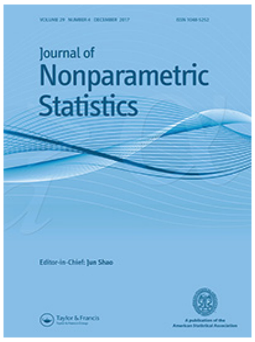

Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: http://www.tandfonline.com/loi/gnst20

Robust estimators for additive models using backfitting

Graciela Boente, Alejandra Martínez & Matías Salibián-Barrera

To cite this article: Graciela Boente, Alejandra Martínez & Matías Salibián-Barrera (2017) Robust estimators for additive models using backfitting, Journal of Nonparametric Statistics, 29:4, 744-767, DOI: 10.1080/10485252.2017.1369077

To link to this article: http://dx.doi.org/10.1080/10485252.2017.1369077

JOURNAL OF NONPARAMETRIC STATISTICS, 2017
VOL. 29, NO. 4, 744–767
https://doi.org/10.1080/10485252.2017.1369077

Taylor & Francis
Taylor & Francis Group

Robust estimators for additive models using backfitting

Graciela Boente[a], Alejandra Martínez[a] and Matías Salibián-Barrera[b]

[a]Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IMAS, CONICET, Buenos Aires, Argentina; [b]Department of Statistics, University of British Columbia, Vancouver, BC, 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. 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 optimisation problem. This formulation allows us to find mild conditions to show Fisher consistency 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.

ARTICLE HISTORY
Received 21 October 2015
Accepted 9 March 2017

KEYWORDS
Fisher-consistency; kernel weights; robust estimation; smoothing

AMS SUBJECT CLASSIFICATION
MSC 62G35; 62G08

1. Introduction

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

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

The error ε is assumed to be independent from $\mathbf{X}$ and centred at zero, while σ_0 is the error scale parameter. When ε has a finite first moment, we have the usual regression representation $E(Y \mid \mathbf{X}) = g_0(\mathbf{X})$. 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 $Y \mid \mathbf{X}$ 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 $Y \mid \mathbf{X}$. In this case, local robust estimators can be used to estimate the regression function as, for example, the local

CONTACT Graciela Boente gboente@dm.uba.ar Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IMAS, CONICET, Ciudad Universitaria, Pabellón 1, 1428 Buenos Aires, Argentina

Supplemental data for this article can be accessed here. https://doi.org/10.1080/10485252.2017.1369077

M-estimators proposed in Boente and Fraiman (1989) and the local regression quantiles studied in Welsh (1996).

Unfortunately both robust and non-robust nonparametric 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 nonparametric 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 jth covariate on the centre of the conditional distribution of Y. Moreover, Linton (1997), Fan, Härdle, and Mammen (1998) and Mammen, Linton, and Nielsen (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, Hastie, and Tibshirani (1989). The backfitting algorithm can be intuitively motivated by observing that, if Equation (2) holds, then

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

Hence, given a sample, the backfitting algorithm iteratively estimates the components $g_{0,j}$, $1 \le j \le d$, using a univariate smoother of the partial residuals in Equation (3) as functions of the jth 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, Linton, and Härdle 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 $m_1(X_1), \ldots, m_d(X_d)$ minimising $\mathbb{E}(Y - \mu_0 - \sum_{i=1}^{d} m_j(X_j))^2$ 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 Equation (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 $\mathbb{E}(Y \mid X_1, \ldots, X_d)$. Furthermore, the backfitting algorithm is a coordinate-wise descent algorithm minimising 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 nonparametric 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 generalised 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, Yao, and Lee (2014), who consider robust fits based on backfitting and penalised splines M-estimators, respectively. In the case of model (1), the approach of Wong et al. (2014) reduces to that of Oh, Nychka, and Lee (2007) which is an alternative based on penalised splines. On the other hand, Croux, Gijbels, and Prosdocimi (2011) provides a robust fit for generalised additive models with nuisance parameters using penalised 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 nonparametric regression estimators of Boente and Fraiman (1989) corresponds to minimising $\mathbb{E}[\rho((Y - \mu_0 - \sum_{i=1}^{d} m_j(X_j))/\sigma_0)]$ 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 Equation (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 optimisation problem are closely related to the robust conditional location functional defined in Boente and Fraiman (1989).

The rest of the paper is organised as follows. In Section 2, we show that the robust backfitting algorithm mentioned above corresponds to a coordinate-descent algorithm to minimise 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. Finally, in Section 5 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 minimise 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 $\mathbf{X} = (X_1, \dots, X_d)^{\mathrm{T}}$, 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 $\mathbb{E} g_{0,j}(X_j) = 0$, $1 \leq j \leq d$. When second moments exist, it is easy to see that the backfitting estimators solve the following minimisation problem

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

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

To reduce the effect of outliers on the regression estimates, we replace the square loss function in Equation (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 $\rho_1 = \rho_{\mathrm{H}}$ with $\rho_{\mathrm{H}}(u) = u^2/2$ if $|u| \leq 1$, $\rho_{\mathrm{H}}(u) = |u| - 1/2$ otherwise, and the Tukey bisquare loss to $\rho_1(u) = \rho_{\mathrm{T}}(u) = \min(3u^2 - 3u^4 + u^6, 1)$. Other 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 $\Upsilon(\nu, m) = \Upsilon(\nu, m, \sigma_0)$ with

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

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

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

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

To prove that the functional in Equation (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, Martin, and Yohai 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 Equation (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 αth coordinate equal to x and the other ones equal to X_j, $j \neq \alpha$ 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 $\int m(x_\alpha, \mathbf{u}_{\underline{\alpha}})\, \mathrm{d}F_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = 0$, where $F_{\mathbf{X}_{\underline{\alpha}}}$ denotes the distribution function of $\mathbf{X}_{\underline{\alpha}}$. Note that since $\mathbb{E}m_j(X_j) = 0$ for all j, $\int m(x_\alpha, \mathbf{u}_{\underline{\alpha}})\, \mathrm{d}F_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = m_\alpha(x_\alpha) + \int \sum_{j\neq\alpha} m_j(u_j)\, \mathrm{d}F_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = m_\alpha(x_\alpha)$. 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 $\int m(x_\alpha, \mathbf{u}_{\underline{\alpha}})\, \mathrm{d}F_{\mathbf{X}_{\underline{\alpha}}}(\mathbf{u}) = 0$. This has already been observed by Hastie and Tibshirani (1990, p. 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 which gives weak conditions for the unique representation and hence, as shown in Theorem 2.1, 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* $\inf_{\mathbf{x}\in\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}) > 0$. *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 optimisation 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 Equation* (4) *and let P stand for its distribution.*

(a) *If* (E1) *and* (R1) *hold, then, for each fixed* $\sigma > 0$, $\Upsilon(\nu, m, \sigma)$ *in Equation* (6) *achieves its unique minimum over* $\mathbb{R} \times \mathcal{H}^{ad}$ *at* $(\mu(P), g(P)) = (\mu(P), \sum_{j=1}^{d} g_j(P))$, *regardless of the value of* σ, *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 minimiser $(\mu(P), g(P))$ of Equation (7) always exists if ρ is a strictly convex function, even if (E1) does not hold. If in addition (A1) holds, the minimiser will have a unique representation.

