Robust estimators in a generalized partly linear regression model under monotony constraints

Graciela Boente[1], Daniela Rodriguez[1] and Pablo Vena[1]

[1] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET

Abstract

In this paper, we consider the situation in which the observations follow an isotonic generalized partly linear model. Under this model, the mean of the responses is modelled, through a link function, linearly on some covariates and nonparametrically on an univariate regressor in such a way that the nonparametric component is assumed to be a monotone function. A class of robust estimates for the monotone nonparametric component and for the regression parameter, related to the linear one, is defined. The robust estimators are based on a spline approach combined with a score function which bounds large values of the deviance. As an application, we consider the isotonic partly linear log–Gamma regression model. Through a Monte Carlo study, we investigate the performance of the proposed estimators under a partly linear log–Gamma regression model with increasing nonparametric component.

1 Introduction

As is well known, semiparametric models may be introduced when the linear model is insufficient to explain the relationship between the response variable and its associated covariates. This approach has been used to extend generalized linear models to allow most predictors to be modelled linearly while one or a small number of them enter the model nonparametrically. In this paper, we deal with observations $(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i)^{\mathrm{T}}$ satisfying a semiparametric generalized partially linear model, denoted GPLM. To be more precise, we assume that $y_i|(\mathbf{x}_i, t_i) \sim F(., \mu_i, \kappa_0)$ where $\mathrm{V\small{AR}}(y_i|(\mathbf{x}_i, t_i)) = A^2(\kappa_0)\, V^2(\mu_i)$, with A and V known functions and $\mu_i = \mathbb{E}(y_i|(\mathbf{x}_i, t_i)) = \mu(\mathbf{x}_i, t_i)$ is such that

$$\mu(\mathbf{x}, t) = H\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t)\right) , \tag{1}$$

where H^{-1} is a known link function, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown parameter and η_0 is an unknown continuous function with support on a compact interval $\mathcal{I}$, which we will assume equal to $[0, 1]$, without loss of generality. The parameter κ_0 which is usually a nuisance parameter, generally lies on a subset of $\mathbb{R}$, for that reason we will assume that $\kappa_0 \in \mathcal{K}$, where $\mathcal{K} \subset \mathbb{R}$ stands for an open set.

When $H(t) = t$, the generalized partially linear model is simply the well known partly linear regression model, that has been considerably studied, and, in this case, κ_0 is the scale parameter. We refer for instance to Härdle *et al.* (2000). Robust estimators for GPLM have been considered for instance by Boente *et al.* (2006) and by Boente and Rodríguez (2010). However, in this paper, we deal with the situation in which there are constraints on the nonparametric component η_0. More precisely, we will assume that η_0, in model (1), is monotone and for simplicity and without loss of generality non–decreasing. Most studies on generalized partly linear models assume that η_0 is an unspecified smooth function. However, in many applications, monotonicity is a property of the function to be fitted. Some examples when $\boldsymbol{\beta}_0 = \mathbf{0}$ can be found for instance in Ramsay (1988) who studied the relation between the incidence of Down's syndrome and the mother's age; see also He and Shi (1998). In Section 6, we analyse a data set considered in Marazzi and Yohai (2004) which aims to study the relationship between the hospital cost of stay and several explanatory variables, including the length of stay in days which we model non–parametrically. The monotone assumption on η_0 is natural in this data set, since the hospital cost increases the longer the stay.

Most estimation developments under monotone constraints were given under a partly linear regression model and we can mention among others, Huang (2002), Sun *et al.* (2012) who considered estimation under constraints and also Lu (2010) who proposed a sieve maximum likelihood estimator based on $B-$splines. Recently, Lu (2015) considered a spline approach to generalized monotone partial linear models. All these methods are sensitive to outliers and some developments were given under a regression model, that is, when $H(t) = t$ to provide robust estimators. For nonparametric isotonic regression models, He and Shi (1998) and Wang and Huang (2002) proposed a robust isotonic estimate procedure based on the median regression, while, to improve the efficiency, Álvarez and Yohai (2012) considered $M-$estimators for isotonic regression. On the other hand, under a partly linear regression model and following the approach given by Lu (2010), Du *et al.* (2013) consider $M-$estimators based on monotone $B-$splines when η_0 is assumed to be a monotone function, the scale parameter is known and the errors have a symmetric distribution. However, in the hospital data set to be considered in Section 6, the errors follow an asymmetric log–Gamma distribution and the proposal considered in Du *et al.* (2013) is not appropriate. Furthermore, the shape parameter is unknown and needs to be estimated in order to calibrate the robust estimators and to downweight large residuals.

In this paper, we provide a general setting to provide a family of estimators for the regression parameter $\boldsymbol{\beta}_0$ and the monotone regression function η_0 under the GPLM model (1) when the nuisance parameter is unknown. This model includes a partly linear isotonic regression model with unknown scale and a partly linear isotonic log–Gamma regression model with unknown shape parameter, as particular cases. In this sense, we generalize the proposal given in Du *et al.* (2013) by considering a preliminary scale estimator. The paper is organized as follows. Section 2 described the proposed robust estimators. In particular, since our approach is based on $B-$splines, a data–driven robust selection method for the knots is described. Consistency and rates of convergence for the proposed estimators are given in Section 3. The particular case of the log–Gamma model is considered in Section 4, while in Section 5, a numerical study is carried out to examine the small sample properties of the proposed procedures. An application to a real data set is provided in Section 6, while concluding remarks are given in Section 7. Some comments regarding the Fisher–consistency of the proposed estimators are given in Appendix

A, while the proofs of the main results are relegated to Appendix B.

2 The robust estimators

Let $w : \mathbb{R}^p \to \mathbb{R}$ be a weight function to control leverage points on the carriers $\mathbf{x}$ and $\rho : \mathbb{R}^2 \to \mathbb{R}$ a loss function. Define the functions

$$L_n(\boldsymbol{\beta}, g, a) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + g(t_i), a\right) w(\mathbf{x}_i) \tag{2}$$

$$L(\boldsymbol{\beta}, g, a) = \mathbb{E}\rho\left(y_1, \mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta} + g(t_1), a\right) w(\mathbf{x}_1)\ . \tag{3}$$

As in Lu (2010, 2015) and Du *et al.* (2013), consider $\mathcal{T}_n = \{t_i\}_{i=1}^{m_n+2\ell}$ where $0 = t_1 = \cdots = t_\ell < t_{\ell+1} < \cdots < t_{m_n+\ell+1} = \cdots = t_{m_n+2\ell} = 1$ is a sequence of knots that partition the closed interval $[0, 1]$ into m_n+1 subintervals $\mathcal{I}_i = [t_{l+i}, t_{l+i+1})$, for $i = 0, \ldots, m_n-1$ and $\mathcal{I}_{m_n} = [t_{m_n+\ell}, t_{m_n+\ell+1}]$.

Denote as $\mathcal{S}_n(\mathcal{T}_n, \ell)$ the class of splines of order $\ell > 1$ with knots $\mathcal{T}_n$. According to Corollary 4.10 of Schumaker (1981), for any $g \in \mathcal{S}_n(\mathcal{T}_n, \ell)$, there exist a class of $B-$spline basis functions $\{B_j : 1 \leq j \leq k_n\}$, with $k_n = m_n + \ell$, such that $g = \sum_{j=1}^{k_n} \lambda_j B_j$. Furthermore, according to Theorem 5.9 of Schumaker (1981), the spline g is monotonically nondecreasing on $[0, 1]$ if nondecreasing constraints are imposed on the coefficients $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{k_n})^{\mathrm{T}}$, i.e., when $\lambda_1 \leq \cdots \leq \lambda_{k_n}$.

Therefore, we can define a collection of monotone non-decreasing splines on $[0, 1]$, $\mathcal{M}_n(\mathcal{T}_n, \ell)$, which is a subclass $\mathcal{S}_n(\mathcal{T}_n, \ell)$, through

$$\mathcal{M}_n(\mathcal{T}_n, \ell) = \left\{ \sum_{i=j}^{k_n} \lambda_j B_j : \quad \lambda_1 \leq \cdots \leq \lambda_{k_n} \right\},$$

where the non-decreasing constraints are imposed on the coefficients to guarantee monotonicity. Hence, the function η_0 can be approximated as $\eta(t) \approx \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}(t)$ with $\mathbf{B}(t) = (B_1(t), \ldots, B_{k_n}(t))^{\mathrm{T}}$ the vector of $B-$spline basis functions, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{k_n})^{\mathrm{T}}$ the spline coefficient vector such that $\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B} \in \mathcal{M}_n(\mathcal{T}_n, \ell)$.

This suggests that estimators of $(\boldsymbol{\beta}_0, \eta_0)$ may be obtained minimizing $L_n(\boldsymbol{\beta}, g, \widehat{\kappa})$ over $\boldsymbol{\beta} \in \mathbb{R}^p$ and $g \in \mathcal{M}_n(\mathcal{T}_n, \ell)$, where $\widehat{\kappa}$ is a robust consistent estimator of κ_0, for instance, previously computed without the monotonicity constraint. More precisely, the estimators $(\widehat{\boldsymbol{\beta}}, \widehat{\eta}) =$

$(\widehat{\boldsymbol{\beta}}, \sum_{j=1}^{k_n} \widehat{\lambda}_j B_j) = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{B})$ are defined through the values $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}})$ such that

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\lambda}\in\mathcal{L}_{k_n}}{\operatorname{argmin}} \; L_n\left(\boldsymbol{\beta}, \sum_{j=1}^{k_n} \lambda_j \, B_j, \widehat{\kappa}\right), \tag{4}$$

where $\mathcal{L}_{k_n} = \{\boldsymbol{\lambda} \in \mathbb{R}^{k_n} : \lambda_1 \leq \cdots \leq \lambda_{k_n}\}$. If we denote $\mathbf{B}_i = (B_1(t_i), \ldots, B_{k_n}(t_i))$, we have that

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\lambda}\in\mathcal{L}_{k_n}}{\operatorname{argmin}} \; \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_i^{\mathrm{T}}\boldsymbol{\lambda}, \widehat{\kappa}\right) w(\mathbf{x}_i)\,. \tag{5}$$

Let $\mathcal{G} = \{g : g$ is a monotonically nondecreasing function on $[0,1]\}$. Throughout the paper, we will assume Fisher–consistency, i.e.,

$$L(\boldsymbol{\beta}_0, \eta_0, \kappa_0) = \min_{\boldsymbol{\beta}\in\mathbb{R}^p, g\in\mathcal{G}} L(\boldsymbol{\beta}, g, \kappa_0)\,, \tag{6}$$

with $(\boldsymbol{\beta}_0, \eta_0)$ being the unique minimum, that is, $L(\boldsymbol{\beta}_0, \eta_0, \kappa_0) < L(\boldsymbol{\beta}, g, \kappa_0)$ for any $(\boldsymbol{\beta}, g) \in \mathbb{R}^p \times \mathcal{G}$, $(\boldsymbol{\beta}, g) \neq (\boldsymbol{\beta}_0, \eta_0)$. This is a usual condition in robustness and it states that our target are indeed the true parameters of the model. A similar condition for generalized linear models was required in Bianco *et al.* (2013a) and for generalized partial linear models in Boente *et al.* (2006) and Boente and Rodríguez (2010) who provide conditions ensuring that $L(\boldsymbol{\beta}_0, \eta_0, \kappa_0) = \min_{\boldsymbol{\beta}\in\mathbb{R}^p, g\in\mathcal{G}} L(\boldsymbol{\beta}, \eta, \kappa_0)$.

Remark 2.1. As mentioned in Lu (2015), if $\lambda_1 \leq \cdots \leq \lambda_{k_n}$, the function $g = \sum_{j=1}^{k_n} \lambda_j B_j$ is non–decreasing, but the linear inequality constraint on the coefficients is not a necessary condition. However, for quadratic $B-$splines, the coefficients condition is sufficient and necessary for monotonicity.

