 \times G_3 \subset \mathbb{R}^3$. Let $(\mathbf{X}_1^{\mathrm{T}}, Y_1)^{\mathrm{T}}, \dots, (\mathbf{X}_n^{\mathrm{T}}, Y_n)^{\mathrm{T}}$ be the considered observations ($n = 111$). The usual leave-one-out cross-validation criterion in this setting is given by $L_{\text{LS}}(\mathbf{h}) = (1/n)\sum_{i=1}^n \left(Y_i - \widehat{g}_{\text{BC},\mathbf{h}}^{-i}(\mathbf{X}_i)\right)^2$, where $\widehat{g}_{\text{BC},\mathbf{h}}^{-i}(\mathbf{X}_i)$ denotes the backfitting predictor for $\mathbf{X}_i$, computed with bandwidth $\mathbf{h} \in \mathcal{G}$ and without using the i-th observation. For the classical backfitting estimator the smallest value of L_{LS} over the grid $\mathcal{G}$ was obtained at $\mathbf{h}_{\text{LS}} = (9.53, 10.67, 91.15)$.

When outliers may be present in the data, it is important to use a robust selection criterion for smoothing parameters, even when considering robust estimators, see, for instance, Cantoni and Ronchetti (2001) for a discussion. Noting that the classical cross–validation criterion combines the squared bias and variance, Bianco and Boente (2007) and Boente and Rodriguez (2008) introduced the following robust cross-validation criterion based on robust estimators of bias and scale. Let $\widehat{g}_{\text{BR,T},\mathbf{h}}^{-i}(\mathbf{X}_i)$ denote the robust backfitting predictor at $\mathbf{X}_i$, computed with the smoothing parameter $\mathbf{h} \in \mathcal{G}$ and without using the i-th observation. The robust cross-validation criterion is:

$$L_{\mathrm{R}}(\mathbf{h}) \;=\; \left(\operatorname*{median}_{1\le i\le n}\{Y_i - \widehat{g}^{\,-i}_{\mathrm{BR,T},\mathbf{h}}(\mathbf{X}_i)\}\right)^2 \;+\; \left(\operatorname*{MAD}_{1\le i\le n}\{Y_i - \widehat{g}^{\,-i}_{\mathrm{BR,T},\mathbf{h}}(\mathbf{X}_i)\}\right)^2 \;.$$

The minimum of L_{R} over $\mathcal{G}$ was obtained at $\mathbf{h}_{\mathrm{R}} = (4.76, 8.89, 136.73)$, which leads to a smaller bandwidth for the first additive component and a larger one for the third than the ones chosen with the classical approach. This suggests that some influential observations may be present, which lead to oversmoothing of the classical estimator of the first additive component.

Figure 7 shows the partial residuals and the estimated regression components for each explanatory variable, both for the classical and robust estimators. Although the shape of the estimated additive components are similar, some important differences in their pattern can be highlighted. On the one hand, the classical estimator appears to magnify the effect of the covariates on the additive components of the regression function. With the classical estimator increasing temperatures correspond to a higher mean ozone concentration, but only for temperatures between 70 and 90 degrees (F). Higher temperatures correspond to lower mean ozone concentrations, and the same happens for increasing wind speeds and low values of solar radiance. At the same time, low wind speeds and solar radiance values between 150 and 250 correspond to higher mean levels of ozone. Intriguingly, lower temperatures are seen to result in a slight increase in mean ozone concentration. On the other hand, the robust estimator suggests covariate effects that are more moderate. For example, in the case of temperature, we note that the corresponding additive component is practically constant for temperatures up to 75 degrees, and for temperatures beyond 90 degrees does not decrease as markedly as the classical one.

We can use the residuals obtained with the robust fit to explore the presence of potential outliers in the data. Figure 8 shows the corresponding residual boxplot which indicates 5 clear outliers (observations 23, 34, 53, 68 and 77). To study the influence of these observations on the classical fit we repeat the analysis without them. The obtained cross–validation bandwidths for the classical estimator are now $\mathbf{h}_{\mathrm{LS}}^{(-5)} = (4.85, 10.52, 138.87)$. Note that these values are very similar to those obtained with the robust estimator combined with the robust cross–validation criterion. Figure 9 shows the estimates, $\widehat{g}_j^{(-5)}$, $j = 1, \ldots, 3$, obtained with the classical estimator using the "cleaned" data together with the robust ones obtained with the original data set. We see that both sets of fits are now very similar. In other words, the robust fit automatically down-weighted potential outliers and returned estimated additive components based on the remaining observations that are almost identical to the classical ones when the outliers are removed by hand. Furthermore, the residuals obtained from the robust fit allow us to identify potential outliers.