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

$$\Gamma_0(\nu, \mathbf{m}) = \mathbb{E}\left[\psi\left(\frac{Y - \nu - \sum_{j=1}^{d} m_j(X_j)}{\sigma_0}\right)\right]$$

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

Our next theorem shows that it is possible to choose the solution $g(P)$ of Equation (7) so that its additive components $g_j = g_j(P)$ satisfy first-order conditions which are generalisations 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 minimiser of* $\Upsilon(\nu, m)$ *over* $\mathbb{R} \times \mathcal{H}^{ad}$ *where* $\mu(P) \in \mathbb{R}$, $g(P) = \sum_{j=1}^{d} g_j(P) \in \mathcal{H}^{ad}$, *that is,* $(\mu(P), g(P))$ *is the solution of Equation* (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 Equation (4) with the errors satisfying (E1), then $\boldsymbol{\Gamma}(\mu_0, \mathbf{g}_0, \mathbf{x}) = \mathbf{0}$. Moreover, if the model is heteroscedastic, that is, if $Y = g_0(\mathbf{X}) + \sigma_0(\mathbf{X})\varepsilon = \mu_0 + \sum_{j=1}^{d} g_{0,j}(X_j) + \sigma_0(\mathbf{X})\varepsilon$, where the errors ε are symmetrically distributed and the score function ψ is odd, then $(\mu_0, \mathbf{g}_0)$ satisfies

$\mathbb{E}\psi((Y-\mu_0-\sum_{j=1}^{d} g_{0,j}(X_j))/\sigma_0(\mathbf{X}))=0$ and for $1\le \ell \le d$,

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

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. 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), that is, $\mathbb{P}(\delta=1\,|\,(\mathbf{X},Y))=\mathbb{P}(\delta=1\,|\,\mathbf{X})=p(\mathbf{X})$. Define $(\mu(P),g(P))=\operatorname{argmin}_{(\nu,m)\in\mathbb{R}\times\mathcal{H}^{ad}}\Upsilon_\delta(\nu,m)$ where $\Upsilon_\delta(\nu,m)=\mathbb{E}\delta\rho((Y-\nu-\sum_{j=1}^{d} m_j(X_j))/\sigma_0(\mathbf{X}))=\mathbb{E}p(\mathbf{X})\rho((Y-\nu-\sum_{j=1}^{d} m_j(X_j))/\sigma_0(\mathbf{X}))$. 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 generalised 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 $\tilde{g}_\ell(x)$ of $\lambda_{\ell,\delta}(x,a)=0$ where

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

More precisely, let $\boldsymbol{\Gamma}_\delta(\nu,\mathbf{m},\mathbf{x})=(\Gamma_{0,\delta}(\nu,\mathbf{m}),\Gamma_{1,\delta}(\nu,\mathbf{m},x_1),\ldots,\Gamma_{d,\delta}(\nu,\mathbf{m},x_d))^{\mathrm{T}}$ with $\mathbf{m}=(m_1,\ldots,m_d)^{\mathrm{T}}$, $\Gamma_{0,\delta}(\nu,\mathbf{m})=\mathbb{E}[p(\mathbf{X})\psi((Y-\nu-\sum_{j=1}^{d} m_j(X_j))/\sigma_0)]$ and, for $1\le \ell\le d$, $\Gamma_{\ell,\delta}(\nu,\mathbf{m},x_\ell)=\mathbb{E}[p(\mathbf{X})\psi((Y-\mu(P)-\sum_{j\neq\ell} m_j(X_j)-m_\ell(X_\ell))/\sigma)|X_\ell=x_\ell]$. Similar arguments to those considered in the proof of Theorem 2.2, allow to show that if there exists a unique minimiser $(\mu(P),g(P))\in\mathbb{R}\times\mathcal{H}^{ad}$ of $\Upsilon_\delta(\nu,m)$, then $(\mu(P),\mathbf{g}(P))$ is a solution of $\boldsymbol{\Gamma}_\delta(\nu,\mathbf{m},\mathbf{x})=\mathbf{0}$. Note that instead of a simplified approach, a propensity score approach may be considered taking $\delta/p(\mathbf{X})$ instead of δ. In this case, $\Upsilon_\delta(\nu,m)=\Upsilon(\nu,m)$ defined in Equation (6) and $\Gamma_{\ell,\delta}=\Gamma_\ell$ defined in Equation (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 Equation (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 Equation (4). By Theorem 2.2, the robust functional $(\mu(P),\mathbf{g}(P))$ satisfies Equation (8). To simplify the notation, in what follows we will put $\mu=\mu(P)$ and $g_j=g_j(P)$, $1\le j\le d$ 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 $R_j^{(\ell)} = Y - \mu^{(\ell-1)} - \sum_{s=1}^{j-1} \tilde{g}_s^{(\ell)}(X_s) - \sum_{s=j+1}^{d} g_s^{(\ell-1)}(X_s)$
        Let $\tilde{g}_j^{(\ell)}$ solve
            $$\mathbb{E}\left[ \psi\left( \frac{R_j^{(\ell)} - \tilde{g}_j^{(\ell)}(X_j)}{\sigma_0} \right) \middle| X_j \right] = 0 \quad \text{a.s.}$$
    end for
    for $j = 1$ to $d$ do
        $g_j^{(\ell)} = \tilde{g}_j^{(\ell)} - \mathbb{E}[\tilde{g}_j^{(\ell)}(X_j)]$.
    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 $\boldsymbol{\ell}$** 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 $(\mu^{(\ell)}, \mathbf{g}^{(\ell)})_{\ell\geq 1} = (\mu^{(\ell)}, g_1^{(\ell)}, \dots, g_d^{(\ell)})_{\ell\geq 1}$ be the sequence obtained with Algorithm* 1. *Then, the sequence $\{\Upsilon(\mu^{(\ell)}, g^{(\ell)})\}_{\ell\geq 1}$ is non-increasing.*

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:

$$\frac{1}{n} \sum_{i=1}^{n} \psi\left(\frac{Y_i - \hat{\mu} - \sum_{j=1}^{d} \hat{g}_j(X_{i,j})}{\hat{\sigma}_0} \right) = 0$$

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

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

Algorithm 2 The sample version of the robust backfitting

```
Let $\ell = 0$ and $\hat{\mathbf{g}}^{(0)} = (\hat{g}_1^{(0)},\ldots,\hat{g}_d^{(0)})^{\mathrm{T}}$ be an initial set of additive components, for
example: $\hat{\mathbf{g}}^{(0)} = \mathbf{0}$, and let $\hat{\sigma}_0$ be a robust residual scale estimator. Moreover, let $\hat{\mu}^{(0)}$
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 $\mathbf{Z}_{i,j}(x_j) = (1,(X_{i,j}-x_j),(X_{i,j}-x_j)^2,\ldots,(X_{i,j}-x_j)^q)^{\mathrm{T}}$ and $\widehat{R}_{i,j}^{(\ell)} = Y_i -$
                $\hat{\mu}^{(\ell)} - \sum_{s=1}^{j-1}\tilde{g}_s^{(\ell)}(X_{i,s}) - \sum_{s=j+1}^{d}\hat{g}_s^{(\ell-1)}(X_{i,s})$.
            end for
            Let $\hat{\boldsymbol{\beta}}_j(x_j) = (\hat{\beta}_{0j}(x_j),\hat{\beta}_{1j}(x_j),\ldots,\hat{\beta}_{qj}(x_j))^{\mathrm{T}}$ be the solution to

                $\frac{1}{n}\sum_{i=1}^{n} K_h(X_{i,j}-x_j)\,\psi\left(\frac{\widehat{R}_{i,j}^{(\ell)}-\hat{\boldsymbol{\beta}}_j(x_j)^{\mathrm{T}}\mathbf{Z}_{i,j}(x_j)}{\hat{\sigma}_0}\right)\mathbf{Z}_{i,j}(x_j)=\mathbf{0}\,.$

            Let $\tilde{g}_j^{(\ell)}(x_j) = \hat{\beta}_{0j}(x_j)$.
        end for
    end for
    for $j = 1$ to $d$ do
        $\hat{g}_j^{(\ell)} = \tilde{g}_j^{(\ell)} - \sum_{i=1}^{n}\tilde{g}_j^{(\ell)}(X_{i,j})/n$.
    end for
    Let $\hat{\mu}^{(\ell)}$ solve

        $\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}^{(\ell)}-\sum_{j=1}^{d}\hat{g}_j^{(\ell)}(X_{i,j})}{\hat{\sigma}_0}\right)=0\,.$

until convergence
```

Remark 3.1: A possible choice of the preliminary scale estimator $\hat{\sigma}_0$ 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 $\hat{\sigma}_0 = \mathrm{mad}_{1\leq i\leq n}\{Y_i-\widehat{Y}_i\}$, 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 rth nearest neighbour among $\{X_{j,k}\}_{j\neq i}$.