2.1 The loss function

Under a fully parametric generalized linear model, the selected loss function ρ aims to bound either large values of the deviance or of the Pearson residuals. We refer to Bianco and Yohai (1996), Croux and Haesbroeck (2003), Bianco *et al.* (2005) and Cantoni and Ronchetti (2001), where different choices for the loss function are given. On the other hand, optimally bounded score functions have been studied in Stefanski *et al.* (1986). We briefly remind the definition of the family which bounds the deviance which is the function used in our simulation study, for more details see, for instance, Boente *et al.* (2006) who considered this family of loss functions to estimate the parameters of a generalized partial linear model using a profile–kernel approach.

Let φ_a be a bounded non–decreasing function with continuous derivative φ_a', a being the tuning constant. Typically, φ_a is a function performing like the identity function in a neighbourhood of 0 but bounding large values of the deviance. Denote as $f(\cdot, s)$ the density of the distribution function $F(\cdot, s)$ with $y|(\mathbf{x}, t) \sim F\left(\cdot, H\left(\eta(t) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)$. In this setting, the *robust deviance–based estimator* are related to the following choice for the function $\rho(y, u, a)$

$$\rho(y, u, a) = \varphi_a[-\log\, f(y, H(u)) + \log\, f(y, y)] + G_a(H(u))\ . \tag{7}$$

The correction term G_a is given by

$$G_a'(s) = \mathbb{E}_s\left(\varphi_a'[-\log\, f(y, s) + \log f(y, y)]\, \frac{f'(y, s)}{f(y, s)}\right)\ ,$$

where $\mathbb{E}_s$ indicates expectation taken under $y \sim F(\cdot, s)$ and $f'(y, s)$ is a shorthand for $\partial\, f(y, s)/\partial s$. It is worth noticing that $\varphi_a(s) = s$, $G_a(u) = 0$ and $w \equiv 1$ when considering the maximum likelihood estimator, under a generalized linear model. For a general function φ_a, the correction factor is included to guarantee Fisher–consistency under the true model, as for generalized linear models. If the correction factor is taken equal to 0, the results stated in Section 3 only ensure that the estimators will be consistent to the minimizer $(\boldsymbol{\beta}_F, \eta_F)$ of $L(\boldsymbol{\beta}, g, \kappa_0)$, where $L(\boldsymbol{\beta}, g, a)$ is defined in (3). However, as discussed in Bianco *et al.* (2005), when considering a continuous family of distributions with strongly unimodal density function, the correction term G_a can be avoided. In this case, κ_0 may play the role of the tuning constant. For instance, for the Gamma distribution, the tuning constant depends on the shape parameter so, if the shape is unknown, initial estimators need to be considered. Further details are given in Section 4.

Note that for the Poisson and logistic regression models, we have $\kappa_0 = 1$, so κ_0 does not need to be estimated, hence $\varphi_a(s) = \varphi(s)$. Furthermore, as noted by Croux and Haesbroeck (2003) for the logistic model, in order to guarantee existence of solution, beyond the overlapping condition required for the maximum likelihood estimator, the derivative φ' of the function $\varphi(s)$ must satisfy additional constraints. More precisely, φ' needs to be increasing on $(-\infty, A_0]$ and decreasing on $[A_0, +\infty)$ for some $A_0 > 0$ or increasing on $\mathbb{R}$ and also to fulfil that $\lim_{s\to+\infty} \varphi'(st)/\varphi'(-s) = \infty$ for any $t > 0$. An example of function φ satisfying these conditions is also given therein.

On the other hand, when $H(u) = u$, the usual square loss function is replaced by a $\rho-$function after scaling the residuals to control the effect of large responses. More precisely, let $\phi : \mathbb{R} \to [0, \infty)$ stands for $\rho-$function as defined in Maronna *et al.* (2006), i.e., an even continuous, non-decreasing function with $\phi(0) = 0$ and such that $\phi(u) < \phi(v)$ when $0 \leq u < v$ with $\phi(v) < \sup_s \phi(s)$. Then, when the link function equals to identity function $\rho(y, u, a) = \phi((y - u)/a)$ and, as mentioned in the Introduction, κ_0 plays the role of the scale parameter.

Remark 2.2. a) As noted in Boente *et al.* (2006), under a logistic partially linear regression model, Fisher–consistency can easily be derived for the loss function given by (7), when φ satisfies the regularity conditions stated in Bianco and Yohai (1996), $w(\mathbf{x}) > 0$, for all $\mathbf{x}$, and

$$\mathbb{P}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = a_0 | t = t_0\right) < 1, \qquad \forall (\boldsymbol{\beta}, a_0) \neq 0 \text{ and for almost all } t_0. \tag{8}$$

Moreover, taking conditional expectations with respect to $(\mathbf{x}, t)$, it is easy to verify that $(\boldsymbol{\beta}_0, \eta_0)$ is the unique minimizer of $L(\boldsymbol{\beta}, g, \kappa_0)$ in this case. Condition (8) does not allow $\boldsymbol{\beta}_0$ to include an intercept, so that the model will be identifiable.

b) Under a generalized partially linear model with responses having a gamma distribution, Theorem 1 of Bianco *et al.* (2005) allows us to derive Fisher–consistency for the nonparametric and parametric components, if the score function is bounded and strictly increasing on the set where it is not constant and if (8) holds (see Section 4).

c) Finally, consider the partially linear model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t_i) + \epsilon_i$ where ϵ_i are independent of $(\mathbf{x}_i, t_i)$, that is, the link function equals $H(u) = u$. In this case, Fisher–consistency holds if, for instance, the errors ϵ_i have a symmetric distribution with density strictly unimodal, the loss function equals $\rho(y, u, a) = \phi((y-u)/a)$ with ϕ a $\rho-$function as defined in Maronna *et al.* (2006), i.e., an even continuous, non-decreasing function with $\phi(0) = 0$ and such that $\phi(u) < \phi(v)$ when $0 \leq u < v$ with $\phi(v) < \sup_s \phi(s)$. Furthermore, we also have that $L(\boldsymbol{\beta}_0, \eta_0, a) = \min_{\boldsymbol{\beta} \in \mathbb{R}^p, g \in \mathcal{G}} L(\boldsymbol{\beta}, g, a)$, for any $a > 0$, see Appendix A for a proof.

2.2 Selection of k_n

A remaining question is the choice of the number knots and their location for the space of $B-$splines. Knot selection is more important for the estimate of η_0 than for the estimate of $\boldsymbol{\beta}_0$. One approach is to use uniform knots which is the approach followed in our simulation study. Uniform knots are usually sufficient when the function η_0 does not exhibit dramatic changes in its derivatives. On the other hand, non–uniform knots are desirable when the function has very different local behaviours in different regions. Another commonly used approach is to consider as knots quantiles of the observed t_i with uniform percentile ranks.

The number of knots m_n or equivalently the number of elements of the basis (recall that $k_n = m_n + \ell$) may be determined by a model selection criterion. Suppose that $(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{\lambda}}^{(k)})$ is the estimator solution of (4) with a $k-$dimensional spline space. As in He and Shi (1996) and He *et al.* (2002), for each k define a criterion analogous to Schwartz (1978) information criterion

$$BIC(k) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k)} + \sum_{j=1}^{k} \widehat{\lambda}_j^{(k)}\, B_j(t_i), \widehat{\kappa}\right) w(\mathbf{x}_i) + \frac{\log n}{2\, n}(k+p)\,.$$

Large values of BIC indicate poor fits. A robust version of the Akaike criterion considered in Lu (2015) can also be considered. As is usual in spline–based procedures the number of knots should increase slowly with the sample size n to attain an optimal rate of convergence. When it is assumed that η is twice continuously differentiable and cubic splines ($\ell = 3$) are considered, as in our simulation study, according to the convergence rate derived in Theorem 3.2, a possible criterion is to search for the first (i.e. smallest k) local minimum of $BIC(k)$ in the range of $\max(n^{1/5}/2, 4) \le k \le 8 + 2\, n^{1/5}$. Within this range, there is usually only one local minimum. The reason for k being larger than 4 is that for cubic splines the smallest possible choice is 4. Also note that the global minimum of $BIC(k)$ actually occurs at a saturated model in which $k = n - p$, so $BIC(k)$ is a valid criterion only for a limited range of k.

3 Consistency

In this section, we will derive, under some regularity conditions, consistency and rates of convergence for the estimators defined in the previous Section. We will begin by fixing some notation.

Let $\|\cdot\|$ the Euclidean norm of $\mathbb{R}^p$ and $\|f\|_2^2 = \left(\mathbb{E} f^2(t_1)\right)^{1/2}$. For any continuous function $v : \mathbb{R} \to \mathbb{R}$ denote $\|v\|_\infty = \sup_t |v(t)|$ and $\mathcal{G} = \{g : g$ is a monotonically nondecreasing function on $[0,1]\}$. From now on, $\mathcal{V}$ stands for a neighbourhood of κ_0 with closure $\overline{\mathcal{V}}$ strictly included in $\mathcal{K}$ and $\mathcal{F}_n$ will denote the family of functions

$$\mathcal{F}_n = \{f(y, \mathbf{x}, t) = \rho\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}(t), a\right) w(\mathbf{x}), \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\lambda} \in \mathcal{L}_{k_n}, a \in \mathcal{V}\}\,.$$

Furthermore, for any measure Q, $N(\epsilon, \mathcal{F}_n, L_s(Q))$ and $N_{[\,]}(\epsilon, \mathcal{F}_n, L_s(Q))$ stand for the covering and bracketing numbers of the class $\mathcal{F}_n$ with respect to the distance in $L_s(Q)$, as defined, for instance, in van der Vaart and Wellner (1996).

3.1 Consistency results

To derive the consistency of our proposal in the general framework we are considering, we will need the following set of assumptions whose validity is discussed in Remark 3.1.

C0. The estimators $\widehat{\kappa}$ of κ_0 are strongly consistent.

C1. $\rho(y, u, a)$ and $w(\cdot)$ are non–negative and bounded functions and $\rho(y, u, a)$ is a continuous function. Moreover, $L^\star(\boldsymbol{\beta}, \boldsymbol{\lambda}, a) = L(\boldsymbol{\beta}, \sum_{j=1}^{k_n} \lambda_j B_j, a)$ satisfies the following equicontinuity condition: for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $a_1, a_2 \in \overline{\mathcal{V}}$,

$$|a_1 - a_2| < \delta \Rightarrow \sup_{\boldsymbol{\beta} \in \mathbb{R}^k, \boldsymbol{\lambda} \in \mathcal{L}_{k_n}} |L^\star(\boldsymbol{\beta}, \boldsymbol{\lambda}, a_1) - L^\star(\boldsymbol{\beta}, \boldsymbol{\lambda}, a_2)| < \epsilon\,.$$

C2. The true function η_0 is nondecreasing and its $r-$th derivative satisfies a Lipschitz condition on $[0,1]$, with $r \geq 1$, that is,

$$\eta_0 \in \mathcal{H}_r = \{g \in C^r[0,1] : \|g^{(j)}\|_\infty \leq C_1,\ 0 \leq j \leq r \ \text{ and }\ |g^{(r)}(z_1) - g^{(r)}(z_2)| \leq C_2 |z_1 - z_2|\}\,.$$

C3. The maximum spacing of the knots is assumed to be of order $O(n^{-\nu})$, $0 < \nu < 1/2$. Moreover, the ratio of maximum and minimum spacings of knots is uniformly bounded.

C4. The class of functions $\mathcal{F}_n$ is such that, for any $0 < \epsilon < 1$, $\log\left(N(\epsilon, \mathcal{F}_n, L_1(P_n))\right) = O_{\mathbb{P}}(1)(k_n) \log(1/\epsilon)$, for some constant $C_1 > 0$ independent of n and ϵ.

For simplicity, denote as $L(\boldsymbol{\theta}_0, \kappa_0) = L(\boldsymbol{\beta}_0, \eta_0, \kappa_0)$, where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \eta_0)$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\eta})$ the estimators defined through (4) with $\widehat{\eta}(t) = \sum_{j=1}^{k_n} \widehat{\lambda}_j B_j(t)$. To measure the closeness between the estimators and the parameters, consider the metric $\pi^2(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}) = \|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\|^2 + \|\eta_0 - \widehat{\eta}\|_{\mathcal{F}}^2$ where $\|\cdot\|_{\mathcal{F}}$ stands for a norm in the space of functions $\mathcal{F} = \{g : [0,1] \to \mathbb{R},$ such that g is a continuous function$\}$, such as $\|f\|_2 = \left(\mathbb{E} f^2(t_1)\right)^{1/2}$ or $\|f\|_\infty = \sup_{t\in[0,1]} |f(t)|$. Let $\mathcal{A}_\epsilon = \{\boldsymbol{\theta} = (\boldsymbol{\beta}, g) : \boldsymbol{\beta} \in \mathbb{R}^p, g \in \mathcal{G} \cap \mathcal{F}, \pi(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \epsilon\}$.