Acknowledgements. This research was partially supported by Grants PIP 112-201101-00339 from CONICET, PICT 0397 from ANPCYT and 20120130100241BA from the Universidad de Buenos Aires at Buenos Aires, Argentina (G. Boente and A. Martínez) and Discovery Grant 250048-11 of the Natural Sciences and Engineering Research Council of Canada (M. Salibián Barrera). This research

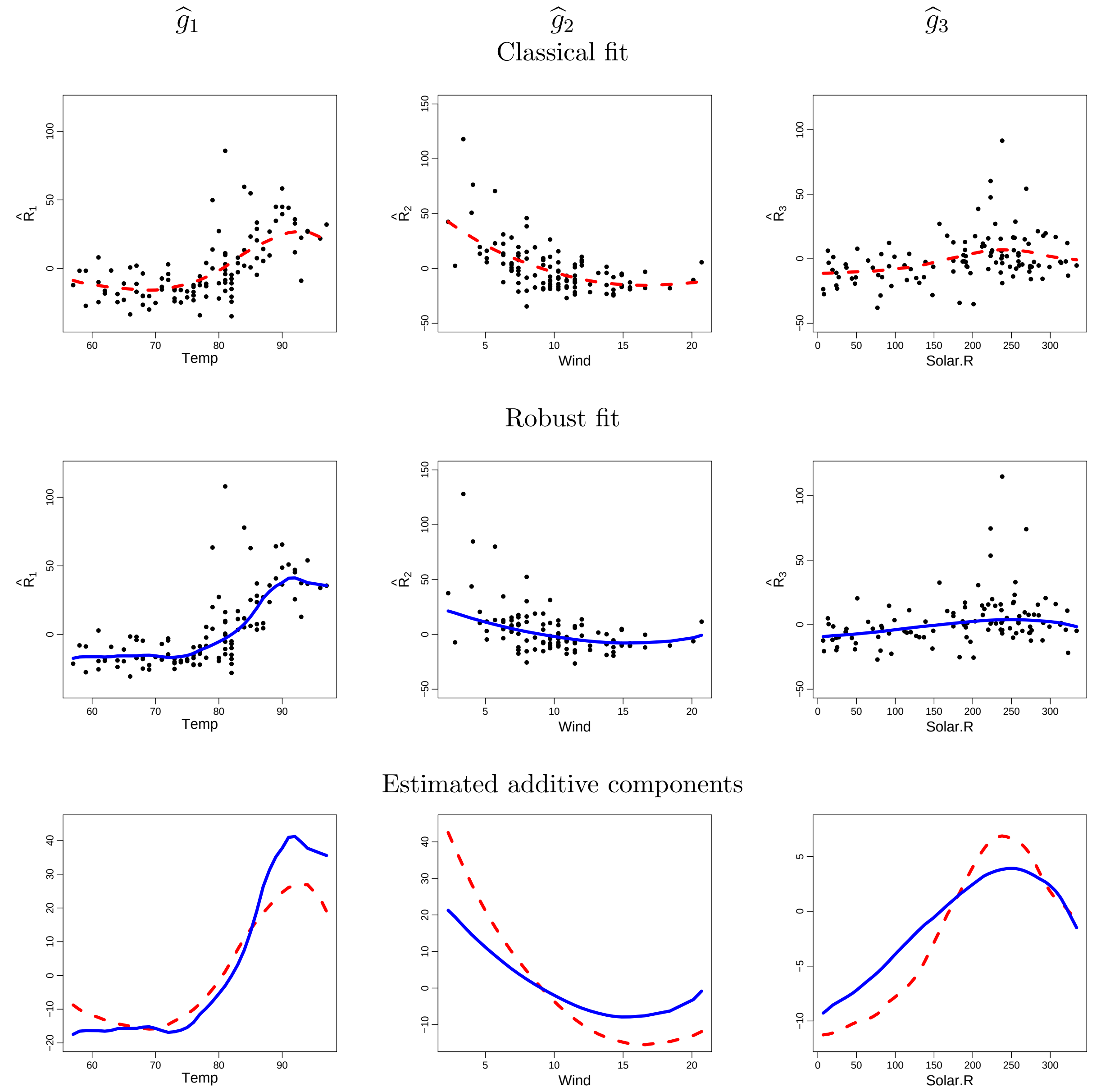


Figure 7: Partial residuals, $\widehat{R}_j$ for $1 \leq j \leq 3$, and estimated curves for the classical (in red dashed lines) and robust (in blue solid lines) backfitting estimators with data-driven bandwidths $\mathbf{h}_{\text{LS}}$ and $\mathbf{h}_{\text{R}}$, respectively.

was begun while Alejandra Martínez was visiting the University of British Columbia supported by the *Emerging Leaders in the Americas Program* from the Canadian Bureau for International Education and Foreign Affairs and International Trade Canada.