3.1. Selection of the smoothing parameter

As with other nonparametric procedures, the selection of the smoothing parameter is an important practical issue when fitting additive models. The importance of using a robust criterion for selecting smoothing parameters, even when one uses robust estimators, has been described, for instance, by Leung, Marriott, and Wu (1993), Wang and Scott (1994), Boente, Fraiman, and Meloche (1997), Cantoni and Ronchetti (2001) and Leung (2005). Several proposals have been made in the literature, including L^1 cross-validation (Wang and Scott 1994), a robust version of C_p and cross-validation (Cantoni and Ronchetti 2001) and a robust plug-in procedure discussed in Boente et al. (1997).

Here we use an intuitively simple robust K-fold cross-validation method related to the procedure described in Bianco and Boente (2007). As usual, first randomly partition the data set into K disjoint subsets of approximately equal sizes, with indices $\mathcal{C}_j$, $1 \leq j \leq K$, so that $\bigcup_{j=1}^{K} \mathcal{C}_j = \{1, \ldots, n\}$. Let $\mathcal{G} \subset \mathbb{R}^d$ be the set of bandwidth combinations to be considered, and let $\hat{g}_{\mathbf{h}}^{(j)}(\mathbf{X})$ be the robust backfitting predictor for $\mathbf{X}$, computed with smoothing parameters $\mathbf{h} = (h_1, \ldots, h_d) \in \mathcal{G}$ and without using the observations with indices in $\mathcal{C}_j$. For each $i = 1, \ldots, n$, the prediction residuals $\hat{e}_i$ are

$$\hat{e}_i = Y_i - \hat{g}_{\mathbf{h}}^{(j)}(\mathbf{X}_i), \quad i \in \mathcal{C}_j, \ \ j = 1, \ldots, K.$$

The classical cross-validation procedure selects the bandwidth minimising the mean squared prediction error:

$$L(\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n} (\hat{e}_i - \bar{\hat{e}})^2 + \bar{\hat{e}}^2, \tag{9}$$

where $\bar{\hat{e}} = \sum_{i=1}^{n} \hat{e}_i/n$. To obtain a more robust cross-validation criterion, one can replace the squared prediction error above with a ρ-function (Leung 2005; Bianco and Boente 2007). This approach has good robustness properties when one uses a bounded ρ-function. However, Bianco and Boente (2007) also showed that in this case the selected bandwidths are noticeably more variable than when one uses a robust alternative to the variance/squared bias expression on the right-hand side of Equation (9). Specifically, let $\mu_n(\hat{e}_1, \ldots, \hat{e}_n)$ and $\sigma_n(\hat{e}_1, \ldots, \hat{e}_n)$ denote robust estimators of location and for the observed prediction errors $\hat{e}_1, \ \ldots, \hat{e}_n$. For example, we can take μ_n and σ_n to be their sample median and MAD (median of the absolute deviations with respect to the median), respectively. The robust cross-validation smoothing parameters are selected by minimising over $\mathbf{h} \in \mathcal{G}$ the following criterion:

$$L_{\text{R}}(\mathbf{h}) = \mu_n^2(\hat{e}_1, \ldots, \hat{e}_n) + \sigma_n^2(\hat{e}_1, \ldots, \hat{e}_n). \tag{10}$$

Leave-one-out cross-validation is a particularly important case of K-fold obtained when $K = n$ and $\mathcal{C}_j = \{j\}$, $1 \leq j \leq n$. This approach has also been considered in Boente and Rodriguez (2008) for partially linear models.

4. Numerical studies

In this section, we report the results of our numerical experiments designed to compare our proposed estimator with other alternatives proposed in the literature. All computations were carried out using an R implementation of our algorithm, publicly available on-line at http://github.com/msalibian/RBF.

We generated data following additive models with $d = 2$ and $d = 4$ components. Our experiments involved $N = 500$ samples for each simulation setting. For models with two additive components ($d = 2$) we include here the results obtained with samples of size $n = 100$ and bandwidths chosen using K-fold cross-validation. For $d = 4$ we used $n = 500$ and fixed bandwidths set to their asymptotic optimal values. Additional results for $d = 2$ and $n = 500$ are reported in Boente, Martinez, and Salibian-Barrera (2015).

We considered samples without outliers, four types of possible data contaminations, independent and correlated covariates, and also cases where the response variable was missing, as described in Remark 2.2. More specifically, 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 $u = \sigma_0 \varepsilon$. We then generated independent Bernoulli random variables $\{\delta_i\}_{i=1}^{n}$ such that $\mathbb{P}(\delta_i = 1|Y_i, \mathbf{X}_i) = \mathbb{P}(\delta_i = 1|\mathbf{X}_i) = p(\mathbf{X}_i)$. For models with $d = 4$ we considered the case without missing data ($p(\mathbf{x}) \equiv 1$) and also used $p(\mathbf{x}) = p_4(\mathbf{x}) = 0.4 + 0.5(\cos(x_1 x_3 + 0.2))^2$, which produces approximately 33% of missing Y_i's. Since other robust estimators proposed in the literature for this model cannot be applied directly to samples with missing observations, we ran a series of experiments with $d = 2$, $n = 100$ and no missing data. Comparisons between the robust and classical backfitting algorithm for $d = 2$ and missing data generated with $p(\mathbf{x}) = p_2(\mathbf{x}) = 0.4 + 0.5(\cos(x_1 + 0.2))^2$ can be found in Boente et al. (2015).

We compared the following estimators:

- The classical backfitting estimator (denoted $\hat{g}_{\mathrm{BC}}$) adapted to missing responses.
- A robust backfitting estimator (denoted $\hat{g}_{\mathrm{BR,H}}$) using Huber's loss function. This loss function ρ_c satisfies $\rho_c'(u) = \psi_c(u) = \min(c, \max(-c, u))$, and we used $c = 1.345$.
- A robust backfitting estimator (denoted $\hat{g}_{\mathrm{BR,T}}$) using Tukey's bisquare loss function. This loss function satisfies $\rho_c'(u) = \psi_c(u) = u(1 - (u/c)^2)^2 \mathbb{I}_{[-c,c]}(u)$, and we used $c = 4.685$. These estimators were computed using the Huber estimator $\hat{g}_{\mathrm{BR,H}}$ as the initial estimator in Step 1 of Algorithm 2.
- The estimator defined in Bianco and Boente (1998) (denoted $\hat{g}_{\mathrm{BB}}$).
- The estimator proposed by Croux et al. (2011) (denoted $\hat{g}_{\mathrm{CR}}$).

The univariate smoothers were computed using the Epanechnikov kernel $K(u) = 0.75(1 - u^2)\mathbb{I}_{[-1,1]}(u)$. We used local linear polynomials with $q = 0$ and $q = 1$ in Algorithm 2. Not surprisingly, our results for $d = 2$ show that in general local linear smoothers outperform locally constant ones. Hence, here we only report the results for $q = 1$, but see Boente et al. (2015) for additional tables. In what follows, classical backfitting estimates obtained using local linear smoothers ($q = 1$) are indicated as $\hat{g}_{\mathrm{BC},1}$, while the robust counterparts based on Tukey's bisquare and Huber's loss functions are denoted $\hat{g}_{\mathrm{BR,T},1}$ and $\hat{g}_{\mathrm{BR,H},1}$, respectively. Note that in order to perform a fair comparison between estimators we adapted the proposal in Bianco and Boente (1998) to the case $q = 1$, which we denote by $\hat{g}_{\text{BB},1}$.

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

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

where X_{ij} is the jth component of $\mathbf{X}_i$ and $\delta_i = 0$ if the ith 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

Our data were generated according to the additive model in Equation (2) with $n = 100$, $\sigma_0 = 0.5$ $\mu_0 = 0$, $g_{0,1}(x_1) = 24(x_1 - 0.5)^2 - 2$ and $g_{0,2}(x_2) = 2\pi\,\sin(\pi x_2) - 4$. The distributions of X_1 and X_2 were $U([0,1])$ and we considered two situations for the distribution of the vector $(X_1, X_2)^{\mathrm{T}}$: independent components, and $\text{cor}(X_1, X_2) = 0.7$. The latter was generated as follows. Let $\mathbf{Z} \sim N_2(0, \mathbf{\Sigma})$ be a bivariate Gaussian random vector where $\mathbf{\Sigma}_{1,1} = \mathbf{\Sigma}_{2,2} = 1$ and $\mathbf{\Sigma}_{1,2} = \mathbf{\Sigma}_{2,1} = 2\sin(\rho\pi/6)$, with $\rho \in (-1, 1)$. Let $X_j = \Phi(Z_j)$, $j = 1,2$, where Φ is the cumulative distribution function of a standard normal distribution. It follows that the marginal distribution of each X_j is $U([0,1])$, $j = 1,2$ and that their correlation equals ρ.