Theorem 3.1. *Let $(y_i, \mathbf{x}_i, t_i)^{\mathrm{T}}$ be i.i.d. observations satisfying* (1)*. Assume that **C0** to **C4** hold and that for any $\epsilon > 0$, $\inf_{\boldsymbol{\theta}\in\mathcal{A}_\epsilon} L(\boldsymbol{\theta}, \kappa_0) > L(\boldsymbol{\theta}_0, \kappa_0)$ and that $k_n = O(n^\nu)$ for $1/(2r+2) < \nu < 1/(2r)$. Then, we have that $\pi(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}) \xrightarrow{a.s.} 0$.*

Remark 3.1. As mentioned above, for the logistic and Poisson model, κ_0 is known and does not need to be estimated, hence **C0** may be omitted. On the other hand, when $H(t) = t$ the scale parameter κ_0 may be estimated using any robust scale estimator computed without using the monotone constraint. To be more precise, let $(\widehat{\boldsymbol{\beta}}, \widehat{\eta})$ be the robust estimators of $(\boldsymbol{\beta}_0, \eta_0)$ defined in Bianco and Boente (2004) and define the residuals as $r_i = y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \widehat{\eta}(t_i)$. The scale estimator $\widehat{\kappa}$ can be taken as $\text{median}_{1\le i\le n} |r_i|$. Another possibility is to consider a scale estimator based on a $\rho-$function as follows. As in Maronna *et al.* (2006), let $\chi : \mathbb{R} \to \mathbb{R}_+$ be a $\rho-$function, that is, an even function, non–decreasing on $|t|$, increasing for $t > 0$ when $\chi(t) < \|\chi\|_\infty$ and such that $\chi(0) = 0$. The estimator $\widehat{\kappa}$ of the scale κ_0 is the solution

$$\frac{1}{n}\sum_{i=1}^{n} \chi_c\left(\frac{r_i}{s}\right) = b\,, \tag{9}$$

where $\chi_c(u) = \chi(u/c)$, $c > 0$ is a user–chosen tuning constant and b is related to the breakdown point of the scale estimator. If χ is bounded, it is usually assumed that $\|\chi\|_\infty = 1$ in which case $0 < b < 1$. For instance, when χ is the Tukey's biweight function, the choice $c = 1.54764$ and $b = 1/2$ leads to an scale estimator Fisher–consistent at the normal distribution with breakdown point 0.5. On the other hand, the choice $\chi(t) = \mathbb{I}_{(1,\infty)}(|t|)$, $c = 1$ and $b = 0.5$ leads to $\text{median}_{1\le i\le n} |r_i|$. Similarly, when the responses have a Gamma distribution the parameter κ_0 corresponds to the tuning constant and is related to the shape parameter. It can be estimated using a preliminary $S-$estimator computed without making use of the monotone restriction, as described in Section 4. Straightforward calculations allow to show that in both situations **C0** holds.

Assumption **C1** is a standard requirement since it states that the weight function controls large values of the covariates and that the score function bounds large residuals, respectively. Moreover, the equicontinuity requirement allows to deal with the nuisance parameter in a general setting and a similar condition appears in Bianco *et al.* (2013a). For the particular case of a partly linear regression model, i.e., when $H(t) = t$, κ_0 is the scale parameter and the function $\rho(y, u, a)$ is usually chosen as $\rho(y, u, a) = \phi((y - u)/a)$ where the function ϕ is an even, bounded function, non–decreasing on $(0, \infty)$. In this case, the equicontinuity condition is satisfied, for instance, if ϕ is continuously differentiable with first derivative ϕ' such that $s\,\phi'(s)$ is bounded.

C2 and **C3** are conditions regarding the smoothness of the nonparametric component and the knots spacing. They are analogous to those considered, for instance, in Lu (2010, 2015). On the other hand, the requirement $\inf_{\boldsymbol{\theta}\in\mathcal{A}_\epsilon} L(\boldsymbol{\theta}, \kappa_0) > L(\boldsymbol{\theta}_0, \kappa_0)$ ensures that $L(\boldsymbol{\theta}_0, \kappa_0)$ does not attain a minimum value at infinite. It was also a requirement in Boente *et al.* (2006) and Boente and Rodríguez (2010) to guarantee strong consistency. It can be replaced by the condition that $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}})$ lie ultimately in a compact set since $(\boldsymbol{\beta}_0, \eta_0)$ is the unique minimizer of $L(\boldsymbol{\beta}, g, \kappa_0)$ as stated in (6).

Assumption **C4** is satisfied for most loss functions ρ. Effectively, assume that κ_0 is known and that the densities are such that the covering number of the class

$$\mathcal{F}_0 = \{g(y, \mathbf{x}) = \log f\left(y, H\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}\right)\right), \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\lambda} \in \mathbb{R}^{k_n}\}$$

grows at a polynomial rate, i.e., it is bounded by $A\epsilon^{-(k_n+p+1)}$. Then, if the functions $\varphi(s)$ and $G(H(s))$ are of bounded variation, we obtain the result using that $N(\epsilon, \mathcal{H}_1 + \mathcal{H}_2, L_r(\mathbb{Q})) \leq N(\epsilon/2, \mathcal{H}_1, L_r(\mathbb{Q}))\, N(\epsilon/2, \mathcal{H}_2, L_r(\mathbb{Q}))$. A similar bound can be obtained for the bracketing numbers. For the score functions usually considered in robustness, such as the Tukey's biweight function or the score function introduced in Croux and Haesbroeck (2002) for the logistic model, φ and $G(H(s))$ have bounded variation and the required condition is easily verified using the permanence properties of $VC-$classes of functions since the class $\{\mathbf{x}^{\mathrm{T}}\mathbf{b} + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}, \mathbf{b} \in \mathbb{R}^p, \boldsymbol{\lambda} \in \mathbb{R}^{k_n}\}$ is a finite–dimensional class and so a $VC-$class. Furthermore, if κ_0 plays the role of the tuning constant or the scale parameter, as in the Gamma model or when $H(t) = t$ and the errors have a symmetric distribution, the same conclusions hold.

3.2 Convergence rates

In order to derive rates of convergence for the estimators, we choose as norm $\|\cdot\|_{\mathcal{F}}$ in the space of functions $\mathcal{F}$, the $L^{\wp}(Q)$ norm, with $2 \leq \wp \leq \infty$, where $t \sim Q$. Hence, we include as possible norms $\|f\|_{\mathcal{F}}^2 = \|f\|_2^2 = \mathbb{E}f^2(t)$ or $\|f\|_{\mathcal{F}}^2 = \|f\|_\infty$, in which case $\pi^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + \|\eta_1 - \eta_2\|_{\wp}^2$ with $\wp = 2$ or $\wp = \infty$, respectively. Furthermore, in this setting we define the distance

$$\pi_{\mathbb{P}}^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}\left(w(\mathbf{x}) \left[\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) + \eta_1(t) - \eta_2(t) \right]^2 \right) ,$$

where for $j = 1, 2$, $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \eta_j) \in \Theta = \mathbb{R}^p \times \mathcal{G}$.

We consider the following additional assumptions. Two possible conditions on the bracketing entropy are stated below and according to them weaker or stronger convergence rates are attained. Conditions under which they hold for some particular models are given in Remark 3.2.

To avoid requiring an order of consistency to the estimator $\widehat{\kappa}$ of κ_0, from now on we will assume that $L(\boldsymbol{\beta}_0, \eta_0, a) < L(\boldsymbol{\beta}, g, a)$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and $g \in \mathcal{M}_n(\mathcal{T}_n, \ell)$, $a \in \mathcal{V}$ such that $(\boldsymbol{\beta}, g) \neq (\boldsymbol{\beta}_0, \eta_0)$. This condition clearly entails Fisher–consistency and holds, for instance, for the log–partly linear regression model and when $H(t) = t$ if the errors have a symmetric distribution.

From now on, for $\boldsymbol{\lambda} \in \mathbb{R}^{k_n}$, $g_{\boldsymbol{\lambda}}(t)$ stands for the spline function $g_{\boldsymbol{\lambda}}(t) = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}(t)$.

C5*. Let $\mathcal{G}_{n,c,\boldsymbol{\lambda}_0} = \{ f(y, \mathbf{x}, t) = [\rho(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g_{\boldsymbol{\lambda}}(t), a) - \rho(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + g_{\boldsymbol{\lambda}_0}(t), a)]\, w(\mathbf{x}) \,,\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \epsilon_0 \,,\ \boldsymbol{\lambda} \in \mathcal{L}_{k_n}, a \in \mathcal{V}, \pi_{\mathbb{P}}((\boldsymbol{\beta}_0, g_{\boldsymbol{\lambda}_0}(t)), \boldsymbol{\theta}) \leq c \}$. For some constant $C_2 > 0$ independent of n, $\boldsymbol{\lambda}_0 \in \mathcal{L}_{k_n}$ and ϵ, we have that $N_{[\,]}(\epsilon, \mathcal{G}_{n,c,\boldsymbol{\lambda}_0}, L_2(P)) \leq C_2\, (c/\epsilon)^{k_n+p+1}$.

C5**. For $n \geq n_0$, the family of functions $\mathcal{F}_{n,c}^\star = \{ f(y, \mathbf{x}, t) = \rho(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g_{\boldsymbol{\lambda}}(t), a)\, w(\mathbf{x}), \boldsymbol{\lambda} \in \mathcal{L}_{k_n}, a \in \mathcal{V}, \pi(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \leq c \}$ is such that for any $0 < \epsilon < 1$, $N_{[\,]}(\epsilon, \mathcal{F}_{n,c}^\star, L_2(P)) \leq C_2 / \epsilon^{k_n+p+1}$, for some constant $C_2 > 0$ independent of n and ϵ.

C6. a) The function ρ is twice continuously differentiable with respect to its second argument with derivatives $\Psi(y, u, a) = \partial\rho(y, u, a)/\partial u$ and $\chi(y, u, a) = \partial\Psi(y, u, a)/\partial u$ such that

$$\|\Psi\|_{\infty,\mathcal{V}} = \sup_{y\in\mathbb{R}, u\in\mathbb{R}, a\in\mathcal{V}} |\Psi(y, u, a)| < \infty \text{ and } \|\chi\|_{\infty,\mathcal{V}} = \sup_{y\in\mathbb{R}, u\in\mathbb{R}, a\in\mathcal{V}} |\chi(y, u, a)| < \infty \,.$$

b) $\mathbb{E}\left\{\Psi\left(y_1, \mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t_1), a\right)|(\mathbf{x}_1, t_1)\right\} = 0$, almost surely, for any $a \in \mathcal{V}$.

C7. $\mathbb{E}w(\mathbf{x}_1)\,\|\mathbf{x}_1\|^2 < \infty$.

C8. There exists $\epsilon_0 > 0$ and a positive constant C_0, such that for any $\boldsymbol{\theta} \in \mathbb{R}^p \times \mathcal{M}_n(\mathcal{T}_n, \ell)$ with $\pi^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < \epsilon_0$ and any $a \in \mathcal{V}$, $L(\boldsymbol{\theta}, a) - L(\boldsymbol{\theta}_0, a) \geq C_0\, \pi^2_{\mathbb{P}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$.