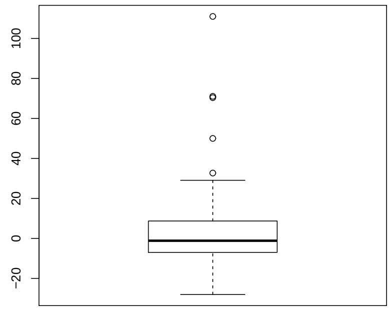


Figure 8: Boxplot of the residuals obtained using the robust fit with data-driven bandwidth $\mathbf{h}_{\mathrm{R}}$.

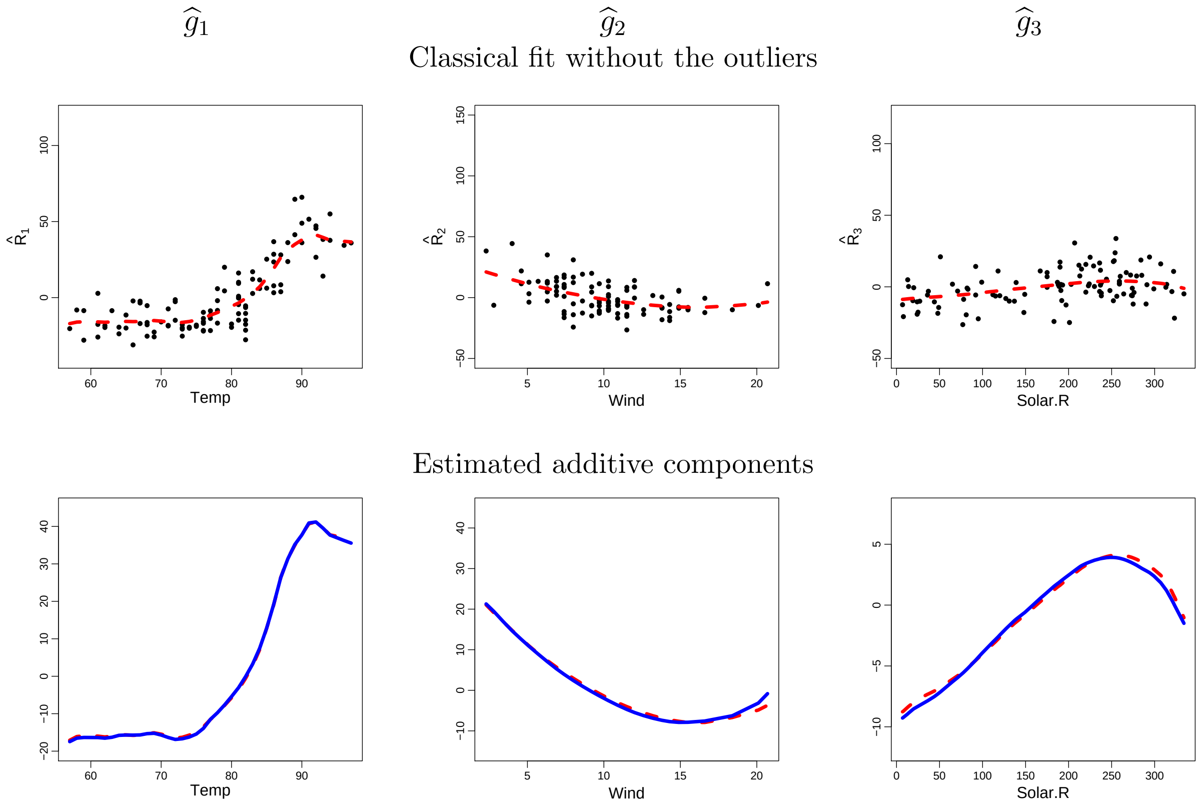


Figure 9: The upper plots show the partial residuals and estimated curves for the classical backfitting estimator, $\widehat{g}_j^{(-5)}$, (in red dashed lines) with data-driven bandwidth $\mathbf{h}_{\mathrm{LS}}^{(-5)}$. The lower plots correspond to the estimated curves for the classical backfitting estimator, $\widehat{g}_j^{(-5)}$ (in red dashed lines) with data-driven bandwidth $\mathbf{h}_{\mathrm{LS}}^{(-5)}$ and for the robust ones (in blue solid lines) computed with all the data and with data-driven bandwidth $\mathbf{h}_{\mathrm{R}}$.