To select the bandwidths of the classical backfitting estimator we used the standard K-fold cross-validation procedure with a square loss function, while for the robust backfitting and the estimator of Bianco and Boente (1998) we used the robust K-fold method described in Section 3.1. In all these cases we set $K = 5$. The parameters involved in the estimators defined by Croux et al. (2011) were chosen as described therein.

For the error distribution, we considered the following settings:

- C_0: $u_i \sim N(0, \sigma_0^2)$.
- C_1: $u_i \sim (1 - 0.15)N(0, \sigma_0^2) + 0.15N(15, 0.01)$.
- 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: $u_i \sim (1 - 0.30)N(0, \sigma_0^2) + 0.30N(15, 0.01)$ for all i's such that $\mathbf{X}_i \in \mathcal{D}_{0.3}$.

The first 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 four contamination settings introduce asymmetrically distributed outliers, which are expected to induce significant bias in all the estimators. The goal of this experiment is to study the magnitude of this bias, which is expected to be bounded for the robust estimators, and lower than that of the classic estimator. 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 asymptotically optimal bandwidths 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 $[0.2, 05] \times [0.2, 0.5]$. Figures 1 and 2 illustrate these contamination scenarios on one randomly generated sample with independent and correlated covariates, respectively. The panels correspond to settings C_2, C_3 and C_4, with solid triangles indicating contaminated cases.

Note that for the case of correlated covariates the contamination setting C_2 produces samples with a very high number of outliers in neighbourhoods of points with one coordinate between 0.2 and 0.5 (see Figure 2). Since all the estimators considered in our experiment were severely affected in this setting, we omit the corresponding results here.

Tables 1 and 2 report the obtained values of the MISE, for 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. Since the values of ISE for some estimators have notably heavy tails which substantially distort their averages, we also report their median and 5% upper trimmed mean:

$$\frac{1}{\lfloor N0.95 \rfloor} \sum_{j=1}^{\lfloor N0.95 \rfloor} \text{ISE}_{(j)},$$

where $\text{ISE}_{(1)} \leq \text{ISE}_{(2)} \leq \cdots \leq \text{ISE}_{(N)}$ and $\lfloor x \rfloor$ denotes the integer part of x.

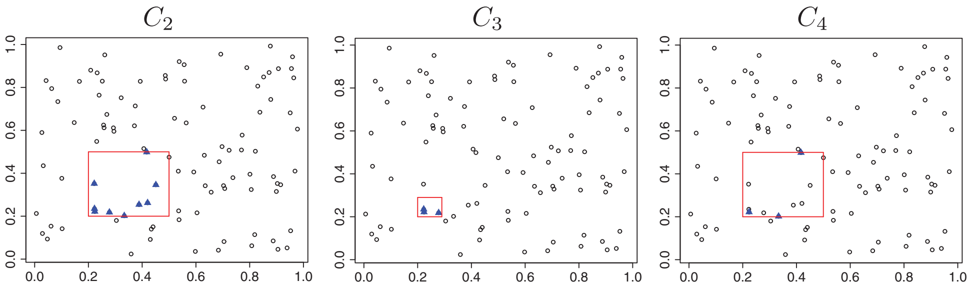

Figure 1. Scatter plots of covariates $(X_1, X_2)^t \sim U([0, 1]^2)$ with solid triangles indicating observations with contaminated response variables, for contamination settings C_2, C_3 and C_4. The square regions indicate the sets $\mathcal{D}_\eta$ for each scenario.

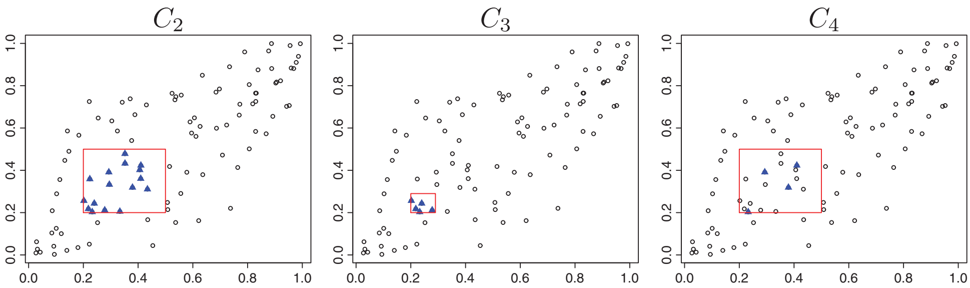

Figure 2. Scatter plots of covariates $(X_1, X_2)^t$ with solid triangles indicating observations with contaminated response variables, for contamination settings C_2, C_3 and C_4. The square regions indicate the sets $\mathcal{D}_\eta$ for each scenario and the covariates have correlation 0.7 with marginal uniform distribution.

Table 1. Summary measures of the ise of the estimators of the regression function $g_0 = \mu_0 + \sum_{j=1}^{2} g_{0,j}$ and the additive components $g_{0,1}$ and $g_{0,2}$ under different contaminations and when the covariates are independent.

		$\hat{g}_{\rm bc}$	$\hat{g}_{\rm cr}$	$\hat{g}_{\rm bb}$	$\hat{g}_{\rm br,h}$	$\hat{g}_{\rm br,t}$	$\hat{g}_{1,\rm bc}$	$\hat{g}_{1,\rm cr}$	$\hat{g}_{1,\rm bb}$	$\hat{g}_{1,\rm br,h}$	$\hat{g}_{1,\rm br,t}$	$\hat{g}_{2,\rm bc}$	$\hat{g}_{2,\rm cr}$	$\hat{g}_{2,\rm bb}$	$\hat{g}_{2,\rm br,h}$	$\hat{g}_{2,\rm br,t}$
mise	C_0	0.032	0.135	0.391	0.038	0.038	0.049	0.093	0.562	0.052	0.052	0.050	0.112	0.462	0.054	0.053
	C_1	7.641	3.676	0.430	0.681	0.336	1.188	0.491	0.420	0.314	0.182	1.130	0.476	0.377	0.255	0.203
	C_2	8.113	2.336	0.474	1.867	0.740	3.223	1.068	0.324	0.930	0.376	3.113	0.502	0.329	0.631	0.338
	C_3	0.153	0.073	0.385	0.041	0.039	0.104	0.064	0.548	0.054	0.053	0.102	0.076	0.452	0.055	0.054
	C_4	1.125	0.138	0.376	0.049	0.037	0.522	0.108	0.475	0.059	0.052	0.452	0.069	0.386	0.055	0.053
5%ise	C_0	0.031	0.033	0.346	0.036	0.035	0.041	0.048	0.502	0.044	0.044	0.041	0.045	0.417	0.045	0.044
	C_1	7.224	3.246	0.329	0.305	0.046	1.032	0.399	0.358	0.095	0.052	0.997	0.383	0.300	0.088	0.051
	C_2	7.562	1.698	0.330	1.090	0.076	2.998	0.862	0.266	0.431	0.051	2.873	0.276	0.225	0.170	0.052
	C_3	0.109	0.033	0.339	0.038	0.036	0.085	0.046	0.492	0.046	0.045	0.079	0.041	0.406	0.046	0.045
	C_4	0.971	0.101	0.333	0.045	0.035	0.451	0.089	0.420	0.051	0.044	0.380	0.052	0.344	0.046	0.044
medise	C_0	0.030	0.031	0.316	0.035	0.034	0.033	0.037	0.477	0.036	0.037	0.031	0.032	0.405	0.035	0.034
	C_1	7.179	2.947	0.313	0.194	0.042	0.915	0.343	0.343	0.076	0.044	0.873	0.305	0.298	0.073	0.039
	C_2	7.498	1.410	0.309	0.220	0.039	2.945	0.767	0.243	0.142	0.041	2.896	0.209	0.214	0.097	0.040
	C_3	0.072	0.032	0.317	0.036	0.035	0.069	0.038	0.467	0.040	0.038	0.057	0.030	0.397	0.037	0.035
	C_4	0.827	0.074	0.312	0.042	0.034	0.397	0.075	0.390	0.043	0.038	0.314	0.041	0.334	0.036	0.035

Note: We report the mise, the 5% left trimmed mean (labelled 5%ise) and the median of the ise, indicated as medise.