Theorem 3.2. *Let $(y_i, \mathbf{x}_i, t_i)^{\mathrm{T}}$ be i.i.d. observations satisfying (1) and $k_n = O(n^{\nu})$ for $1/(2r+2) < \nu < 1/(2r)$. Assume that **C1** to **C3** and **C6** to **C8** hold and that $\pi(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \xrightarrow{a.s.} 0$. Then, we have that*

*a) if **C5*** holds, $\gamma_n\, \pi_{\mathbb{P}}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$, where $\gamma_n = n^{\min(r\nu, (1-\nu)/2)}$, so if $\nu = 1/(1+2r)$, the estimators converge at the optimal rate $n^{r/(1+2r)}$.*

*b) if **C5**** holds, $\gamma_n\, \pi_{\mathbb{P}}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$, for any γ_n, such that $\gamma_n \leq O(n^{r\nu})$ and $\gamma_n \log(\gamma_n) \leq O(n^{(1-\nu)/2})$.*

Remark 3.2. Note that condition **C6**b) is analogous to the conditional Fisher–consistency stated in Kunsch *et al.* (1989), while condition **C5*** is analogous to assumption C3$'$ in Shen and Wong (1994). Similar arguments to those considered in Shen and Wong (1994) when analysing the Case 3 in page 596, allow to show that **C5*** holds, for instance, when $H(t) = t$ when ϕ is continuously differentiable with first derivative ϕ' such that $s\,\phi'(s)$ is bounded. It also holds for the logistic model and for the gamma model when $w(\mathbf{x})\,\|\mathbf{x}\|^2$ is bounded using **C6**a).

4 The log–Gamma regression model

Among generalized linear models, the Gamma distribution with a log–link, usually denoted log–Gamma regression, plays an important role, see Chapter 8 of McCullagh and Nelder (1989). For any $\alpha > 0$ and $\mu > 0$, denote as $\Gamma(\alpha, \mu)$ the parametrization of the Gamma distribution given by the density

$$f(y, \alpha, \mu) = \alpha^{\alpha}\, y^{\alpha-1}\, \exp\left(-(\alpha/\mu)y\right)\{\mu^{\alpha}\,\Gamma(\alpha)\}^{-1}\, I_{y\geq 0}\,.$$

Under a log–Gamma model, $y_i|\mathbf{x}_i \sim \Gamma(\alpha, \mu_i)$, where $\mu_i = \mathbb{E}(y_i|(\mathbf{x}_i, t_i))$ with link function $\log(\mu_i) = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i + \eta_0(t_i)$. As it is well known, in this case, the responses can be transformed so that they are modelled through a linear regression model with asymmetric errors (see for instance Cantoni and Ronchetti, 2006). Let $z_i = \log(y_i)$ be the transformed responses, then

$$z_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \eta_0(t_i) + u_i\,, \tag{10}$$

where u_i and $(\mathbf{x}_i, t_i)$ are independent. Moreover, $u_i \sim \log(\Gamma(\alpha, 1))$ with density

$$g(u,\alpha) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)}\, \exp\left[\alpha(u - \exp(u))\right]\,. \tag{11}$$

This density is asymmetric and unimodal with maximum at $u_0 = 0$. For fully parametric linear models. i.e., when $\eta_0(t) = \gamma_0\, t$, a description on robust estimators based on deviances was given in Bianco *et al.* (2005), while Heritier *et al.* (2009) considered $M-$type estimators based on Pearson residuals. For the sake of completeness, we will describe how to adapt the estimators based on deviances to the present situation.

We will consider the transformed model (10) and denote by $d_i(\boldsymbol{\beta}_0, \eta_0, \alpha)$ the deviance component of the i-th observation, i.e.,

$$d_i(\boldsymbol{\beta}_0, \eta_0, \alpha) = 2\alpha\; d\left(z_i - \left[\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \eta_0(t_i)\right]\right)$$

where $d(u) = \exp(u) - u - 1$.

In this setting, the classical estimators to be considered below are not based on the quasi–likelihood but on the deviance and they correspond to the choice $\varphi_a(u) = \varphi(u) = u$ in (7), since no tuning constant is needed. Thus, the loss function equals $\rho(z,s) = d(z-s)$, while $\Psi(z,s) = \partial\rho(z,s)/\partial s = 1 - \exp(z-s)$, $\chi(z,s) = \partial\Psi(z,s)/\partial s = \exp(z-s)$. Hence, if $\mathbf{B}(t) = (B_1(t), \dots, B_{k_n}(t))$, the classical estimators of $(\boldsymbol{\beta}_0, \eta_0)$ without any restriction are obtained as $(\widehat{\boldsymbol{\beta}}, \widehat{\eta})$ where $\widehat{\eta}(t) = \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{B}(t)$ with

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}) = \underset{\boldsymbol{\beta},\boldsymbol{\lambda}}{\mathrm{argmin}} \sum_{i=1}^{n} d\left(z_i - \left[\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{B}_i\right]\right)\,,$$

where, for the sake of simplicity, we have denoted as $\mathbf{B}_i = (B_1(t_i), \dots, B_{k_n}(t_i))^{\mathrm{T}}$, so $\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{B}_i = \sum_{i=j}^{k_n} \lambda_j B_j(t_i)$.

On the other hand, robust estimators are obtained controlling large values of the deviance, with a $\rho-$ function ϕ, as defined in Maronna *et al.* (2006), i.e., an even function, non–decreasing on $|y|$, increasing for $y > 0$ when $\phi(y) < \lim_{t\to+\infty} \phi(t)$ and such that $\phi(0) = 0$. An example of such functions is the Tukey's biweight score function, $\phi(y) = \phi_{\text{T}}(y) = \min\left(3y^2 - 3y^4 + y^6, 1\right)$. Hence, in this case

$$\rho(z, s, a) = \phi\left(\frac{\sqrt{d\,(z - s)}}{a}\right) ,$$

so the tuning constant a needs to be chosen, unless it is fixed by the practitioner. Note that with this notation, the classical estimator corresponds to $\phi(u) = u^2$.

To provide an algorithm to compute the estimators with an adaptive constant, let us consider the situation in which we have fixed k_n so that we seek for $\boldsymbol{\lambda}$ such that $\sum_{i=j}^{k_n} \lambda_j B_j(t)$ provides a good approximation for $\eta_0(t)$. As in Bianco *et al.* (2005), a three step procedure can be considered to compute initial estimators of the parameters. First note that, since the tuning constant of the loss function depends on the unknown parameter α, Bianco *et al.* (2005) introduce an adaptive sequence of tuning constants $\widehat{c}_{\text{M},n}$ to define a sequence of $M-$estimators, $\widehat{\boldsymbol{\theta}}_{\text{M},n} = (\widehat{\boldsymbol{\beta}}_{\text{M},n}, \widehat{\boldsymbol{\lambda}}_{\text{M},n})$. When k_n is fixed, these estimators, which satisfy

$$\widehat{\boldsymbol{\theta}}_{\text{M},n} = \underset{\boldsymbol{\beta},\boldsymbol{\lambda}}{\operatorname{argmin}} \sum_{i=1}^{n} \phi\left(\frac{\sqrt{d\left(z_i - \left[\mathbf{x}_i^{\text{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}^{\text{T}}\mathbf{B}_i\right]\right)}}{\widehat{c}_{\text{M},n}}\right) ,$$

for constants $\widehat{c}_{\text{M},n} \xrightarrow{p} c_0$, have as asymptotic covariance matrix $\left(B(\phi, \alpha, c_0)/A^2(\phi, \alpha, c_0)\right) \boldsymbol{\Sigma}_0$ where $\boldsymbol{\Sigma}_0$ is the asymptotic covariance matrix of the classical estimators obtained when $\phi(u) = u^2$. The constants $B(\phi, \alpha, c_0)$ and $A^2(\phi, \alpha, c_0)$ depend only on the derivative of the score function ϕ and the shape parameter α, but not on the covariates. Hence, the estimators can be calibrated to attain a given efficiency. From now on, denote $C_e(\alpha)$ the value of the tuning constant c_0 such that the $M-$ estimator has efficiency e with respect to the classical one. Note that in particular, e will be the efficiency of the regression estimator $\widehat{\boldsymbol{\beta}}_{\text{M},n}$.

In our modification, we consider the following four step algorithm to compute a generalized $MM-$estimator. It is worth noticing that the method to be described below is just the proposal considered in Bianco *et al.* (2005) applied to the finite–approximation of η_0 but taking into account the order restrictions.

- **Step 1.** We first compute an initial $S-$estimates $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\lambda}}_n)$ and the corresponding scale estimate $\widehat{\sigma}_n$ taking $b = \sup \phi/2$. To be more precise, for each value of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ let

$\sigma_n(\boldsymbol{\beta},\boldsymbol{\lambda})$ be the $M-$scale estimate of $\sqrt{d\left(z_i - \left[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}_i\right]\right)}$ given by

$$\frac{1}{n}\sum_{i=1}^{n}\phi\left(\frac{\sqrt{d\left(z_i - \left[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}_i\right]\right)}}{\sigma_n(\boldsymbol{\beta},\boldsymbol{\lambda})}\right) = b\ ,$$

where ϕ is the Tukey bisquare function, ϕ_{T}.

The $S-$estimate of $(\boldsymbol{\beta}_0,\boldsymbol{\lambda}_0)$ for the considered model is defined as $\widetilde{\boldsymbol{\theta}}_n = \mathrm{argmin}_{\boldsymbol{\beta},\boldsymbol{\lambda}}\ \sigma_n(\boldsymbol{\beta},\boldsymbol{\lambda})$ and the corresponding scale estimate by $\widehat{\sigma}_n = \min_{\boldsymbol{\beta},\boldsymbol{\lambda}}\ \sigma_n(\boldsymbol{\beta},\boldsymbol{\lambda})$. Let u be a random variable with density (11) and write $\sigma^*(\alpha)$ for the solution of

$$\mathbb{E}_G\left[\phi\left(\frac{\sqrt{d(u_1)}}{\sigma^*(\alpha)}\right)\right] = b\,.$$

Similar arguments to those considered in Theorem 5 in Bianco *et al.* (2005) combined with the results of Theorem 3.1 allow to show that under mild conditions $\widetilde{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_0$, $\|\widetilde{\eta}-\eta_0\|_{\mathcal{F}}^2 \xrightarrow{a.s.} 0$, where $\widetilde{\eta} = \sum_{i=1}^{k_n}\widetilde{\lambda}_i B_i$ and that $\widehat{\sigma}_n \xrightarrow{a.s.} \sigma^*(\alpha)$. Moreover, as in Bianco *et al.* (2005), $\sigma^*(\alpha)$ is a continuous and strictly decreasing function and so, an estimator of α can be defined as $\widehat{\alpha}_n = \sigma^{*-1}(\widehat{\sigma}_n)$ leading to a a strongly consistent estimator for α.

- **Step 2.** In the second step, we compute $\widehat{\tau}_n = \sigma^{*-1}(\widehat{\sigma}_n)$ and

$$\widehat{c}_n = \max(\widehat{\sigma}_n, C_e(\widehat{\tau}_n)) = \max(\widehat{\sigma}_n, C_e(\sigma^{*-1}(\widehat{\sigma}_n))\,.$$

We then have that $\widehat{c}_n \xrightarrow{p} c_0 = \max\{\sigma^*(\alpha), C_e(\alpha)\}$.

- **Step 3.** Let $\widehat{\boldsymbol{\theta}}_n^{(0)} = \left(\widehat{\boldsymbol{\beta}}^{(0)\,\mathrm{T}}, \widehat{\boldsymbol{\lambda}}^{(0)\,\mathrm{T}}\right)^{\mathrm{T}}$ be the adaptive $MM-$estimator without restrictions defined by

$$\widehat{\boldsymbol{\theta}}_n^{(0)} = \underset{\boldsymbol{\nu}=(\boldsymbol{\beta},\boldsymbol{\lambda})}{\mathrm{argmin}}\sum_{i=1}^{n}\phi\left(\frac{\sqrt{d\left(z_i - \left[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}_i\right]\right)}}{\widehat{c}_n}\right)w(\mathbf{x}_i). \tag{12}$$

where the weight function $w(\mathbf{x})$ controls large leverage points in the $\mathbf{x}-$covariate space.

