Supplementary material to: Robust estimators for additive models using backfitting*

Graciela Boente, Alejandra Martínez and Matías Salibián–Barrera

Abstract

To illustrate the sensitivity of the robust backfitting to the presence of single outliers, Section S.1 reports a numerical study of its empirical influence function. Some additional figures for the real data set are provided in Section S.2. Proofs of Theorems 2.1 to 2.3 are given in Section S.3. In this supplement, Figures and Tables are numbered S.1, S.2, ... References to Sections on the main body of the paper are indicated without the capital "S".

*Graciela Boente (Email: gboente@dm.uba.ar) is Full Professor at the Departamento de Matemáticas from the Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Buenos Aires 1428, Argentina (Tel./FAX: 54-11-45763375). She also has a researcher position at the CONICET. Alejandra Martínez (Email: ale_m_martinez@hotmail.com) has a postdoctoral position at the CONICET and is teaching assistant at the Departamento de Matemáticas from the Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Buenos Aires 1428, Argentina (Tel./FAX: 54-11-45763375). Matías Salibián Barrera (Email: matias@stat.ubc.ca) is Full Professor at the Department of Statistics from the University of British Columbia, 3182 Earth Sciences Building, 2207 Main Mall, Vancouver, BC, V6T 1Z4, Canada (Tel./FAX: 1-604-8223410). This research was partially supported by Grants PIP 112-201101-00339 from CONICET, PICT 2014-0351 from 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.

S.1 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

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

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 (S.1), 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 S.1.

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 S.2, while the results for $\tau=0.6$ and 0.8 are given in Figure S.3.

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 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

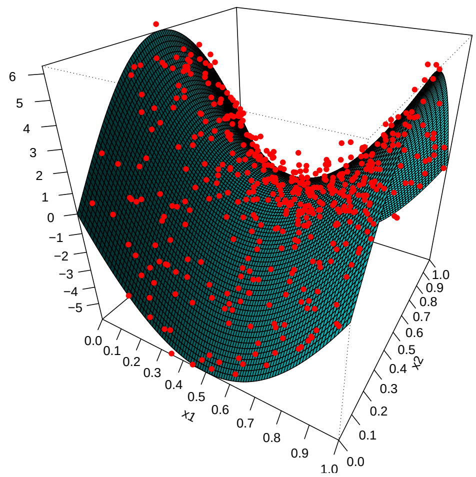

Figure S.1: Data used for the influence function study, and the corresponding regression function g_0.

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 S.4 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 S.5. 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.