A Appendix: Proofs

Proof of Theorem 2.1. (a) We will show that if $(\nu, m) \in \mathbb{R} \times \mathcal{H}^{ad}$ is such that either $\nu \neq \mu_0$ or $\mathbb{P}(\sum_{j=1}^d m_j(X_j) = \sum_{j=1}^d g_{0,j}(X_j)) < 1$ then $\Upsilon(\nu, m) > \Upsilon(\mu_0, g_0)$. For any $(\nu, m) \in \mathbb{R} \times \mathcal{H}^{ad}$ we have

$$\Upsilon(\nu, m) = \mathbb{E}\rho\left(\frac{Y - \nu - \sum_{j=1}^d m_j(X_j)}{\sigma_0}\right) = \mathbb{E}_{\mathbf{X}}\left(\mathbb{E}_{\varepsilon|\mathbf{X}}\left\{\rho\left(\varepsilon - \frac{b(\mathbf{X})}{\sigma_0}\right)\right\}\right),$$

where $b(\mathbf{x}) = \nu - \mu + \sum_{j=1}^d (m_j(x_j) - g_{0,j}(x_j))$. Furthermore, since ε is independent of $\mathbf{X}$, it follows that $\Upsilon(\nu, m) = \mathbb{E}_{\mathbf{X}}\mathbb{E}_{\varepsilon}\{\rho(\varepsilon - [b(\mathbf{X})/\sigma_0])\}$. To simplify the notation, let $a(\mathbf{x}) = b(\mathbf{x})/\sigma_0$ and $\mathcal{B}_0 = \{\mathbf{x} : b(\mathbf{x}) = 0\}$. We have

$$\Upsilon(\nu, m) = \int_{\mathcal{B}_0} \mathbb{E}_{\varepsilon}(\rho(\varepsilon))\, dF_{\mathbf{X}}(\mathbf{x}) + \int_{\mathcal{B}_0^c} \mathbb{E}_{\varepsilon}(\rho(\varepsilon - a(\mathbf{x})))\, dF_{\mathbf{X}}(\mathbf{x})\,. \qquad (A.1)$$

Note that if either $\nu \neq \mu_0$ or $\mathbb{P}(\sum_{j=1}^d m_j(X_j) = \sum_{j=1}^d g_{0,j}(X_j)) < 1$ then $\mathbb{P}(\mathcal{B}_0) < 1$. To see this, assume that $\mathbb{P}(\mathcal{B}_0) = 1$ which implies that $\mathbb{E}[b(\mathbf{X})] = 0$. Since $\mathbb{E}[m_j(X_j)] = \mathbb{E}[g_{0,j}(X_j)] = 0$, for all $1 \leq j \leq d$, we have that $\nu = \mu_0$. Moreover, it then follows that $\mathbb{P}(\sum_{j=1}^d m_j(X_j) = \sum_{j=1}^d g_{0,j}(X_j)) = 1$, which is a contradiction.

In addition, Lemma 3.1 of Yohai (1987) and assumptions **E1** and **R1** imply that for all $a \neq 0$, $\mathbb{E}_{\varepsilon}[\rho(\varepsilon - a)] > \mathbb{E}_{\varepsilon}[\rho(\varepsilon)]$.

Hence, if $(\nu, m) \in \mathbb{R} \times \mathcal{H}^{ad}$ is such that either $\nu \neq \mu_0$ or $\mathbb{P}(\sum_{j=1}^d m_j(X_j) = \sum_{j=1}^d g_{0,j}(X_j)) < 1$ we have $\mathbb{P}(\mathcal{B}_0) < 1$, and then from (A.1) it follows that

$$\Upsilon(\nu, m) > \int_{\mathcal{B}_0} \mathbb{E}_{\varepsilon}(\rho(\varepsilon))\, dF_{\mathbf{X}}(\mathbf{x}) + \int_{\mathcal{B}_0^c} \mathbb{E}_{\varepsilon}(\rho(\varepsilon))\, dF_{\mathbf{X}}(\mathbf{x}) = \mathbb{E}_{\varepsilon}(\rho(\varepsilon)) = \Upsilon(\mu_0, g_0)\,.$$