- **Step 4.** If $\widehat{\lambda}_1^{(0)} \le \widehat{\lambda}_2^{(0)} \le \cdots \le \widehat{\lambda}_{k_n}^{(0)}$, the final estimators are $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\eta}(t) = \sum_{j=1}^{k_n}\widehat{\lambda}_j^{(0)}B_j(t)$. Otherwise, the final estimators are obtained using a standard non–linear minimization algorithm with restrictions choosing as initial value $(\widehat{\boldsymbol{\beta}}_n^{(0)}, \boldsymbol{\lambda}^{(0)})$, where $\boldsymbol{\lambda}^{(0)} \in \mathcal{L}_{k_n}$. One possible choice for $\boldsymbol{\lambda}^{(0)}$ is $\lambda_1^{(0)} = \lambda_2^{(0)} = 0$ and $\lambda_i^{(0)} = i-2$ for $i = 3,\ldots,k_n$, in which case the matrix $\mathbf{A}$ below equals $\mathbf{A} = (1,-1,0,\ldots,0)$.

We briefly describe below an algorithm to approximate the minimizer of $L_n(\boldsymbol{\theta}, \widehat{c}_n)$ under the considered restrictions.

– Denote $\widehat{\boldsymbol{\nabla}}(\boldsymbol{\beta},\boldsymbol{\lambda}) = (\widehat{\boldsymbol{\nabla}}_1(\boldsymbol{\beta},\boldsymbol{\lambda})^{\mathrm{T}}, \widehat{\boldsymbol{\nabla}}_2(\boldsymbol{\beta},\boldsymbol{\lambda})^{\mathrm{T}})^{\mathrm{T}}$ the gradient function and $\widehat{\mathbf{H}}(\boldsymbol{\beta},\boldsymbol{\lambda}) = (\widehat{\mathbf{H}}_{ij}(\boldsymbol{\beta},\boldsymbol{\lambda}))_{1\le i,j\le 2}$ the gradient vector and negative Hessian matrix of the objective function, that is,

$$\begin{aligned}
\widehat{\boldsymbol{\nabla}}_1(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\Psi\left(z_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_i^{\mathrm{T}}\boldsymbol{\lambda}, \widehat{c}_n\right) w(\mathbf{x}_i)\mathbf{x}_i \\
\widehat{\boldsymbol{\nabla}}_2(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\Psi\left(z_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_i^{\mathrm{T}}\boldsymbol{\lambda}, \widehat{c}_n\right) w(\mathbf{x}_i)\mathbf{B}_i \\
\widehat{\mathbf{H}}_{11}(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\chi\left(z_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_i^{\mathrm{T}}\boldsymbol{\lambda}, \widehat{c}_n\right) w(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \\
\widehat{\mathbf{H}}_{12}(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\chi\left(z_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_i^{\mathrm{T}}\boldsymbol{\lambda}, \widehat{c}_n\right) w(\mathbf{x}_i)\mathbf{B}_i\mathbf{x}_i^{\mathrm{T}} \\
\widehat{\mathbf{H}}_{21}(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\chi\left(z_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_i^{\mathrm{T}}\boldsymbol{\lambda}, \widehat{c}_n\right) w(\mathbf{x}_i)\mathbf{x}_i\mathbf{B}_i^{\mathrm{T}} \\
\widehat{\mathbf{H}}_{22}(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\chi\left(z_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_i^{\mathrm{T}}\boldsymbol{\lambda}, \widehat{c}_n\right) w(\mathbf{x}_i)\mathbf{B}_i\mathbf{B}_i^{\mathrm{T}}
\end{aligned}$$

where

$$\Psi(z,s,a) = \partial\rho(z,s,a)/\partial s = \frac{1}{2\,a\sqrt{d\,(z-s)}}\phi'\left(\frac{\sqrt{d\,(z-s)}}{a}\right)(1-\exp(z-s))$$

with ϕ' the first derivative of ϕ and $\chi\,(z,u,a) = \partial\Psi(z,u,a)/\partial u$. Let $\mathcal{A} = \{i_1,\ldots,i_m\}$ the set of indices such that $\lambda_{i_j}^{(0)} = \lambda_{i_j+1}^{(0)}$. If $m > 0$ define the working matrix as $\mathbf{A} \in \mathbb{R}^{m\times(k_n+p)}$ in which the $j-$th row is the vector with its i_j-th element equal to 1 and the $(i_j+1)-$th element equal to -1, the remaining ones equal to 0.

– Fix an initial value $\boldsymbol{\theta}$ (in the first step, $\boldsymbol{\theta} = (\widehat{\boldsymbol{\beta}}_n^{(0)}, \boldsymbol{\lambda}^{(0)})$ and denote $\widehat{\mathbf{H}} = \widehat{\mathbf{H}}(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\nabla}} = \widehat{\boldsymbol{\nabla}}(\boldsymbol{\theta})$.

– **Step 4.1.** Find the feasible direction as

$$\boldsymbol{\eta} = -\left(\mathbf{I} - \widehat{\mathbf{H}}^{-1}\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\widehat{\mathbf{H}}^{-1}\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{A}\right)\widehat{\mathbf{H}}^{-1}\widehat{\boldsymbol{\nabla}}$$

– **Step 4.2.** If $\|\boldsymbol{\eta}\| < \epsilon$ for some $\epsilon > 0$ small enough, compute the Lagrange multipliers

$$\boldsymbol{\mu} = -\left(\mathbf{A}\widehat{\mathbf{H}}^{-1}\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{A}\widehat{\mathbf{H}}^{-1}\widehat{\boldsymbol{\nabla}}$$

Let μ_i be the $i-$th component of $\boldsymbol{\mu}$.

* If $\mu_i \geq 0$, for all $i \in \mathcal{A}$, then $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.
* If there exists at least one $i \in \mathcal{A}$ such that $\mu_i < 0$, determine the index corresponding to the largest μ_i and remove it from $\mathcal{A}$ and go to **S1**.

– **Step 4.3** Compute

$$\nu_1 = \min_{\eta_i > \eta_{i+1}, i \notin \mathcal{A}, 1 \leq i \leq k_n - 1} \frac{-(\lambda_{i+1} - \lambda_i)}{\eta_{i+1} - \eta_i}$$

and find the smallest r such that $L_n(\boldsymbol{\theta} + 2^{-r}\boldsymbol{\eta}, \widehat{\kappa}) < L_n(\boldsymbol{\theta}, \widehat{\kappa})$. Then replace $\boldsymbol{\theta}$ by $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \min(2^{-r}, \nu_1)\boldsymbol{\eta}$, update $\mathcal{A}$ and $\mathbf{A}$ and go to **Step 4.1**.

The following Lemma states the Fisher–consistency of the functionals related to the estimators $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\eta})$ and $(\widehat{\boldsymbol{\beta}}_n, \widehat{\eta})$. Its proof is given in the Appendix A and is a consequence of Lemma 1 in Bianco *et al.* (2005).

Lemma 4.1. *If the score function $\phi : \mathbb{R} \to [0, \infty)$ is a continuous, non-decreasing and even function such that $\phi(0) = 0$. Moreover, if $0 \leq s < v$ with $\phi(v) < \sup_s \phi(s)$ then $\phi(s) < \phi(v)$. Assume that, for almost any t_0, $\mathbb{P}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = c \ \cup \ w(\mathbf{x}) = 0 | t = t_0) < 1$, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, and $c \in \mathbb{R}$, $(\boldsymbol{\beta}, c) \neq \mathbf{0}$. Then, we have that the functionals related to the estimators $(\widehat{\boldsymbol{\beta}}_n, \widehat{\eta})$ are Fisher–consistent. Furthermore, the functionals related to $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\eta})$ are Fisher–consistent when* (8) *holds.*

5 Monte Carlo study

In this Section, we summarize the results of a simulation study designed to compare the performance of the proposed estimators with the classical ones under a log–Gamma partly linear isotonic regression model. In all Tables, the estimators in this paper are indicated as ROB while their classical counterparts are indicated as CL, since they correspond to the estimators based on the deviance. To be more precise, the robust estimators correspond to those controlling large values of the deviance as described in Section 4 and they were computed using the Tukey's biweight score function. The weight functions w used to control high leverage points was taken

w used to control high leverage points was taken as the Tukey's biweight function with tuning constant $c_w = 4.685$

$$w(x) = \begin{cases} \left(1 - \left[\dfrac{x - \widehat{\mu}_n}{c_w\, s_n}\right]^2\right)^2 & |x - \widehat{\mu}_n| \leq c_w s_n \\ 0 & |x - \widehat{\mu}_n| \geq c_w\, s_n\ , \end{cases} \tag{13}$$

with $\widehat{\mu}_n$ the median of x_i and $s_n = \text{MAD}(x_i)$, since we have considered $x_i \in \mathbb{R}$. On the other hand, the classical estimators correspond to the choice $\varphi(t) = t$ in (7) and $w \equiv 1$.

We have performed $NR = 1000$ replications with samples of size $n = 100$. The value of k_n was chosen as described in Section 2.2. The central model denoted C_0 in Tables corresponds to select (x_i, t_i) independent of each other such that $x_i \sim \text{N}(0, 1)$, $t_i \sim \mathcal{U}(0, 1)$. The response variable was generated as $y_i|(x_i, t_i) \sim \Gamma(3, \lambda_i)$, where

$$\mathbb{E}\left(y_i|(x_i, t_i)\right) = \frac{3}{\lambda_i} = \exp\{\beta_0 x_i + \eta_0(t_i)\}$$

with $\beta_0 = 2$. Hence, the transformed log–Gamma model is

$$z_i = \beta_0 x_i + \eta_0(t_i) + u_i\ ,$$

where $u_i \sim \log(\Gamma(3, 1))$. Two choices for the nonparametric component have been considered, $\eta_{0,1}(t) = \sin(\pi t/2)$ and $\eta_{0,2}(t) = \pi\, t + 0.25\, \sin(4\pi t)$ which leads to Models 1 and 2, respectively.

For each sample generated, we have considered three contaminations labelled C_1, C_2 and C_3 that lead to contaminated samples $(z_{i,c}, x_{i,c}, t_i)$. We have first generated a sample $v_i \sim \mathcal{U}(0, 1)$ for $1 \leq i \leq n$ and then, we have considered the following contamination scheme:

- C_1 introduces *bad* high leverage points in the carriers x, without changing the responses already generated, i.e., $z_{i,c} = z_i$, $1 \leq i \leq n$, while

  $$x_{i,c} = \begin{cases} x_i & \text{if } v_i \leq 0.90 \\ x_i^\star & \text{if } v_i > 0.90\,, \end{cases}$$

  where $x_i^\star \sim \text{N}\,(5, 1/16)$.

- C_2 introduces outlying observations in the responses generated according to the model but with an incorrect carrier x.

$$z_{i,c} = \begin{cases} z_i & \text{if } v_i \leq 0.90 \\ z_i^\star & \text{if } v_i > 0.90 \end{cases}$$

where $z_i^\star = \beta_0 x_i^\star + \eta_0(t_i) + u_i^\star$ with $u_i^\star \sim \log(\Gamma(3,1))$ and $x_i^\star$ a new observation from a $\mathrm{N}(5, 1/16)$. Note that the carriers are not contaminated in this situation, i.e., $x_{i,c} = x_i$.