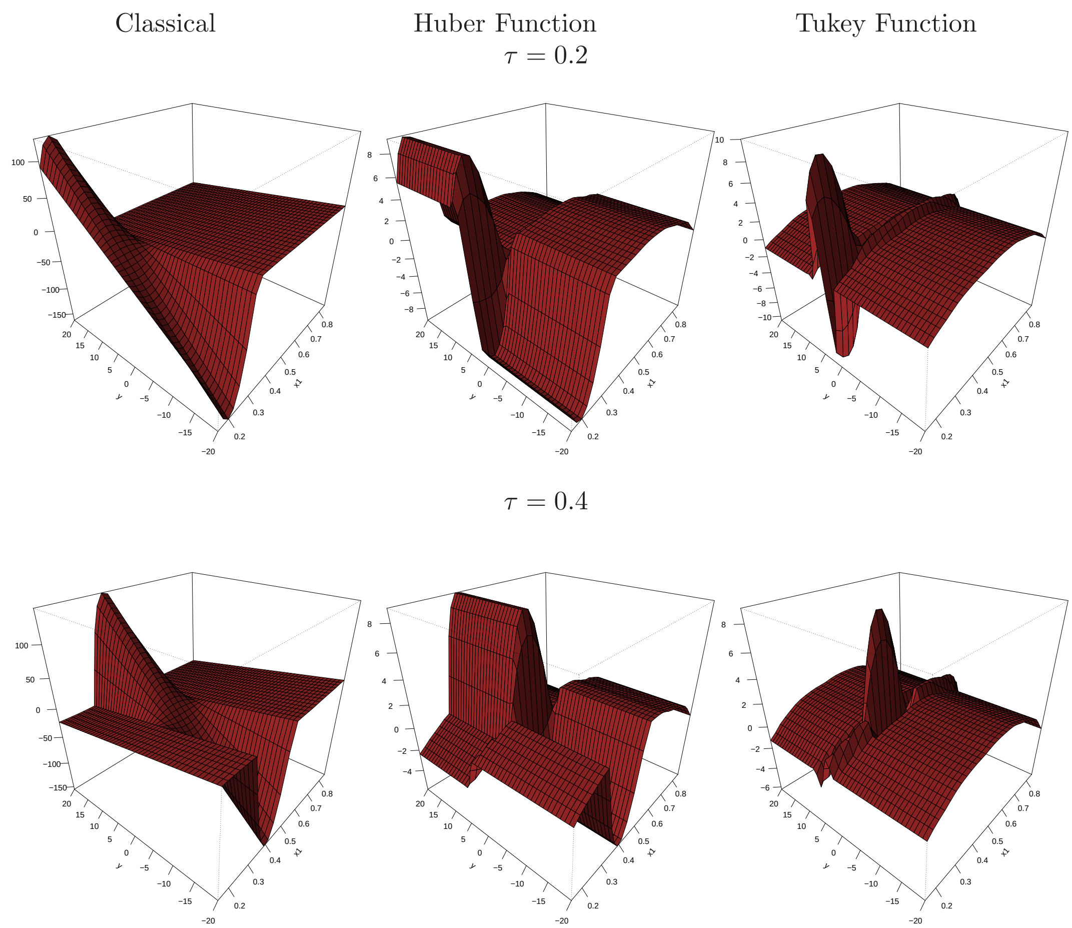

Figure S.2: Empirical influence for the classical and robust estimators, $\mathrm{EIF}_{1,\tau}(\mathbf{x}, y)$ when $\tau = 0.2$ and 0.4 and $\mathbf{x} = (x_1, 0.5)$.