Table 2. Summary measures of the ise of the estimators of the regression function $g_0 = \mu_0 + \sum_{j=1}^{2} g_{0,j}$ and the additive components $g_{0,1}$ and $g_{0,2}$ under different contaminations and when the covariates have correlation 0.7.

		$\hat{g}_{\rm bc}$	$\hat{g}_{\rm cr}$	$\hat{g}_{\rm bb}$	$\hat{g}_{\rm br,h}$	$\hat{g}_{\rm br,t}$	$\hat{g}_{1,\rm bc}$	$\hat{g}_{1,\rm cr}$	$\hat{g}_{1,\rm bb}$	$\hat{g}_{1,\rm br,h}$	$\hat{g}_{1,\rm br,t}$	$\hat{g}_{2,\rm bc}$	$\hat{g}_{2,\rm cr}$	$\hat{g}_{2,\rm bb}$	$\hat{g}_{2,\rm br,h}$	$\hat{g}_{2,\rm br,t}$
mise	C_0	0.035	0.107	1.299	0.043	0.045	0.062	0.120	1.581	0.070	0.073	0.065	0.141	1.358	0.074	0.078
	C_1	7.581	3.756	1.417	0.605	0.327	1.773	0.918	1.467	0.329	0.260	1.674	0.955	1.289	0.252	0.159
	C_3	0.286	0.056	1.281	0.049	0.044	0.173	0.071	1.561	0.075	0.072	0.168	0.074	1.345	0.076	0.076
	C_4	1.690	0.234	1.182	0.064	0.043	0.710	0.182	1.430	0.084	0.071	0.617	0.102	1.247	0.076	0.074
5%ise	C_0	0.033	0.033	1.235	0.039	0.041	0.053	0.058	1.502	0.061	0.064	0.056	0.058	1.297	0.065	0.068
	C_1	7.120	3.255	1.114	0.299	0.051	1.557	0.752	1.310	0.143	0.073	1.414	0.752	1.145	0.126	0.073
	C_3	0.229	0.039	1.226	0.046	0.041	0.147	0.058	1.484	0.066	0.063	0.141	0.052	1.286	0.067	0.067
	C_4	1.504	0.183	1.133	0.057	0.040	0.628	0.154	1.359	0.074	0.062	0.531	0.082	1.190	0.066	0.064
medise	C_0	0.034	0.032	1.253	0.038	0.040	0.047	0.048	1.519	0.054	0.055	0.047	0.047	1.301	0.056	0.059
	C_1	7.135	2.919	1.107	0.201	0.046	1.372	0.589	1.298	0.113	0.068	1.266	0.534	1.155	0.101	0.063
	C_3	0.123	0.036	1.238	0.043	0.038	0.124	0.051	1.493	0.060	0.057	0.108	0.045	1.281	0.058	0.058
	C_4	1.383	0.138	1.145	0.054	0.038	0.542	0.132	1.379	0.067	0.055	0.458	0.070	1.207	0.058	0.058

Note: We report the mise, the 5% left trimmed mean (labelled 5%ise) and the median of the ise, indicated as medise.

As expected, when the data do not contain outliers the robust backfitting estimators $\hat{g}_{\rm BR,H}$ and $\hat{g}_{\rm BR,T}$ are slightly less efficient than the least squares one, although the differences are very small. On the other hand, the estimators $\hat{g}_{\rm CR}$ and $\hat{g}_{\rm BB}$ show much larger mean square errors than our proposal. Also note that the performance of $\hat{g}_{\rm BB}$ deteriorates further when the covariates are correlated (Table 2) since these estimators are biased unless $Y - g_{0,j}(X_j)$ is independent from X_j, $j = 1{,}2$.

For the contamination cases C_1 and C_2, when using the backfitting algorithm combined with local linear smoothers, the MISE of the classical estimator for g_0 is notably larger than those of all robust estimators (more than 20 times larger) for independent covariates. This difference is smaller when estimating each component $g_{0,1}$ and $g_{0,2}$, but remains fairly large nonetheless. A similar behaviour is observed under C_1 when the covariates are correlated. This contamination causes the most damage to the estimator of Croux et al. (2011), with a resulting MISE which is 10 times that of the robust backfitting and 30 times that obtained under C_0.

Comparing the three summary measures in Tables 1 and 2 we see that the ISE's for $\hat{g}_{\text{BR,H}}$ and $\hat{g}_{\text{BR,T}}$ have very heavy tails. By looking at the 5% upper trimmed mean and median of the ISE's, we note that as we progressively reduce the impact of the tails, the performance summaries of the estimator based on Tukey's bisquare loss function under the different contamination settings (C_1 through C_4) are very close to that observed with clean samples. The second most stable robust estimator was $\hat{g}_{\text{BR,H}}$.

In general, when the data contain outliers, we note that the robust backfitting estimators give noticeably better regression estimators (both for g_0 and its components) than the classical one and outperforms the estimators proposed in Bianco and Boente (1998) and Croux et al. (2011). Based on the above results, from its stability with respect to the studied contaminations and the lower bias under the central model, we recommend the robust backfitting algorithm combined with local linear smoothers computed with Tukey's loss function.

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 $\varepsilon_i \sim N(0,1)$, set $\mu_0 = 0$ and $\sigma_0 = 0.15$. The additive components were: $g_{0,1}(x_1) = x_1^3/12$, $g_{0,2}(x_2) = \sin(-x_2)$, $g_{0,3}(x_3) = x_3^2/2 - 1.5$, $g_{0,4}(x_4) = \mathrm{e}^{x_4}/4 - (\mathrm{e}^3 - \mathrm{e}^{-3})/24$. Based on the results obtained for two additive components, with $d=4$ we only compared the classical backfitting estimator and the robust proposal described in this paper. 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 $1 \le j \le 4$.

Due to the intensive computational effort required to perform $K-$fold cross–validation with 4 bandwidths, we report here results obtained using fixed bandwidths set to their optimal asymptotic value. These fixed bandwidths were computed assuming that the other components in the model are known (Härdle, Müller, Sperlich, and Werwatz 2004), 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 (using this vector of bandwidths we expect an average of 5 points in each 4-dimensional neighbourhood). We then applied the backfitting algorithm with the optimal bandwidths $\mathbf{h}_{\text{opt}}^{\text{MISE}}$.

Tables 3–5 report the MISE for the different estimators, contamination settings and missing mechanisms. Our experiments with and without missing responses yield similar conclusions regarding the advantage of the robust procedure over the classical backfitting. As expected, when responses are missing, all the ISE's are slightly inflated. 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. The different summary measures show that the ISE's 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. A similar

Table 3. Summary measures for the ise of the estimators of the regression function $g_0 = \mu_0 + \sum_{j=1}^{4} g_{0,j}$ under different contaminations and missing mechanisms.

		$p(\mathbf{x}) \equiv 1$			$p(\mathbf{x}) = p_4(\mathbf{x})$		
		$\hat{g}_{\text{bc}}$	$\hat{g}_{\text{br,h}}$	$\hat{g}_{\text{br,t}}$	$\hat{g}_{\text{bc}}$	$\hat{g}_{\text{br,h}}$	$\hat{g}_{\text{br,t}}$
mise	C_0	0.0023	0.0023	0.0023	0.0033	0.0033	0.0033
	C_1	7.6095	0.0497	0.0046	8.8581	0.1563	0.0932
	C_2	4.8224	0.0221	0.0025	6.0121	0.0258	0.0038
5%ise	C_0	0.0023	0.0023	0.0023	0.0032	0.0032	0.0032
	C_1	7.4435	0.0430	0.0027	8.6458	0.0618	0.0085
	C_2	4.6718	0.0209	0.0024	5.8007	0.0226	0.0035
medise	C_0	0.0023	0.0023	0.0023	0.0033	0.0033	0.0033
	C_1	7.5283	0.0421	0.0027	8.7854	0.0435	0.0040
	C_2	4.7560	0.0203	0.0024	5.9358	0.0215	0.0035

Table 4. Summary measures for the ise of the estimators of the additive component $g_{0,1}$ and $g_{0,2}$ under different contaminations and missing mechanisms.