- C_3 corresponds to increasing the variance of the carriers x and also to introduce large values on the responses

$$x_{i,c} = \begin{cases} x_i & \text{if } v_i \leq 0.90 \\ x_i^\star & \text{if } v_i > 0.90, \end{cases} \qquad z_{i,c} = \begin{cases} z_i & \text{if } v_i \leq 0.90 \\ z_i^\star & \text{if } v_i > 0.90\,, \end{cases}$$

where $x_i^\star$ is a new observation from a $\mathrm{N}(0, 25)$ and $z_i^\star = 3\,\log(10) + u_i^\star$ with $u_i^\star \sim \log(\Gamma(3,1))$

Table 1 summarize the obtained results and report the mean over replication of $\widehat{\beta} - \beta_0$, denoted bias$(\widehat{\beta})$, its standard deviation denoted SD$(\widehat{\beta})$ and the mean square error, that is, the mean over replications of $(\widehat{\beta} - \beta_0)^2$. To study the performance of the estimators of the regression function η_0, denoted $\widehat{\eta}$, we have considered the mean square error (MISE$(\widehat{\eta})$), i.e, the mean over replications of an approximation of the integrated square error (ISE) given by

$$\mathrm{ISE}(\widehat{\eta}) = n^{-1} \sum_{i=1}^{n} \left[\widehat{\eta}(t_i) - \eta_0(t_i)\right]^2 \,.$$

The classical estimator shows its sensitivity under all contaminations, the effect being worst in this case on the estimation of the regression function η_0 when contaminating the responses as in C_2 or C_3. For these two contamination the mean square errors of the classical estimators of η_0 are more than one thousand times those obtained by the robust procedure which are quite close to the corresponding ones under C_0. On the other hand, contaminating only on the carriers duplicates of the mean square error of the classical estimators $\widehat{\eta}_{\mathrm{CL}}$. Therefore, as expected large responses affect the estimators of the nonparametric component more than leverage points. It is worth noting that for the studied log–Gamma model, both the bias and the dispersion of the classical estimators of β_0 are increased under C_2 enlarging the mean square error. On the other hand, the increased mean square error obtained under C_3 is mainly due to the bias. The effect of the different contaminations is also striking in Figures 1 and 2 which gives the boxplots of $\widehat{\boldsymbol{\beta}}$ under Models 1 and 2, respectively. For instance, under C_1 and C_3, the whole

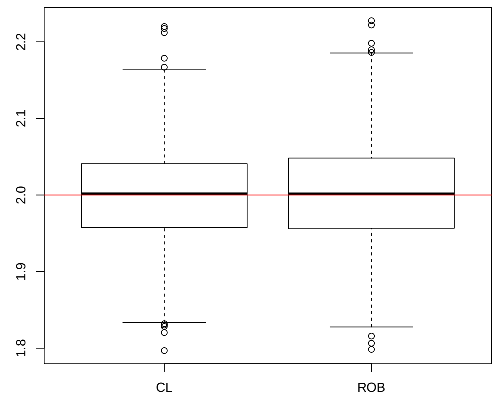
	Estimator	Model 1				Model 2			
		Bias($\widehat{\beta}$)	SD($\widehat{\beta}$)	MSE($\widehat{\beta}$)	MISE($\widehat{\eta}$)	mean($\widehat{\beta}$)	SD($\widehat{\beta}$)	MSE($\widehat{\beta}$)	MISE($\widehat{\eta}$)
C_0	CL	0.0002	0.0608	0.0037	0.0088	0.0000	0.0636	0.0040	0.0324
	ROB	0.0021	0.0672	0.0045	0.0096	0.0019	0.0700	0.0049	0.0340
C_1	CL	-0.5497	0.2170	0.3492	0.0265	-0.5549	0.2215	0.3570	0.0556
	ROB	-0.0016	0.0706	0.0050	0.0100	-0.0020	0.0728	0.0053	0.0344
C_2	CL	-1.8359	0.9343	4.2426	54.3390	-1.8168	0.9665	4.2340	52.8369
	ROB	0.0002	0.0711	0.0051	0.0103	-0.0001	0.0736	0.0054	0.0348
C_3	CL	-1.9400	0.2721	3.8376	15.0401	-1.9116	0.2581	3.7207	10.1817
	ROB	0.0043	0.0727	0.0053	0.0146	0.0020	0.0749	0.0056	0.0350

Table 1: Summary results for the estimators of β_0 and η_0, under a Gamma model. The estimators are obtained when k_n is the data–driven number of knots that minimizes $BIC(k)$.

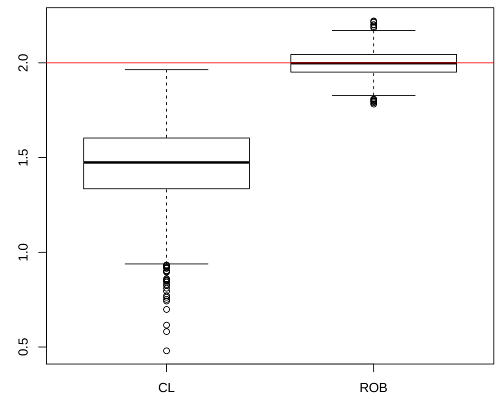
boxplot is under the horizontal line which corresponds to the true value $\beta_0 = 2$. On the other hand, the robust estimators are quite stable across all contaminated scenarios. Furthermore, the stability of the robust procedure is clearly illustrated in Figure 3 which plots the density estimators of $\widehat{\beta}_{\rm CL}$ and $\widehat{\beta}_{\rm R}$ under the different contamination schemes. The solid black lines correspond to the uncontaminated samples, while the red dashed, the blue dotted and the maroon dashed-dotted lines to contaminations C_1 to C_3 respectively. Besides, the dashed green line corresponds to the normal density with mean 2 and standard deviation equal to 0.0608 and 0.0672 for the classical and robust estimators, respectively. Note that these values correspond to SD($\widehat{\boldsymbol{\beta}}$) reported in Table 1, for clean samples. For the robust estimators all the density estimators are over-imposed showing that the contaminations have a mild effect on the estimations. On the other hand, when using the classical procedure based on the deviance, the densities of the estimators computed with contaminated samples move away from that obtained when clean data are considered, leading to unreliable estimates.

6 Real data example: Hospital Costs Data

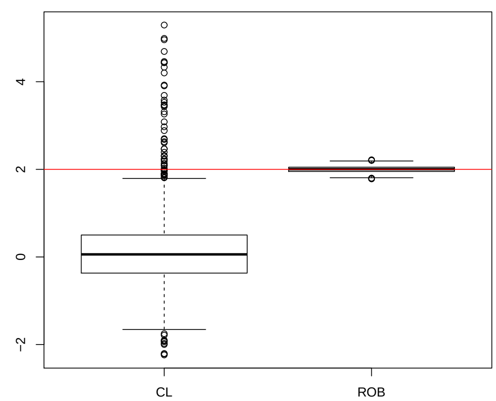
Marazzi and Yohai (2004) introduced a data set that corresponds to the costs of 100 patients in a Swiss hospital in 1999 for *medical back problems*. They concerned on the relationship between the hospital cost of stay, y, (Cost, in Swiss francs) and the following administrative explanatory variables:



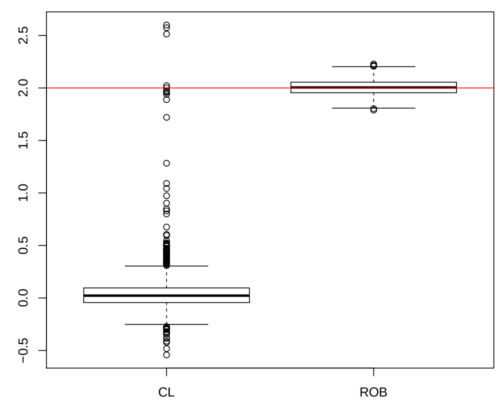
(a) C_0



(b) C_1



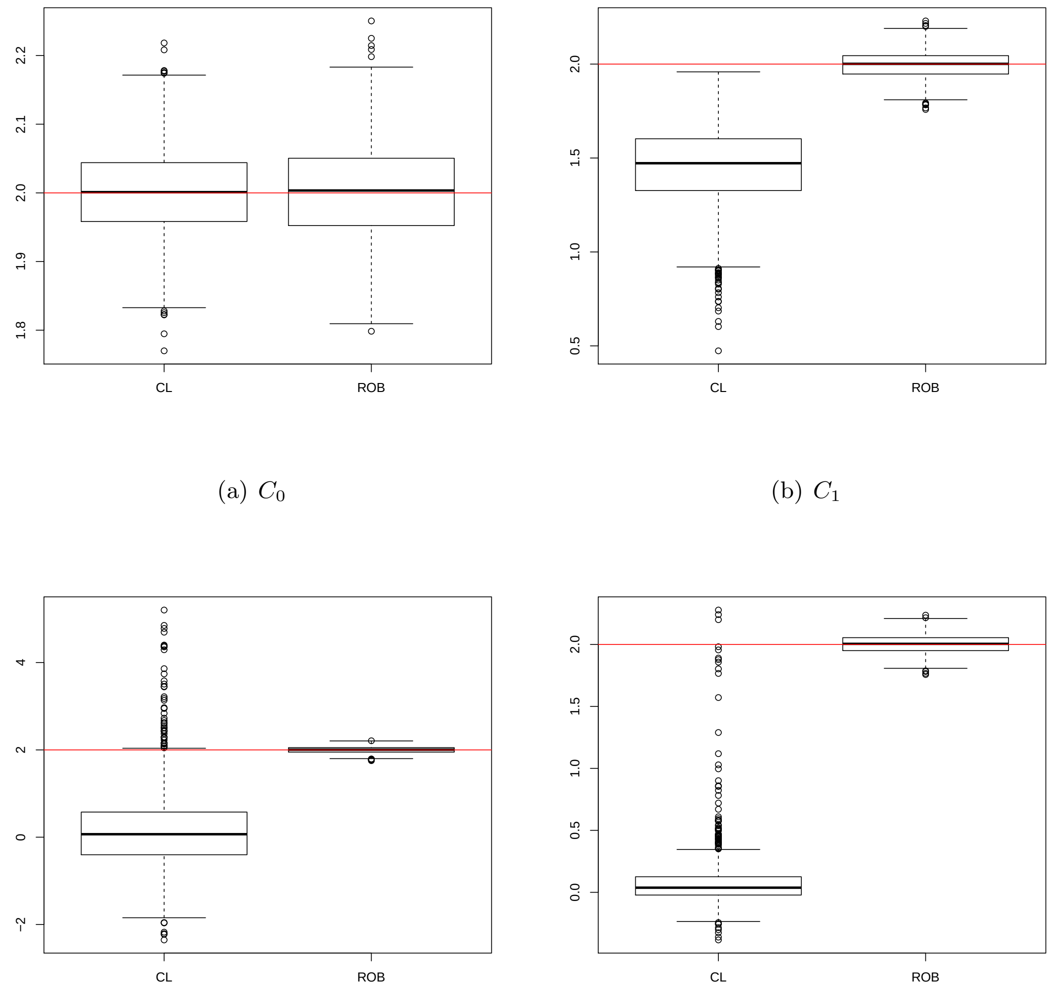
(c) C_2



(d) C_3

Figure 1: Boxplots of the estimators $\widehat{\beta}$ of β_0, under a log–Gamma Model with $\eta_0 = \eta_{0,1}$.

- LOS: length of stay in days
- ADM: admission type (0 = planned; 1 = emergency)
- INS: insurance type (0 = regular; 1 = private)
- AGE: years
- SEX: (0 = female; 1 = male)
- $DEST$: discharge destination (1 = home; 0 = other)

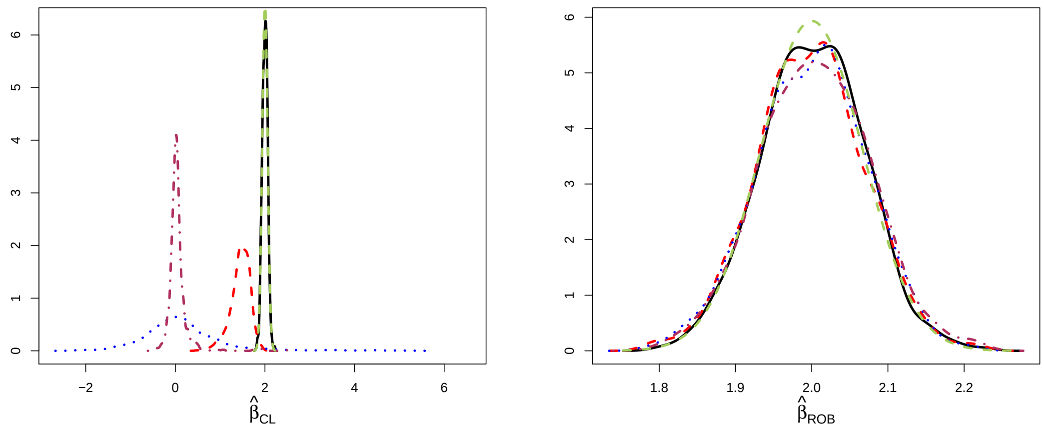


(a) C_0 (b) C_1

(c) C_2 (d) C_3

Figure 2: Boxplots of the estimators $\widehat{\beta}$ of β_0, under a log–Gamma Model with $\eta_0 = \eta_{0,2}$.