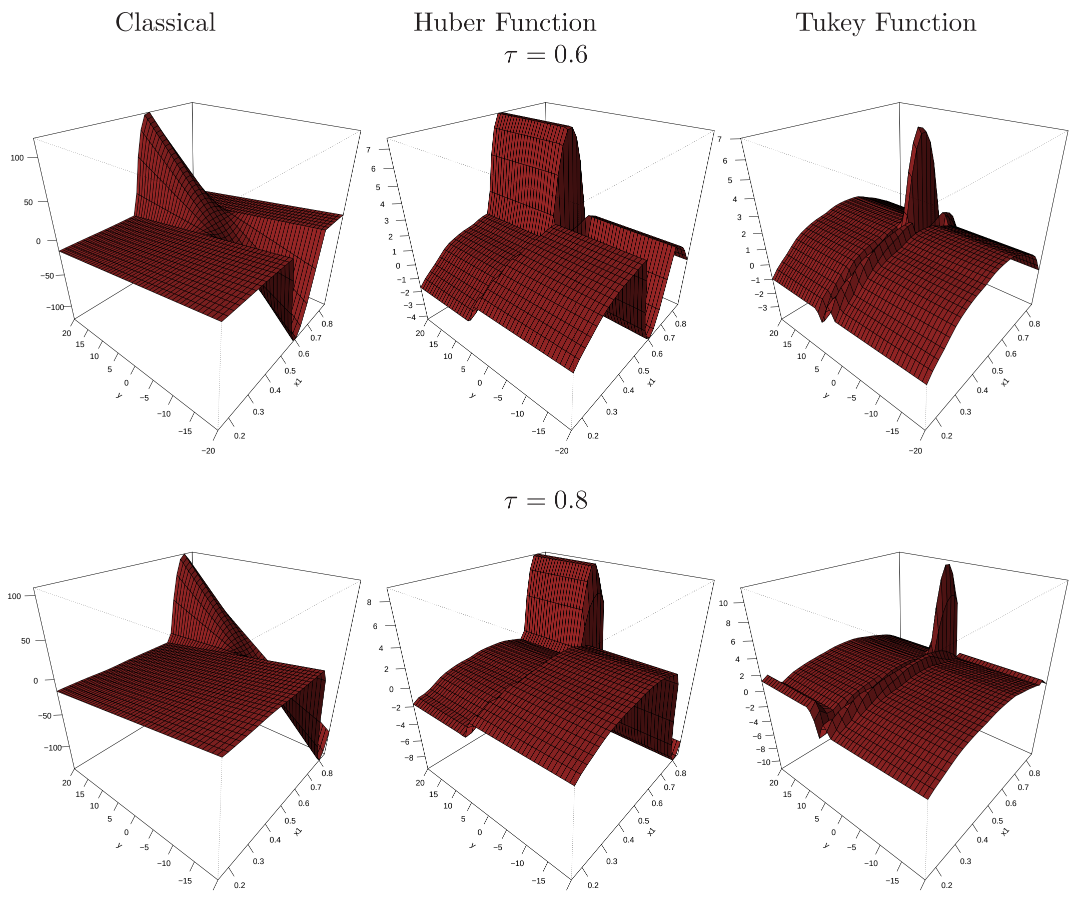

Figure S.3: Empirical influence for the classical and robust estimators, $\mathrm{EIF}_{1,\tau}(\mathbf{x}, y)$ when $\tau = 0.6$ and 0.8 and $\mathbf{x} = (x_1, 0.5)$

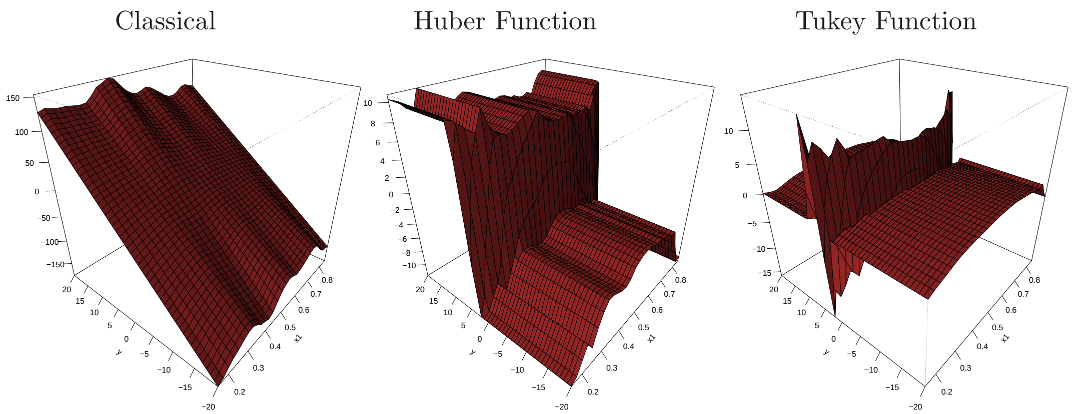

Figure S.4: Empirical influence $\mathrm{EIF}_{1,x_1}((x_1, 0.5), y)$ for the classical and robust estimators.