(b) Follows immediately from (a) and **A1** noting that $g_j(P) - g_{0,j} \in \mathcal{H}_j$, $1 \leq j \leq d$. □

Proof of Theorem 2.2. For the sake of simplicity, denote $\mu = \mu(P)$ and $g_j = g_j(P)$. Note that $\Upsilon(\mu, g) \leq \Upsilon(\nu, g)$, since $\Upsilon(\mu, g) \leq \Upsilon(\nu, m)$. Then, if we denote $L(\nu) = \Upsilon(\nu, g)$, we have that $\mu = \mathrm{argmin}_{\nu \in \mathbb{R}} L(\nu)$ which leads to $L'(\mu) = 0$. Noting that

$$L'(\nu) = -\frac{1}{\sigma_0}\mathbb{E}\,\psi\left(\frac{Y - \nu - \sum_{j=1}^d g_j(X_j)}{\sigma_0}\right)$$

we obtain that $\Gamma_0(\mu, \mathbf{g}(P)) = 0$, as desired.

Let $1 \leq j \leq d$ be fixed and consider the problem of minimizing $\Upsilon(\mu, m)$ with respect to m_j for any $m(\mathbf{x}) \in \mathcal{H}^{ad}$ such that its $j-$th component is $m_j(X_j)$, the other ones been equal to g_s. To be more precise, for any $m_j \in \mathcal{H}_j$ let $m^{(j)} \in \mathcal{H}^{ad}$ be defined as $m^{(j)}(\mathbf{x}) = m_j(x_j) +$

$\sum_{s\neq j} g_s(x_s)$. Denote $L_j(m_j) = \Upsilon(\mu, m^{(j)}) = \mathbb{E}\,\rho\left((Y-\mu-m_j(X_j)-\sum_{s\neq j} g_s(X_s))/\sigma_0\right)$. Note that the fact that $\Upsilon(\mu, g) \leq \Upsilon(\nu, m)$ for any $m \in \mathcal{H}^{ad}$, entails that $L_j(g_j) \leq L_j(m_j)$. Hence, for any direction $\eta \in \mathcal{H}_j$, the partial Gateaux derivative of L_j at g_j along η should vanish. Denote this Gateaux derivative as $\partial L_j(g_j;\eta)$. Furthermore, let $\nu_\eta(t) = L_j(g_j + t\eta)$ and note that $\partial L_j(g_j;\eta) = \nu_\eta'(0)$, where

$$\nu_\eta'(0) \;=\; \lim_{t\to 0}\frac{1}{t}\,\mathbb{E}\left[\rho\left(\frac{R_j - g_j(X_j) - t\eta(X_j)}{\sigma_0}\right) - \rho\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\right], \tag{A.2}$$

with $R_j = Y - \mu - \sum_{s\neq j} g_s(X_s)$. Then, the first order condition states that $\nu_\eta'(0) = 0$, for any $\eta \in \mathcal{H}_j$. Note that for any $(x_1, x_2, \ldots, x_d, y)^{\mathrm{T}}$ we have

$$\frac{\partial}{\partial t}\left\{\rho\left(\frac{r_j - g_j(x_j) - t\,\eta(x_j)}{\sigma}\right)\right\} = \psi\left(\frac{r_j - g_j(x_j) - t\,\eta(x_j)}{\sigma}\right)\left(-\frac{\eta(x_j)}{\sigma}\right),$$

where $r_j = y - \mu - \sum_{\ell\neq j} g_\ell(x_\ell)$. Now we use (A.2) and the Dominating Convergence Theorem to obtain $\nu_\eta'(t) = -(1/\sigma_0)\mathbb{E}\left[\psi\left((R_j - g_j(X_j) - t\,\eta(X_j))/\sigma_0\right)\eta(X_j)\right]$, so that $\partial L_j(g_j;\eta) = -(1/\sigma_0)\,\mathbb{E}\left[\psi\left((R_j - g_j(X_j))/\sigma_0\right)\eta(X_j)\right]$. Hence, the first order condition $\nu_\eta'(0) = 0$ is

$$\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\eta(X_j)\right] = 0\,, \qquad \forall \eta \in \mathcal{H}_j\,. \tag{A.3}$$