		$p(\mathbf{x}) \equiv 1$						$p(\mathbf{x}) = p_4(\mathbf{x})$					
		$\hat{g}_{1,BC}$	$\hat{g}_{1,\text{br,h}}$	$\hat{g}_{1,\text{br,t}}$	$\hat{g}_{2,\text{bc}}$	$\hat{g}_{2,\text{br,h}}$	$\hat{g}_{2,\text{br,t}}$	$\hat{g}_{1,\text{bc}}$	$\hat{g}_{1,\text{br,h}}$	$\hat{g}_{1,\text{br,t}}$	$\hat{g}_{2,\text{bc}}$	$\hat{g}_{2,\text{br,h}}$	$\hat{g}_{2,\text{br,t}}$
mise	C_0	0.0020	0.0020	0.0020	0.0016	0.0016	0.0016	0.0030	0.0030	0.0030	0.0024	0.0024	0.0024
	C_1	0.6356	0.0066	0.0021	0.6189	0.0081	0.0026	0.9703	0.0250	0.0178	0.8982	0.0226	0.0136
	C_2	0.9897	0.0060	0.0020	0.9482	0.0055	0.0016	1.1960	0.0073	0.0030	1.2339	0.0067	0.0024
5%ise	C_0	0.0017	0.0017	0.0017	0.0014	0.0014	0.0014	0.0025	0.0025	0.0025	0.0020	0.0020	0.0020
	C_1	0.6023	0.0061	0.0018	0.5865	0.0057	0.0014	0.9201	0.0083	0.0027	0.8493	0.0070	0.0022
	C_2	0.9520	0.0056	0.0017	0.9154	0.0051	0.0014	1.1434	0.0066	0.0026	1.1887	0.0062	0.0020
medise	C_0	0.0013	0.0013	0.0013	0.0011	0.0011	0.0011	0.0019	0.0019	0.0019	0.0016	0.0016	0.0016
	C_1	0.5909	0.0059	0.0014	0.5879	0.0054	0.0012	0.9060	0.0075	0.0021	0.8413	0.0064	0.0018
	C_2	0.9677	0.0056	0.0013	0.9260	0.0050	0.0011	1.1591	0.0061	0.0020	1.2041	0.0059	0.0016

Notes: $p(\mathbf{x}) \equiv 1$ corresponds to the case of no missing responses and $p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 x_3 + 0.2)$ to missing responses according to p_4. We report the mise, the 5% left trimmed mean (labelled 5%ise) and the median of the ise, indicated as medise.

Table 5. Summary measures for the ise of the estimators of the additive component $g_{0,3}$ and $g_{0,4}$ under different contaminations and missing mechanisms.

		$p(\mathbf{x}) \equiv 1$						$p(\mathbf{x}) = p_4(\mathbf{x})$					
		$\hat{g}_{3,\text{bc}}$	$\hat{g}_{3,\text{br,h}}$	$\hat{g}_{3,\text{br,t}}$	$\hat{g}_{4,\text{bc}}$	$\hat{g}_{4,\text{br,h}}$	$\hat{g}_{4,\text{br,t}}$	$\hat{g}_{3,\text{bc}}$	$\hat{g}_{3,\text{br,h}}$	$\hat{g}_{3,\text{br,t}}$	$\hat{g}_{4,\text{bc}}$	$\hat{g}_{4,\text{br,h}}$	$\hat{g}_{4,\text{br,t}}$
mise	C_0	0.0042	0.0042	0.0042	0.0036	0.0036	0.0036	0.0082	0.0082	0.0082	0.0058	0.0058	0.0058
	C_1	0.6741	0.0106	0.0052	0.7558	0.0097	0.0037	1.0679	0.0449	0.0337	1.2310	0.0583	0.0429
	C_2	1.0007	0.0085	0.0042	1.0592	0.0078	0.0036	1.2256	0.0126	0.0082	1.3877	0.0117	0.0061
5%ise	C_0	0.0032	0.0032	0.0032	0.0029	0.0029	0.0029	0.0065	0.0065	0.0065	0.0046	0.0046	0.0046
	C_1	0.6406	0.0081	0.0034	0.7226	0.0085	0.0030	1.0099	0.0136	0.0070	1.1701	0.0135	0.0055
	C_2	0.9660	0.0074	0.0033	1.0247	0.0071	0.0029	1.1756	0.0109	0.0066	1.3374	0.0094	0.0047
medise	C_0	0.0023	0.0023	0.0023	0.0022	0.0022	0.0022	0.0044	0.0044	0.0044	0.0034	0.0034	0.0034
	C_1	0.6297	0.0075	0.0024	0.7153	0.0078	0.0023	1.0010	0.0115	0.0046	1.1628	0.0112	0.0037
	C_2	0.9907	0.0069	0.0023	1.0447	0.0066	0.0022	1.1953	0.0096	0.0045	1.3391	0.0084	0.0035

Notes: $p(\mathbf{x}) \equiv 1$ corresponds to the case of no missing responses and $p_4(\mathbf{x}) = 0.4 + 0.5\cos^2(x_1 x_3 + 0.2)$ to missing responses according to p_4. We report the mise, the 5% left trimmed mean (labelled 5%ise) and the median of the ise, indicated as medise.

behaviour was observed for models with missing data and $d = 2$ (see Boente et al. 2015). Based on these observations, we also recommend using our robust backfitting method using local linear smoothers and Tukey's loss function.

5. 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, Cleveland, Kleiner, and Tukey 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 nonlinear, 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 jth covariate, $1 \leq j \leq 3$, 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 $(\mathbf{X}_1^{\mathrm{T}}, Y_1)^{\mathrm{T}}, \ldots, (\mathbf{X}_n^{\mathrm{T}}, Y_n)^{\mathrm{T}}$ be the considered observations ($n = 111$). The usual leave-one-out cross-validation criterion in this setting is given by $L_{\text{LS}}(\mathbf{h}) = (1/n) \sum_{i=1}^{n} (Y_i - \hat{g}_{\text{BC},\mathbf{h}}^{-i}(\mathbf{X}_i))^2$, where $\hat{g}_{\text{BC},\mathbf{h}}^{-i}(\mathbf{X}_i)$ denotes the backfitting predictor for $\mathbf{X}_i$, computed with bandwidth $\mathbf{h} \in \mathcal{G}$ and without using the ith observation. For the classical backfitting estimator the smallest value of L_{LS} over the grid $\mathcal{G}$ was obtained at $\mathbf{h}_{\text{LS}} = (9.53, 10.67, 91.15)$.