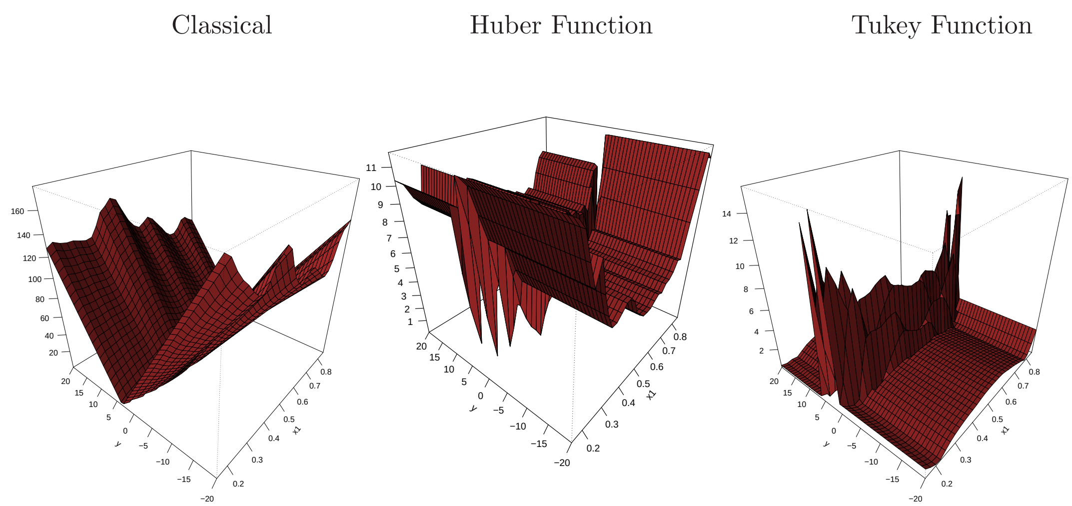

Figure S.5: Absolute value of the empirical influence, $|\mathrm{EIF}_{1,x_1}((x_1, 0.5), y)|$ for the classical and robust estimators.

S.2 Real data example

In this section, we give the plots for the partial residuals obtained using the classical and robust estimators with all the data (Figure S.6) and when using the classical estimators on the data set without the 5 detected atypical observations (Figure S.7).