Let h be any measurable function such that $\mathbb{E}|h(X_j)| < \infty$ and denote $a_h = \mathbb{E}h(X_j)$. Then, $\eta = h - a_h \in \mathcal{H}_j$, so from (A.3) we get that

$$\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)h(X_j)\right] = a_h\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\right]. \tag{A.4}$$

Recall that we have shown that $\Gamma_0(\mu, \mathbf{g}(P)) = 0$, i.e.,

$$\mathbb{E}\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right) = 0\,. \tag{A.5}$$

Therefore, from (A.4) and (A.5), we obtain that $\mathbb{E}\left[\psi\left((R_j - g_j(X_j))/\sigma_0\right)h(X_j)\right] = 0$, for any integrable function h, which implies that $\mathbb{E}\left[\psi\left((R_j - g_j(X_j))/\sigma_0\right)\middle|\, X_j = x\right] = 0$ a.s. concluding the proof since $\Gamma_j(\mu, \mathbf{g}, x_j) = \mathbb{E}\left[\psi\left((R_j - g_j(x_j))/\sigma_0\right)\middle|\, X_j = x_j\right]$. □

Proof of Theorem 2.3. Since the value of the objective function is not changed, we will assume that $\mathbb{E}\widetilde{g}_j^{(\ell)}(X_j) = 0$. Hence, $g_j^{(\ell)} = \widetilde{g}_j^{(\ell)}$ and $\mu^{(\ell)} = \widetilde{\mu}^{(\ell)}$. Note that the last equation in the $\ell-$th iteration of the algorithm is equivalent to solving $\mu^{(\ell)} = \operatorname{argmin}_{\mu\in\mathbb{R}}\mathbb{E}\rho\left((R_0^{(\ell)} - \mu)/\sigma_0\right)$, where $R_0^{(\ell)} = Y - \sum_{j=1}^d g_j^{(\ell)}(X_j)$, since ψ is strictly increasing so that the equation has a unique solution. On the other hand, in the $(k+1)-$th equation of the $\ell-$th iteration, we seek for a solution $a = g_k(X_k) \in \mathcal{H}_k$ of

$$\mathbb{E}\left[\psi\left(\frac{Y - \mu^{(\ell-1)} - \sum_{j=1}^{k-1} g_j^{(\ell)}(X_j) - \sum_{j=k+1}^{d} g_j^{(\ell-1)}(X_j) - a}{\sigma_0}\right)\middle|\, X_k\right] = 0\,.$$

which corresponds to finding the M-conditional location functional, as defined in Boente and Fraiman (1989), of the partial residuals $R_k^{(\ell)} = Y - \mu^{(\ell-1)} - \sum_{j=1}^{k-1} g_j^{(\ell)}(X_j) - \sum_{j=k+1}^{d} g_j^{(\ell-1)}(X_j)$. Using again that ψ is strictly increasing, we obtain that

$$g_k^{(\ell)}(X_k) = \underset{m_k \in \mathcal{H}_k}{\operatorname{argmin}} \mathbb{E}\left[\rho\left(\frac{R_k^{(\ell)} - m_k(X_k)}{\sigma_0}\right) \Big| X_k\right] .$$

Hence, taking expectation with respect to X_k, we get that

$$g_k^{(\ell)}(X_k) = \underset{m_k \in \mathcal{H}_k}{\operatorname{argmin}} \mathbb{E}\left[\rho\left(\frac{R_k^{(\ell)} - m_k(X_k)}{\sigma_0}\right)\right] .$$

Hence, for the $\ell-$th iteration, the system of equations in Algorithm 1 is equivalent to the following system of equations

$$\begin{cases} g_k^{(\ell)}(X_k) = \underset{m_k \in \mathcal{H}_k}{\operatorname{argmin}} \mathbb{E}\left[\rho\left(\dfrac{R_k^{(\ell)} - m_k(X_k)}{\sigma_0}\right)\right] & 1 \le k \le d \\ \mu^{(\ell)} = \underset{\nu \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\rho\left(\dfrac{R_0^{(\ell)} - \nu}{\sigma_0}\right) & \end{cases} \tag{A.6}$$

Let us show that this entails that $\{\upsilon_\ell\}_{\ell \ge 1}$ is a decreasing sequence where $\upsilon_\ell = \Upsilon(\mu^{(\ell)}, g^{(\ell)})$. Let $\mathbf{1}_d$ be the $d-$dimensional vector with all its components equal to 1. To reinforce the additive structure, denote $\Phi\left(\nu, \mathbf{m}\right) = \Upsilon(\nu, \mathbf{1}^{\mathrm{T}}\mathbf{m}) = \mathbb{E}\rho\left((Y - \nu - \sum_{j=1}^{d} m_j(X_j))/\sigma_0\right)$, where $\mathbf{m} = (m_1, \dots, m_d)^{\mathrm{T}}$.