As mentioned in Section 3, 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. For this real data set, we have considered the robust leave–one–out cross-validation criterion defined through $L_{\text{R}}(\mathbf{h})$ in Equation (10) taking μ_n as the median and σ_n as the MAD. More precisely, let $\hat{g}_{\text{BR,T},\mathbf{h}}^{-i}(\mathbf{X}_i)$ denote the robust backfitting predictor at $\mathbf{X}_i$, computed with the smoothing parameter $\mathbf{h} \in \mathcal{G}$ and without using the ith observation. The robust cross-validation criterion used is

$$L_{\text{R}}(\mathbf{h}) = \left(\text{median}_{1 \leq i \leq n}\{Y_i - \hat{g}_{\text{BR,T},\mathbf{h}}^{-i}(\mathbf{X}_i)\}\right)^2 + \left(\text{MAD}_{1 \leq i \leq n}\{Y_i - \hat{g}_{\text{BR,T},\mathbf{h}}^{-i}(\mathbf{X}_i)\}\right)^2 .$$

The minimum of L_{R} over $\mathcal{G}$ was obtained at $\mathbf{h}_{\text{R}} = (4.76, 8.89, 136.73)$, 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 3 shows the estimated regression components for each explanatory variable, both for the classical and robust estimators. The plots of the partial residual are given in the supplemental file available on-line. 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

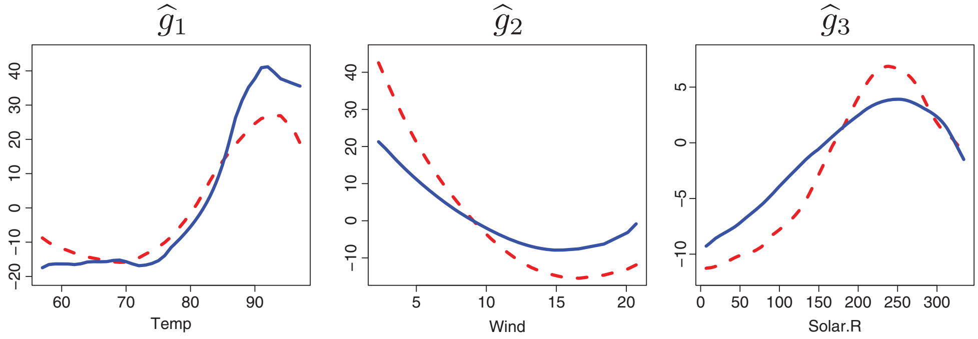

Figure 3. Estimated curves for the classical (in dashed lines) and robust (in solid lines) backfitting estimators with data-driven bandwidths $\mathbf{h}_{ls}$ and $\mathbf{h}_r$, respectively.

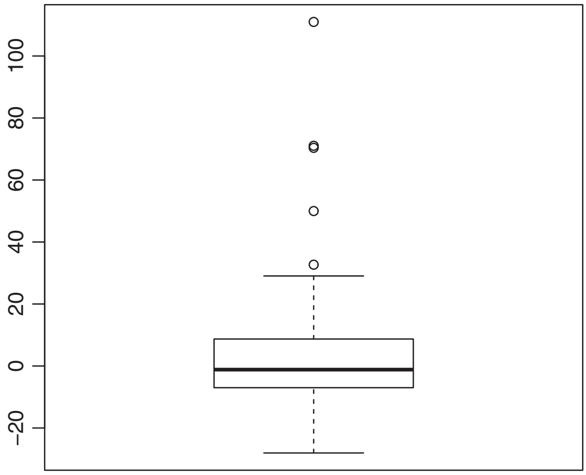

Figure 4. Boxplot of the residuals obtained using the robust fit with data-driven bandwidth $\mathbf{h}_r$.

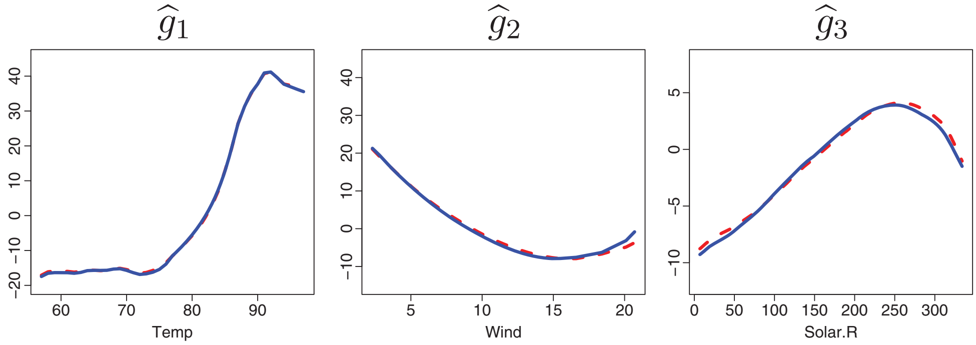

Figure 5. Estimated curves for the classical backfitting estimator, $\hat{g}_j^{(-5)}$ (in dashed lines) with data-driven bandwidth $\mathbf{h}_{\text{ls}}^{(-5)}$ and for the robust ones (in solid lines) computed with all the data and with data-driven bandwidth $\mathbf{h}_{\text{r}}$.

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 4 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 $\mathbf{h}_{\text{LS}}^{(-5)} = (4.85, 10.52, 138.87)$. Note that these values are very similar to those obtained with the robust estimator combined with the robust cross-validation criterion. Figure 5 shows the estimates, $\hat{g}_j^{(-5)}, j = 1, \ldots, 3$, 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.

Supplemental material

The supplementary file includes two Sections labelled S.1 and S.2

[S.1: *Empirical influence*] To study the sensitivity of the robust backfitting with respect to single outliers, we provide a numerical study of the empirical influence function.

[S.2: *Real data example*] This section contains partial residuals plots for each explanatory variable, both for the classical and robust estimators. For the classical estimators, partial residuals are given for the complete data set and for the data without the outliers.

Acknowledgements

The authors wish to thank the Associate Editor and two anonymous referees for valuable comments which led to an improved version of the original paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was partially supported by Grants PIP 112-201101-00339 from Consejo Nacional de Investigaciones Científicas y Técnicas, PICT 2014-0351 from Fondo para la Investigación Científica y Tecnológica, ANPCYT and 20020130100279BA 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 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.

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.E. (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 Semi-Parametric Partly Linear Autoregression: Asymptotic Behaviour 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.

Boente, G., Fraiman, R., and Meloche, J. (1997), 'Robust Plug-in Bandwidth Estimators in Nonparametric Regression', *Journal of Statistical Planning and Inference*, 57, 109–142.

Boente, G., Martinez, A., and Salibian–Barrera, M. (2015), 'Robust Estimators for Additive Models using Backfitting.' Technical report available at http://www.stat.ubc.ca/ ~ matias/RBF.

Breiman, L., and Friedman, J.H. (1985), 'Estimating Optimal Transformations for Multiple Regression and Correlation', *Journal of the American Statistical Association*, 809, 580–598.

Buja, A., Hastie, T., and Tibshirani, R. (1989), 'Linear Smoothers and Additive Models (with discussion)', *Annals of Statistics*, 17, 453–510.

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*, New Jersey: Bell Telephone Laboratories Inc.

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.

Härdle, W. (1990), 'Applied Nonparametric Regression, Econometric Society Monographs, Vol. 19, Cambridge: Cambridge University Press, 1990.

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

Härdle, W., and Tsybakov, A. B. (1988), 'Robust Nonparametric Regression with Simultaneous Scale Curve Estimation', *Annals of Statistics*, 16, 120–135.

Hastie, T.J., and Tibshirani, R.J., *Generalized Additive Models*, Monographs on Statistics and Applied Probability, Vol. 43, London: Chapman and Hall, 1990.

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.

Leung, D.H.-Y. (2005), 'Cross-Validation in Nonparametric Regression with Outliers', *Annals of Statistics*, 33, 2291–2310.

Leung, D.H.-Y., Marriott, F.H.C., and Wu, E.K.H. (1993), 'Bandwidth Selection in Robust Smoothing', *Journal of Nonparametric Statistics*, 4, 333–339.

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.

Maronna, R., Martin, R., and Yohai, V. (2006), *Robust Statistics, Theory and Methods*, Chichester: 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.

Wand, M.P. (1999), 'A Central Limit Theorem for Local Polynomial Backfitting Estimators', *Journal of Multivariate Analysis*, 70, 57–65.

Wang, F., and Scott, D. (1994), 'The L_1 Method for Robust Nonparametric Regression', *Journal of the American Statistical Association*, 89, 65–76.

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.J. (1987), 'High Breakdown-Point and High Efficiency Robust Estimates for Regression', *Annals of Statistics*, 15, 642–656.

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 $\Upsilon(\nu, m, \sigma) > \Upsilon(\mu_0, g_0, \sigma)$. For any $(\nu, m) \in \mathbb{R} \times \mathcal{H}^{ad}$ we have

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

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

$b(\mathbf{x})/\sigma$ and $\mathcal{B}_0 = \{\mathbf{x} : b(\mathbf{x}) = 0\}$. We have

$$\Upsilon(\nu, m, \sigma) = \int_{\mathcal{B}_0} \mathbb{E}_\varepsilon(\rho(\tilde{\epsilon}))\, \mathrm{d}F_{\mathbf{X}}(\mathbf{x}) + \int_{\mathcal{B}_0^c} \mathbb{E}_\varepsilon(\rho(\tilde{\epsilon} - a(\mathbf{x})))\, \mathrm{d}F_{\mathbf{X}}(\mathbf{x}). \tag{A1}$$

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.