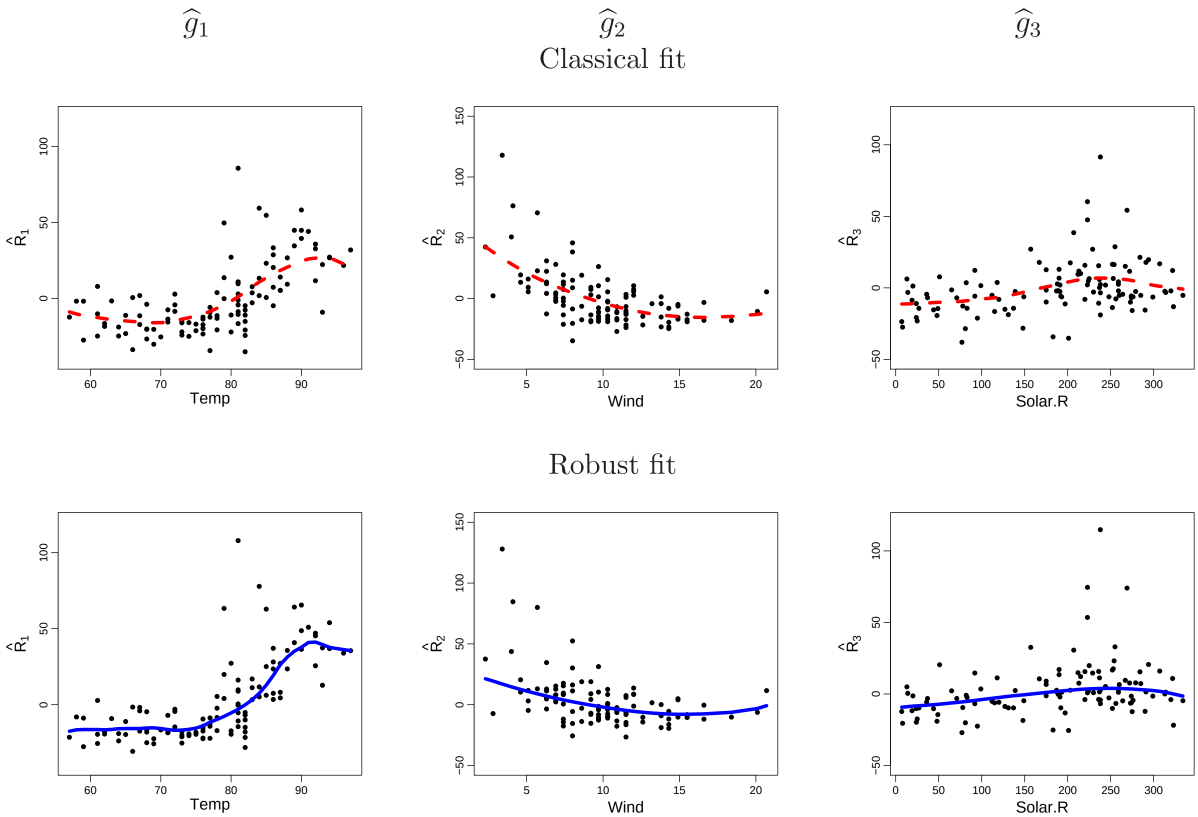

Figure S.6: Partial residuals, $\widehat{R}_j$ for $1 \leq j \leq 3$, and estimated curves for the classical (in red dashed lines) and robust (in blue solid lines) backfitting estimators with data-driven bandwidths $\mathbf{h}_{\text{LS}}$ and $\mathbf{h}_{\text{R}}$, respectively.

S.3 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 $\Upsilon(\nu, m) = \mathbb{E}_{\mathbf{X}}\, \mathbb{E}_{\varepsilon}\left\{\rho\left(\varepsilon - [b(\mathbf{X})/\sigma_0]\right)\right\}$. To simplify the notation, let

$\widehat{g}_1$ $\widehat{g}_2$ $\widehat{g}_3$

Classical fit without the outliers

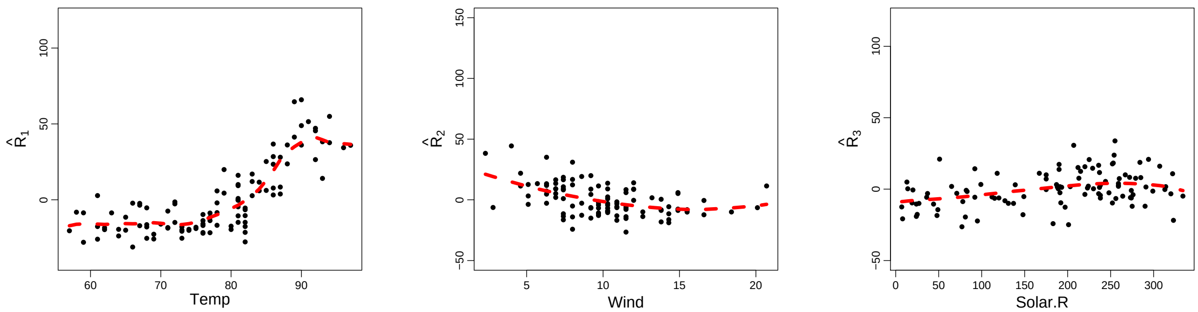

Figure S.7: Partial residuals and estimated curves for the classical backfitting estimator, $\widehat{g}_j^{(-5)}$, (in red dashed lines) with data-driven bandwidth $\mathbf{h}_{\rm LS}^{(-5)}$.

$a(\mathbf{x}) = b(\mathbf{x})/\sigma_0$ and $\mathcal{B}_0 = \{\mathbf{x} : b(\mathbf{x}) = 0\}$. We have

$$\Upsilon(\nu, m) = \int_{\mathcal{B}_0} \mathbb{E}_\varepsilon\left(\rho\left(\varepsilon\right)\right) dF_{\mathbf{X}}(\mathbf{x}) + \int_{\mathcal{B}_0^c} \mathbb{E}_\varepsilon\left(\rho\left(\varepsilon - a(\mathbf{x})\right)\right) dF_{\mathbf{X}}(\mathbf{x})\,. \tag{S.2}$$

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 $\mathbb{P}\left(\mathcal{B}_0\right) < 1$. To see this, assume that $\mathbb{P}(\mathcal{B}_0) = 1$ which implies that $\mathbb{E}\left[b(\mathbf{X})\right] = 0$. Since $\mathbb{E}\left[m_j(X_j)\right] = \mathbb{E}\left[g_{0,j}(X_j)\right] = 0$, 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 $a \neq 0$, $\mathbb{E}_\varepsilon\left[\rho\left(\varepsilon - a\right)\right] > \mathbb{E}_\varepsilon\left[\rho\left(\varepsilon\right)\right]$.