We begin with Step 1. The first equation of the first iteration seeks for the first additive component through $g_1^{(1)}(X_1) = \operatorname{argmin}_{m_1 \in \mathcal{H}_1} \mathbb{E}\rho\left((R_1^{(1)} - m_1(X_1))/\sigma_0\right)$. Hence, choosing $m_1 = g_1^{(0)}$, we get that

$$\Phi\left(\mu^{(0)}, g_1^{(1)}, g_2^{(0)}, \dots, g_d^{(0)}\right) \le \Phi\left(\mu^{(0)}, g_1^{(0)}, g_2^{(0)}, \dots, g_d^{(0)}\right) = \Phi\left(\mu^{(0)}, \mathbf{g}^{(0)}\right) \le \Phi\left(\mu^{(0)}, \mathbf{g}^{(0)}\right) .$$

Assume that $\Phi\left(\mu^{(0)}, g_1^{(1)}, \dots, g_{k-1}^{(1)}, g_k^{(0)}, \dots, g_d^{(0)}\right) \le \Phi\left(\mu^{(0)}, \mathbf{g}^{(0)}\right)$ and consider the $k-$th equation of the first iteration. Then, as $g_k^{(1)}(X_k) = \operatorname{argmin}_{m_k \in \mathcal{H}_k} \mathbb{E}\left[\rho\left((R_k^{(1)} - m_k(X_k))/\sigma_0\right)\right]$, we get $\Phi\left(\mu^{(0)}, g_1^{(1)}, \dots, g_k^{(1)}, g_{k+1}^{(0)} \dots, g_d^{(0)}\right) \le \Phi\left(\mu^{(0)}, g_1^{(1)}, \dots, g_{k-1}^{(1)}, g_k^{(0)}, \dots, g_d^{(0)}\right)$, choosing $m_k = g_k^{(0)}$. Applying these arguments for $1 \le k \le d$ we finally get for $k = d$ that

$$\Phi\left(\mu^{(0)}, \mathbf{g}^{(1)}\right) = \Phi\left(\mu^{(1)}, g_1^{(1)}, \dots, g_d^{(1)}\right) \le \Phi\left(\mu^{(0)}, g_1^{(1)}, \dots, g_{d-1}^{(1)}, g_d^{(0)}\right) \le \Phi\left(\mu^{(0)}, \mathbf{g}^{(0)}\right) . \tag{A.7}$$

Finally, using the last equation in (A.6), we have that $\mu^{(1)} = \operatorname{argmin}_{\nu \in \mathbb{R}} \mathbb{E}\rho\left((R_0^{(1)} - \nu)/\sigma_0\right) = \arg\min_{\nu \in \mathbb{R}} \Phi\left(\nu, \mathbf{g}^{(1)}\right)$, which entails that for any $\nu \in \mathbb{R}$, $\Phi\left(\mu^{(1)}, \mathbf{g}^{(1)}\right) \le \Phi\left(\nu, \mathbf{g}^{(1)}\right)$. In particular, taking $\nu = \mu^{(0)}$ we obtain that $\Phi\left(\mu^{(1)}, \mathbf{g}^{(1)}\right) \le \Phi\left(\mu^{(0)}, \mathbf{g}^{(1)}\right) \le \Phi\left(\mu^{(0)}, \mathbf{g}^{(0)}\right)$, where the last inequality follows from (A.7). Therefore, we have shown that $\upsilon_1 \le \upsilon_0$.