Cantoni and Ronchetti (2006) fitted to the complete data set the model $\log(\mathbb{E}(y_i|\mathbf{x}_i)) = \boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{x}_i$ which for Gamma responses is equivalent to $z_i = \log(y_i) = \boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{x}_i + u_i$, where $u_i \sim \log\Gamma(\alpha, 1)$ and $\mathbf{x} = (ADM, INS, AGE, SEX, DEST, \log(LOS), 1)$. Using their robust proposal, they identified 5 outliers corresponding to observations labelled as $14, 21, 28, 44$ and 63, whose weights are less or equal than 0.5. They realized that the atypical points affected the classical estimates of the coefficient of variable INS and the shape parameter. Bianco *et al.* (2013b) also analysed this data set to perform tests for the covariates SEX and $DEST$.



(a) Classical (b) Robust

Figure 3: Density estimator of the classical and robust estimators, $\widehat{\beta}_{\text{CL}}$ and $\widehat{\beta}_{\text{R}}$, of β_0, under a log–Gamma Model with $\eta_0 = \eta_{0,1}$. The solid black lines correspond to the uncontaminated samples, while the red dashed, the blue dotted and the maroon dashed-dotted lines to contaminations C_1 to C_3 respectively.

In this example, we do not impose a linear relation between $z_i = \log(y_i)$ and the $\log LOS$ but we consider the more general isotonic partial linear model

$$z_i = \boldsymbol{\beta}_0^{\text{T}} \mathbf{x}_i + \eta_0(t_i) + u_i$$

where u_i has $\log \Gamma(\alpha, 1)$ and $\mathbf{x} = (ADM, INS, AGE, SEX, DEST)$, while $t = \log(LOS)$ and η_0 is non–decreasing. The monotone assumption on η_0 is natural in this example, since the hospital cost increases the longer the stay. The obtained results for the estimators of $\boldsymbol{\beta}_0$ are reported in Table 2. For the classical estimators, denoted $\widehat{\boldsymbol{\beta}}_{\text{CL}}$, the BIC criterion selected $k_n = 4$, while for the robust ones, denoted $\widehat{\boldsymbol{\beta}}_{\text{R}}$, the best choice was $k_n = 5$ and the tuning constant for the $\rho-$function bounding the deviances equal $c_\rho = 0.3515$. As in the linear fit, the classical estimator of $\boldsymbol{\beta}_0$ are very sensitive to the 5 outliers, which were also detected in our study. In particular, the shape parameter and the coefficient related to the insurance type are highly affected. After removing these 5 data points, the classical estimators $\widehat{\boldsymbol{\beta}}_{\text{CL}}^{-\{5\}}$ are very similar to those obtained using $\widehat{\boldsymbol{\beta}}_{\text{R}}$, showing the good performance of the robust proposal in presence of outliers. We have computed the jackknife estimators of the standard deviation for the estimators of $\boldsymbol{\beta}$ which are reported between brackets.

Figure 4 shows the plot for the estimators of η_0 obtained using the classical (in red) and robust estimators (in blue) together with the linear fit provided by $\widehat{\boldsymbol{\beta}}_{\text{GM}}$, i.e., $\eta(t) = 0.8892\, t + 7.1268$. The linear fit seems to be a good choice for this data set, however, some discrepancies appear near the boundary which may be caused by a different shape of the regression function for large values of the $\log(LOS)$. It is worth noting that in this case, the shape of the classical estimator is quite close to that of the robust one and this can be mainly explained by the isotonic structure imposed.

	$\widehat{\boldsymbol{\beta}}_{\mathrm{CL}}$		$\widehat{\boldsymbol{\beta}}_{\mathrm{CL}}^{-\{5\}}$		$\widehat{\boldsymbol{\beta}}_{\mathrm{R}}$	
ADM	0.2148	(0.0560)	0.2172	(0.0418)	0.1979	(0.0294)
INS	0.0984	(0.1308)	-0.0324	(0.0514)	-0.0207	(0.0407)
AGE	-0.0009	(0.0014)	-0.0016	(0.0010)	-0.0019	(0.0006)
SEX	0.1088	(0.0523)	0.0820	(0.0352)	0.0615	(0.0329)
$DEST$	-0.1358	(0.0585)	-0.1608	(0.0499)	-0.1673	(0.0304)
$\widehat{\alpha}$	21.0809	-	45.7560	-	46.0088	-

Table 2: Analysis of Hospital Costs data under a log–Gamma isotonic partly linear regression model.

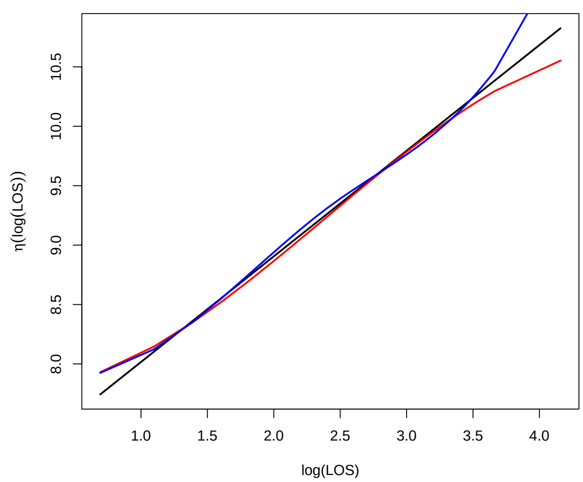


Figure 4: Classical (red) and robust (blue) estimators of the regression function $\eta(t)$ with the linear fit (black).

7 Final comments

The problem of estimating the nonparametric component η_0 and the regression parameter $\boldsymbol{\beta}_0$ under a generalized partly linear model has been extensively studied. Among other methods, $B-$splines have been considered to approximate the unknown function η_0. One advantage of $B-$splines is that they provide an estimation procedure that can be extended to the situation in which there are monotone constraints on the nonparametric component by imposing non–decreasing constraints on the coefficients. To overcome the sensitivity to atypical responses of the classical procedure based on the deviance, we have introduced a family of robust estimators for the components of a generalized partly linear model based on monotone $B-$splines, using a bounded loss function to control large deviance residuals. One of the advantages of our proposal is that it also allows for an unknown nuisance parameter, such as the scale parameter in partly linear regression models or the shape parameter in a Log–Gamma partly linear regression setting. Estimation of the nuisance parameter is an important issue since it allows to calibrate the robust estimators and to down–weight large residuals. Indeed, as in linear regression, to decide if an observation is an outlier it is necessary to determine the size of the residuals which strongly depends on the nuisance parameter estimator.

The obtained estimators are consistent and rates of convergence are also derived. The inadequate behaviour of the classical method when atypical data arise in the sample is confirmed through our simulation results. On the other hand, the robust procedure gives more reliable estimators leading to almost results either under the central log–Gamma model or under the studied contaminations.

8 Appendix A: Fisher–consistency

In this section, we discuss conditions ensuring the Fisher–consistency of the proposed estimators, i.e., $L(\boldsymbol{\beta}_0, \eta_0, \kappa_0) = \min\limits_{\boldsymbol{\beta}\in\mathbb{R}^p, g\in\mathcal{G}} L(\boldsymbol{\beta}, g, \kappa_0)$ where $L(\boldsymbol{\beta}, g, a)$ is defined in (3).

8.1 The logistic case

Let us first consider the situation of a logistic partially linear isotonic model. In this case, the loss function ρ given in (7) can be written as

$$\rho(y,u) = y\varphi\left(-\log\left[H(u)\right]\right) + (1-y)\varphi\left(-\log\left[1-H(u)\right]\right) + G(H(u))\,, \tag{A.1}$$

with $G(t) = G_1(t) + G_1(1-t)$, $G_1(t) = \int_0^t \varphi'(-\log u)\,du$ and $H(u) = 1/(1+\exp(-u))$.

More generally, we have the following results

Lemma 8.1. Let $\rho : \mathbb{R}^2 \to \mathbb{R}$ be defined as in (A.1) where the function $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is such that $\varphi(0) = 0$ and

a) $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is bounded with continuous and bounded derivative φ'.

b) $\varphi'(t) \geq 0$ and there exists some $c \geq \log 2$ such that $\varphi'(t) > 0$ for all $0 < t < c$.

Furthermore, assume that

$$\mathbb{P}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = a_0 \cup w(\mathbf{x}) = 0 | t = t_0\right) < 1, \qquad \forall (\boldsymbol{\beta}, a_0) \neq \mathbf{0} \text{ and for almost all } t_0. \tag{A.2}$$

Then, $(\boldsymbol{\beta}_0, \eta_0)$ is the unique minimizer of $L(\boldsymbol{\beta}, g)$.

Proof. The proof is a direct consequence of Lemma 2.1 in Bianco and Yohai (1996) and (A.2). As in Lemma 2.1 in Bianco and Yohai (1996), let z be a random Bernoulli variable such that $\mathbb{P}(z = 1) = \pi_0$ and define

$$M(\pi_0, \pi) = \mathbb{E} z \varphi\left(-\log \pi\right) + (1 - z)\varphi\left(-\log\left[1 - \pi\right]\right) + G(\pi)\,.$$

Then we have that $M(\pi_0, \pi_0) < M(\pi_0, \pi)$ for any $\pi \neq \pi_0$. Taking conditional expectation, and noticing that $\mathbb{P}(y = 1 | (\mathbf{x}, t)) = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t))$, we get that

$$\mathbb{E}\rho(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g(t)) w(\mathbf{x}) = \mathbb{E} w(\mathbf{x}) M\left[H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t)), H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g(t))\right]\,.$$

For a fixed value $(\mathbf{x}, t)$, denote $\pi = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g(t))$ and $\pi_0 = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t))$, the function $M(\pi_0, \pi)$ reaches its unique minimum when $\pi = \pi_0$ and the proof follows now easily from (A.2). □

8.2 The partially linear regression model

The partially linear model corresponds to the situation in which the link function equals $H(s) = s$. In this case, the model can be written as

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t_i) + \sigma_0 u_i\,,$$

where u_i are independent of $(\mathbf{x}_i, t_i)$ and σ_0 is the scale parameter.

As mentioned in Section 2.1, the loss function may be taken as $\rho(y, u, a) = \phi((y - u)/a)$ for an appropriate function ϕ. Furthermore, the nuisance parameter κ_0 plays the role of the scale parameter. In this section, we consider the situation in which the errors have a symmetric distribution and the function ϕ is an even function.

More precisely, to obtain Fisher–consistency results, we will need the following set of assumptions

F1 The random variable u has a density function $g_0(u)$ that is even, non-increasing in $|u|$, and strictly decreasing for $|u|$ in a neighbourhood of 0.

F2 The function $\phi : \mathbb{R} \to [0, \infty)$ is a continuous, non-decreasing and even function such that $\phi(0) = 0$. Moreover, if $0 \leq s < v$ with $\phi(v) < \sup_s \phi(s)$ then $\phi(s) < \phi(v)$. When ϕ is bounded we assume that $\sup_s \phi(s) = 1$.

F3 For almost any t_0, $\mathbb{P}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = c \,\cup\, w(\mathbf{x}) = 0 | t = t_0) < 1$, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, and $c \in \mathbb{R}$, $(\boldsymbol{\beta}, c) \neq \mathbf{0}$.

The following Lemma entails the Fisher–consistency of the proposed estimators.