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 $\mathbb{P}\left(\mathcal{B}_0\right) < 1$, and then from (S.2) it follows that

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

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

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

$$\nu_\eta'(0) \;=\; \lim_{t\to 0}\frac{1}{t}\,\mathbb{E}\left[\rho\left(\frac{R_j - g_j(X_j) - t\eta(X_j)}{\sigma_0}\right) - \rho\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\right], \tag{S.3}$$

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 (S.3) 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

$$\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\eta(X_j)\right] = 0\,, \qquad \forall \eta \in \mathcal{H}_j\,. \tag{S.4}$$

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 (S.4) we get that

$$\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)h(X_j)\right] = a_h\mathbb{E}\left[\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right)\right]. \tag{S.5}$$

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

$$\mathbb{E}\psi\left(\frac{R_j - g_j(X_j)}{\sigma_0}\right) = 0\,. \tag{S.6}$$

Therefore, from (S.5) and (S.6), 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 $\Gamma_j(\mu, \mathbf{g}, x_j) = \mathbb{E}\left[\psi\left((R_j - g_j(x_j))/\sigma_0\right)\middle|\, X_j = x_j\right]$. $\square$

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

$$\begin{cases} g_k^{(\ell)}(X_k) = \underset{m_k \in \mathcal{H}_k}{\operatorname{argmin}}\, \mathbb{E}\left[\rho\left(\frac{R_k^{(\ell)} - m_k(X_k)}{\sigma_0}\right)\right] & 1 \le k \le d \\ \mu^{(\ell)} = \underset{\nu \in \mathbb{R}}{\operatorname{argmin}}\, \mathbb{E}\rho\left(\frac{R_0^{(\ell)} - \nu}{\sigma_0}\right) & \end{cases} \tag{S.7}$$

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 $\Phi\left(\nu, \mathbf{m}\right) = \Upsilon(\nu, \mathbf{1}^{\mathrm{T}}\mathbf{m}) = \mathbb{E}\rho\left((Y - \nu - \sum_{j=1}^{d} m_j(X_j))/\sigma_0\right)$, 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

$$\Phi\left(\mu^{(0)}, \mathbf{g}^{(1)}\right) = \Phi\left(\mu^{(1)}, g_1^{(1)}, \dots, g_d^{(1)}\right) \le \Phi\left(\mu^{(0)}, g_1^{(1)}, \dots, g_{d-1}^{(1)}, g_d^{(0)}\right) \le \Phi\left(\mu^{(0)}, \mathbf{g}^{(0)}\right) . \tag{S.8}$$

Finally, using the last equation in (S.7), 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 (S.8). 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 $k-$th equation in (S.7) leads to

$$\Phi\left(\mu^{(\ell-1)}, g_1^{(\ell)}, \ldots, g_k^{(\ell)}, g_{k+1}^{(\ell-1)}, \ldots, g_d^{(\ell-1)}\right) \leq \Phi\left(\mu^{(\ell-1)}, g_1^{(\ell)}, \ldots, g_{k-1}^{(\ell)}, g_k^{(\ell-1)}, g_{k+1}^{(\ell-1)}, \ldots, g_d^{(\ell-1)}\right) . \tag{S.9}$$

Using (S.9) iteratively for $k = 1, \ldots d$, we get $\Phi\left(\mu^{(\ell-1)}, \mathbf{g}^{(\ell)}\right) \leq \Phi\left(\mu^{(\ell-1)}, \mathbf{g}^{(\ell-1)}\right) = \upsilon_{\ell-1}$. Finally, using similar arguments as those considered above, we get easily that $\upsilon_\ell = \Phi\left(\mu^{(\ell)}, \mathbf{g}^{(\ell)}\right) \leq \Phi\left(\mu^{(\ell-1)}, \mathbf{g}^{(\ell)}\right)$, so that $\upsilon_\ell \leq \upsilon_{\ell-1}$. $\square$

References

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

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.

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

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

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