It is worth noticing that $\tilde{\epsilon}$ satisfies (E1) since ε does. In addition, Lemma 3.1 of Yohai (1987) and assumptions (E1) and (R1) imply that for all $a \neq 0$, $\mathbb{E}_\varepsilon[\rho(\tilde{\epsilon} - a)] > \mathbb{E}_\varepsilon[\rho(\tilde{\epsilon})]$.

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 Equation (A1) it follows that

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

(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 $\mu = \operatorname{argmin}_{\nu \in \mathbb{R}} L(\nu)$ which leads to $L'(\mu) = 0$. Noting that $L'(\nu) = -(1/\sigma_0)\mathbb{E}\psi((Y - \nu - \sum_{j=1}^d g_j(X_j))/\sigma_0)$, 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 minimising $\Upsilon(\mu, m)$ with respect to m_j for any $m(\mathbf{x}) \in \mathcal{H}^{ad}$ such that its jth 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 $L_j(m_j) = \Upsilon(\mu, m^{(j)}) = \mathbb{E}\rho((Y - \mu - m_j(X_j) - \sum_{s \neq j} g_s(X_s))/\sigma_0)$. 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

$$\nu'_\eta(0) = \lim_{t \to 0} \frac{1}{t} \mathbb{E}\left[\rho\left(\frac{R_j - g_j(X_j) - t\eta(X_j)}{\sigma_0}\right) - \rho\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\right], \tag{A2}$$

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 Equation (A2) and the Dominating Convergence Theorem to obtain $\nu'_\eta(t) = -(1/\sigma_0)\mathbb{E}[\psi((R_j - g_j(X_j) - t\eta(X_j))/\sigma_0)\eta(X_j)]$, so that $\partial L_j(g_j; \eta) = -(1/\sigma_0)\mathbb{E}[\psi((R_j - g_j(X_j))/\sigma_0)\eta(X_j)]$. Hence, the first-order condition $\nu'_\eta(0) = 0$ is

$$\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\eta(X_j)\right] = 0, \quad \forall \eta \in \mathcal{H}_j. \tag{A3}$$

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 Equation (A3) we get that

$$\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)h(X_j)\right] = a_h \mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\right]. \tag{A4}$$

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

$$\mathbb{E}\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right) = 0. \tag{A5}$$

Therefore, from Equations (A4) and (A5), we obtain that $\mathbb{E}[\psi((R_j - g_j(X_j))/\sigma_0)h(X_j)] = 0$, for any integrable function h, which implies that $\mathbb{E}[\psi((R_j - g_j(X_j))/\sigma_0) \mid X_j = x] = 0$ a.s. concluding the proof since $\Gamma_j(\mu, \mathbf{g}, x_j) = \mathbb{E}[\psi((R_j - g_j(x_j))/\sigma_0) \mid X_j = x_j]$. ■

Proof of Theorem 2.3: Since the value of the objective function is not changed, we will assume that $\mathbb{E}\tilde{g}_j^{(\ell)}(X_j) = 0$. Hence, $g_j^{(\ell)} = \tilde{g}_j^{(\ell)}$ and $\mu^{(\ell)} = \tilde{\mu}^{(\ell)}$. Note that the last equation in the ℓth iteration of the algorithm is equivalent to solving $\mu^{(\ell)} = \operatorname{argmin}_{\mu\in\mathbb{R}} \mathbb{E}\rho((R_0^{(\ell)} - \mu)/\sigma_0)$, 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 ℓth 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

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

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

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

Hence, for the ℓth iteration, the system of equations in Algorithm 1 is equivalent to the following system of equations

$$\begin{aligned} g_k^{(\ell)} &= \operatorname*{argmin}_{m_k\in\mathcal{H}_k} \mathbb{E}\left[\rho\left(\frac{R_k^{(\ell)} - m_k(X_k)}{\sigma_0}\right)\right] \qquad 1 \le k \le d \\ \mu^{(\ell)} &= \operatorname*{argmin}_{\nu\in\mathbb{R}} \mathbb{E}\rho\left(\frac{R_0^{(\ell)} - \nu}{\sigma_0}\right). \end{aligned} \tag{A6}$$

Let us show that this entails that $\{\upsilon_\ell\}_{\ell\ge1}$ 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 $\Phi(\nu, \mathbf{m}) = \Upsilon(\nu, \mathbf{1}^\mathrm{T}\mathbf{m}) = \mathbb{E}\rho((Y - \nu - \sum_{j=1}^d m_j(X_j))/\sigma_0)$, 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 $g_1^{(1)} = \operatorname{argmin}_{m_1\in\mathcal{H}_1} \mathbb{E}\rho((R_1^{(1)} - m_1(X_1))/\sigma_0)$. Hence, choosing $m_1 = g_1^{(0)}$, we get that $\Phi(\mu^{(0)}, g_1^{(1)}, g_2^{(0)}, \dots, g_d^{(0)}) \le \Phi(\mu^{(0)}, g_1^{(0)}, g_2^{(0)}, \dots, g_d^{(0)}) = \Phi(\mu^{(0)}, \mathbf{g}^{(0)}) \le \Phi(\mu^{(0)}, \mathbf{g}^{(0)})$.

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

$$\Phi(\mu^{(0)}, \mathbf{g}^{(1)}) = \Phi(\mu^{(1)}, g_1^{(1)}, \dots, g_d^{(1)}) \le \Phi(\mu^{(0)}, g_1^{(1)}, \dots, g_{d-1}^{(1)}, g_d^{(0)}) \le \Phi(\mu^{(0)}, \mathbf{g}^{(0)}). \tag{A7}$$

Finally, using the last equation in (A6), we have that $\mu^{(1)} = \operatorname{argmin}_{\nu\in\mathbb{R}} \mathbb{E}\rho((R_0^{(1)} - \nu)/\sigma_0) = \arg\min_{\nu\in\mathbb{R}} \Phi(\nu, \mathbf{g}^{(1)})$, which entails that for any $\nu \in \mathbb{R}$, $\Phi(\mu^{(1)}, \mathbf{g}^{(1)}) \le \Phi(\nu, \mathbf{g}^{(1)})$. In particular, taking $\nu = \mu^{(0)}$ we obtain that $\Phi(\mu^{(1)}, \mathbf{g}^{(1)}) \le \Phi(\mu^{(0)}, \mathbf{g}^{(1)}) \le \Phi(\mu^{(0)}, \mathbf{g}^{(0)})$, where the last inequality follows from Equation (A7). Therefore, we have shown that $\upsilon_1 \le \upsilon_0$.

Let us consider $\ell > 1$ and assume that $\upsilon_s \leq \upsilon_{s-1}$ for $s = 1, \ldots, \ell$. As above, the kth equation in (A6) leads to

$$\Phi(\mu^{(\ell-1)}, g_1^{(\ell)}, \ldots, g_k^{(\ell)}, g_{k+1}^{(\ell-1)}, \ldots, g_d^{(\ell-1)}) \leq \Phi(\mu^{(\ell-1)}, g_1^{(\ell)}, \ldots, g_{k-1}^{(\ell)}, g_k^{(\ell-1)}, g_{k+1}^{(\ell-1)}, \ldots, g_d^{(\ell-1)}). \tag{A8}$$

Using Equation (A8) iteratively for $k = 1, \ldots, d$, we get $\Phi(\mu^{(\ell-1)}, \mathbf{g}^{(\ell)}) \leq \Phi(\mu^{(\ell-1)}, \mathbf{g}^{(\ell-1)}) = \upsilon_{\ell-1}$. Finally, using similar arguments as those considered above, we get easily that $\upsilon_\ell = \Phi(\mu^{(\ell)}, \mathbf{g}^{(\ell)}) \leq \Phi(\mu^{(\ell-1)}, \mathbf{g}^{(\ell)})$, so that $\upsilon_\ell \leq \upsilon_{\ell-1}$. ■