Let us consider $\ell > 1$ and assume that $\upsilon_s \le \upsilon_{s-1}$ for $s = 1, \dots, \ell$. As above, the $k-$th equation in (A.6) leads to

$$\Phi\left(\mu^{(\ell-1)}, g_1^{(\ell)}, \dots, g_k^{(\ell)}, g_{k+1}^{(\ell-1)}, \dots, g_d^{(\ell-1)}\right) \le \Phi\left(\mu^{(\ell-1)}, g_1^{(\ell)}, \dots, g_{k-1}^{(\ell)}, g_k^{(\ell-1)}, g_{k+1}^{(\ell-1)}, \dots, g_d^{(\ell-1)}\right) . \tag{A.8}$$

Using (A.8) iteratively for $k = 1, \dots d$, we get $\Phi\left(\mu^{(\ell-1)}, \mathbf{g}^{(\ell)}\right) \le \Phi\left(\mu^{(\ell-1)}, \mathbf{g}^{(\ell-1)}\right) = \upsilon_{\ell-1}$. Finally, using similar arguments as those considered above, we get easily that $\upsilon_\ell = \Phi\left(\mu^{(\ell)}, \mathbf{g}^{(\ell)}\right) \le \Phi\left(\mu^{(\ell-1)}, \mathbf{g}^{(\ell)}\right)$, so that $\upsilon_\ell \le \upsilon_{\ell-1}$. □

References

Alimadad, A. and Salibián-Barrera, M. (2012). An outlier-robust fit for generalized additive models with applications to disease outbreak detection. *Journal of the American Statistical Association*, **106**, 719-731.

Baek, J. and Wehrly, T. (1993). Kernel estimation for additive models under dependence. *Stochastic Processes and their Applications*, **47**, 95-112.

Bianco, A. and Boente, G. (1998). Robust kernel estimators for additive models with dependent observations. *The Canadian Journal of Statistics*, **6**, 239-255.

Bianco, A. and Boente, G. (2007). Robust estimators under a semiparametric partly linear autoregression model: asymptotic behavior and bandwidth selection. *Journal of Time Series Analysis*, **28**, 274-306.

Boente, G. and Fraiman, R. (1989). Robust nonparametric regression estimation. *Journal of Multivariate Analysis*, **29**, 180-198.

Boente, G. and Rodriguez, D. (2008). Robust bandwidth selection in semiparametric partly linear regression models: Monte Carlo study and influential analysis. *Computational Statistics and data Analysis*, **52**, 2808-2828.

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, **809**, 580-597.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics*, **17**, 453-555.

Cantoni, E. and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, **11**(2), 141-146.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Belmont, CA : Wadsworth.

Cleveland, W. (1985). *The elements of graphing data.* Bell Telephone Laboratories Inc., New Jersey.

Croux, C., Gijbels, I. and Prosdocimi, I. (2011) Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*, **68**, 31-44.

Dengyi, G. and Kawagochi, S. (1986). Relationship between the increase temperature and variation of ozone level over the Antarctica and Tibetan plateau in spring. *Advances in Atmospheric Sciences*, **3**, 489-498.

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

Friedman, J. H. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association* **76**, 817-823.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Härdle (1990). *Applied Nonparametric Regression.* Econometric Society Monographs, 19, Cambridge University Press, Cambridge.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models.* Springer.

Härdle, W. and Tsybakov, B. (1988). Robust nonparametric regression with simultaneous scale curve estimation. *Annals of Statistics*, **16**, 120-135.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* Monographs on Statistics and Applied Probability No. 43. Chapman and Hall, London.

Lacour, S.A., Monte, M., Diot, P., Brocca, J., Veron, N., Colin, P. and Leblond, V. (2006). Relationship between ozone and temperature during the 2003 heat wave in France: consequences for health data analysis. *BMC Public Health*, **6**, 261.

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

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

Manchester, L. (1996). Empirical Influence for robust smoothing. *Australian Journal of Statistics*, **38**, 275-296.

Maronna, R., Martin, R. and Yohai, V. (2006) *Robust Statistics, Theory and Methods.* John Wiley & Sons, Ltd.

Oh, H-S., Nychka, D.W. and Lee, T.C.M. (2007). The role of pseudo data for robust smoothing with applications to wavelet regression. *Biometrika*, **94**:4, 893-904.

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**, 166-179.

Opsomer, J.D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, **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 regression and other nonparametric models. *Annals of Statistics*, **13**, 689-705.

Tukey, J. (1977). *Exploratory Data Analysis.* Reading, MA: Addison–Wesley.

Wand, M.P. (1999). A Central Limit Theorem for local polynomial backfitting estimators. *Journal of Multivariate Analysis*, **70**, 57-65.

Welsh, A.H. (1996). Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica*, **6**, 347-366.

Wong, R. K. W., Yao, F. and Lee, Th. C. M. (2014). Robust estimation for generalized additive models. *Journal of Computational and Graphical Statistics*, **23**, 270-289.

Yohai, V. (1987). High breakdown–point and high efficiency estimates for regression. *Annals of Statistics*, **15**, 642-656.