Lemma 8.2. Let $\mathcal{G}_0 = \{g : [0,1] \to \mathbb{R} \text{ measurable}\}$. Under **F1** to **F3**, we have that, for any $\sigma > 0$, $(\boldsymbol{\beta}_0, \eta_0)$ is the unique minimizer over $\mathbb{R}^p \times \mathcal{G}_0$ of

$$L(\boldsymbol{\beta}, g, a) = \mathbb{E}\phi\left(\frac{y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} - g(t)}{a}\right) w(\mathbf{x})\,.$$

Proof. Let $\Upsilon(\mathbf{x}, t) = \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + g(t) - \eta_0(t)$, then, we have that

$$L(\boldsymbol{\beta}, \eta, a) = \mathbb{E}\phi\left(\frac{\sigma_0}{a}\,u - \frac{\Upsilon(\mathbf{x}, t)}{a}\right) w(\mathbf{x})$$

Denote as $\mathcal{A}_0 = \{(\mathbf{x}, t) : \Upsilon(\mathbf{x}, t) = 0\}$ and $b(\mathbf{x}, t) = \Upsilon(\mathbf{x}, t)/a$. Taking into account that the errors are independent of the covariates, we have that

$$L(\boldsymbol{\beta}, \eta, a) = \mathbb{E}\phi\left(u\frac{\sigma_0}{a}\right)\, \mathbb{E}\left(w(\mathbf{x})\mathbb{I}_{\mathcal{A}_0}(\mathbf{x}, t)\right) \,+\, \mathbb{E}\left\{\mathbb{E}\left[\phi\left(u\frac{\sigma_0}{a} - b(\mathbf{x}, t)\right)\bigg|(\mathbf{x}, t)\right] w(\mathbf{x})\mathbb{I}_{\mathcal{A}_0^c}(\mathbf{x}, t)\right\}\,.$$

Note that $\widetilde{u} = u\sigma_0/a$ also satisfies **F1**, hence Lemma 3.1 of Yohai (1987) together with **F2** imply for all $b \neq 0$ the following strict inequality holds

$$\mathbb{E}\left[\phi\left(u\frac{\sigma_0}{a} - b\right)\right] > \mathbb{E}\left[\phi\left(u\frac{\sigma_0}{a}\right)\right]\,. \tag{A.3}$$

Then, for any $(\mathbf{x}, t) \in \mathcal{A}_0^c$, we get

$$\mathbb{E}\left[\phi\left(u\frac{\sigma_0}{a} - b(\mathbf{x}, t)\right)\Big|(\mathbf{x}, t) = (\mathbf{x}_0, t_0)\right] = \mathbb{E}\left[\phi\left(u\frac{\sigma_0}{a} - b(\mathbf{x}_0, t_0)\right)\right] > \mathbb{E}\left[\phi\left(u\frac{\sigma_0}{a}\right)\right]$$

where the equality follows from the fact that the errors are independent of the covariates.

Note that **F3** immediately implies that $\mathbb{P}(\mathcal{A}_0^c \cap \{w(\mathbf{x}) \neq 0\}) > 0$. Then, putting all together, we obtain that

$$\begin{aligned} L(\boldsymbol{\beta}, \eta, a) &= \mathbb{E}\phi\left(u\frac{\sigma_0}{a}\right)\, \mathbb{E}\left(w(\mathbf{x})\mathbb{I}_{\mathcal{A}_0}(\mathbf{x}, t)\right) \,+\, \mathbb{E}\left\{\mathbb{E}\left[\phi\left(u\frac{\sigma_0}{a} - a(\mathbf{x}, t)\right)\bigg|(\mathbf{x}, t)\right] w(\mathbf{x})\mathbb{I}_{\mathcal{A}_0^c}(\mathbf{x}, t)\right\} \\ &> \mathbb{E}\phi\left(u\frac{\sigma_0}{a}\right)\, \mathbb{E}\left(w(\mathbf{x})\mathbb{I}_{\mathcal{A}_0}(\mathbf{x}, t)\right) \,+\, \mathbb{E}\left\{\mathbb{E}\left[\phi\left(u\frac{\sigma_0}{\sigma}\right)\right] w(\mathbf{x})\mathbb{I}_{\mathcal{A}_0^c}(\mathbf{x}, t)\right\} = \mathbb{E}\left(\phi\left(u\frac{\sigma_0}{a}\right) w(\mathbf{x})\right) \\ &> L(\boldsymbol{\beta}_0, \eta_0, a)\,, \end{aligned}$$

concluding the proof. ∎

8.3 The log–Gamma model

Under a generalized partially linear model with responses having a gamma distribution, that is, when $y_i|\mathbf{x}_i \sim \Gamma(\alpha, \mu_i)$, with $\mu_i = \mathbb{E}(y_i|(\mathbf{x}_i, t_i))$ and $\log(\mu_i) = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i + \eta_0(t_i)$, the responses can be transformed as $z_i = \log(y_i)$ so as to deal with the regression model with asymmetric errors given by (10), i.e.,

$$z_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_0(t_i) + u_i\,, \tag{A.4}$$

where u_i and $(\mathbf{x}_i, t_i)$ are independent. Recall that, under a log–Gamma model, the errors are such that $u_i \sim \log(\Gamma(\alpha, 1))$ and their density is strongly unimodal function.

In this setting, the loss function equals $\rho(z, s, a) = \phi\left(\sqrt{d\,(z-s)}/a\right)$, where $d(u) = \exp(u) - u - 1$.

We will derive Fisher–consistency results that include other skewed distributions with strongly unimodal densities for the errors. For that reason, we will consider the following additional assumption.

F4 The random variable u has a density function $g_0(u)$ that is strictly unimodal, continuous and $g_0(u) > 0$ for all u.

The following lemma gives a stronger result than the one stated in Lemma 4.1, since it shows that for any nuisance parameter the true parameters $(\boldsymbol{\beta}_0, \eta_0)$ minimize the objective function. This result corresponds to the condition required in Section 3.2 to avoid requiring any consistency order to the nuisance parameter estimator.

Lemma 8.3. Let $\mathcal{G}_0 = \{g : [0,1] \to \mathbb{R} \text{ measurable}\}$ and consider the partial linear regression model (A.4), where the density of the error u satisfies **F4**. Assume that **F2** and **F3** hold, then we have $(\boldsymbol{\beta}_0, \eta_0)$ is the unique minimizer over $\mathbb{R}^p \times \mathcal{G}_0$ of

$$L(\boldsymbol{\beta}, g, a) = \mathbb{E}\left[\phi\left(\frac{\sqrt{d(z - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} - g(t))}}{a}\right) w(\mathbf{x})\right]$$

Proof. As above, let $\Upsilon(\mathbf{x}, t) = \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + g(t) - \eta_0(t)$ and $\mathcal{A}_0 = \{(\mathbf{x}, t) : \Upsilon(\mathbf{x}, t) = 0\}$. Then, we have that

$$L(\boldsymbol{\beta}, g, a) = \mathbb{E}\left(\phi\left(\frac{\sqrt{d(u + \Upsilon(\mathbf{x}, t))}}{a}\right) w(\mathbf{x})\right).$$

Using that the errors are independent of the covariates, we conclude that

$$L(\boldsymbol{\beta}, g, a) = \mathbb{E}\left(\phi\left(\frac{\sqrt{d(u)}}{a}\right)\right)\mathbb{E}\left(w(\mathbf{x})\mathbb{I}_{\mathcal{A}_0}(\mathbf{x}, t)\right) + \mathbb{E}\left\{\mathbb{E}\left[\phi\left(\frac{\sqrt{d(u + \Upsilon(\mathbf{x}, t))}}{a}\right)\Bigg|(\mathbf{x}, t)\right] w(\mathbf{x})\,\mathbb{I}_{\mathcal{A}_0^c}(\mathbf{x}, t)\right\}. \tag{A.5}$$

Taking into account that the errors verify **F4**, from Lemma 1 in Bianco *et al.* (2005) we may bound the second term in (A.5). Effectively, for any $(\mathbf{x}, t) \in \mathcal{A}_0^c$ and for any fixed $a > 0$, we get

$$\mathbb{E}\left(\phi\left(\frac{\sqrt{d(u+\Phi(\mathbf{x},t))}}{a}\right)\middle|(\mathbf{x},t)\right) > \mathbb{E}\left(\phi\left(\frac{\sqrt{d(u)}}{a}\right)\middle|(\mathbf{x},t)\right) = \mathbb{E}\left(\phi\left(\frac{\sqrt{d(u)}}{a}\right)\right),$$

where the last equality follows from the fact that the errors are independent of the covariates. Using **F3**, we get that the strict inequality occurs on a set with positive probability and the result follows as in Lemma 8.2. □

9 Appendix B

Throughout this section we will denote as $\|\rho\|_\infty = \sup_{y\in\mathbb{R},u\in\mathbb{R},a\in\mathcal{V}} \rho(y,u,a)$ and $\|w\|_\infty = \sup_{\mathbf{x}\in\mathbb{R}^p} w(\mathbf{x})$.

9.1 Proof of Theorem 3.1.

Let $V_{\boldsymbol{\beta},g,a} = \rho\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g(t), a\right) w(\mathbf{x})$ and denote as P the probability measure of $(y_1, \mathbf{x}_1, t_1)$ and as P_n its corresponding empirical measure. Then, $L_n(\boldsymbol{\beta}, g, a) = P_n V_{\boldsymbol{\beta},g,a}$ and $L(\boldsymbol{\beta}, g, a) = P V_{\boldsymbol{\beta},g,a}$.

Recall that $\mathcal{M}_n(\mathcal{T}_n, \ell) = \left\{\sum_{i=j}^{k_n} \lambda_j B_j : \quad \lambda_1 \leq \cdots \leq \lambda_{k_n}\right\} = \left\{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B} : \boldsymbol{\lambda} \in \mathcal{L}_{k_n}\right\}$. The consistency of $\widehat{\kappa}$ entails that given any neighbourhood $\mathcal{V}$ of κ_0, there exists a null set $\mathcal{N}_{\mathcal{V}}$, such that for $\omega \notin \mathcal{N}_{\mathcal{V}}$, there exists $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$ we have that $\widehat{\kappa} \in \mathcal{V}$.

The proof follows similar steps as those used in the proof of Theorem 5.7 of van der Vaart (1998). Let us begin showing that

$$A_n = \sup_{\boldsymbol{\beta}\in\mathbb{R}^p, g\in\mathcal{M}_n(\mathcal{T}_n,\ell), a\in\mathcal{V}} |L_n(\boldsymbol{\beta}, g, a) - L(\boldsymbol{\beta}, g, a)| \xrightarrow{a.s.} 0\,. \tag{B.1}$$

Note that $A_n = \sup_{f\in\mathcal{F}_n}(P_n - P)f$, where $\mathcal{F}_n$ is defined in **C4**. Furthermore, **C1** entails that $\sup_{f\in\mathcal{F}_n}|f| = \|\rho\|_\infty\|w\|_\infty$ and **C4** and the fact that $k_n = O(n^\nu)$ with $\nu < 1/(2r) < 1$ imply that

$$\frac{1}{n}\log N(\epsilon, \mathcal{F}_n, L_1(P_n)) = O_{\mathbb{P}}(1)\,\frac{k_n}{n}\log\left(\frac{1}{\epsilon}\right) \xrightarrow{p} 0\,.$$

Hence, we get that (B.1) holds (see, for instance, exercise 3.6 in van der Geer, 2000 with $b_n = \max(1, \|\rho\|_\infty\|w\|_\infty)$).

Since $L(\boldsymbol{\theta}_0, \kappa_0) = \inf_{\boldsymbol{\beta}\in\mathbb{R}^p, g\in\mathcal{G}} L(\boldsymbol{\beta}, g, \kappa_0)$, where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \eta_0)$, we have that

$$0 \leq L(\widehat{\boldsymbol{\theta}}, \kappa_0) - L(\boldsymbol{\theta}_0, \kappa_0) = \sum_{j=1}^{3} A_{n,j}\,, \tag{B.2}$$

with $A_{n,1} = L(\widehat{\boldsymbol{\theta}}, \widehat{\kappa}) - L_n(\widehat{\boldsymbol{\theta}}, \widehat{\kappa})$, $A_{n,2} = L_n(\widehat{\boldsymbol{\theta}}, \widehat{\kappa}) - L(\boldsymbol{\theta}_0, \kappa_0)$ and $A_{n,3} = L(\widehat{\boldsymbol{\theta}}, \kappa_0) - L(\widehat{\boldsymbol{\theta}}, \widehat{\kappa})$. Noting that $|A_{n,1}| \leq A_n$, we obtain that $A_{n,1} = o_{\text{a.s.}}(1)$. On the other hand, since $L(\widehat{\boldsymbol{\theta}}, a) = L^\star(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}, a)$ the equicontinuity of $L^\star$ stated in **C1** and the consistency of $\widehat{\kappa}$ entails that $A_{n,3} = o_{\text{a.s.}}(1)$.

We will now bound $A_{n,2}$. Using Lemma A1 of Lu *et al.* (2007), we get that there exists $g_n \in \mathcal{M}_n(\mathcal{T}_n, \ell)$ with $\ell \geq r+2$, such that $\|g_n - \eta\|_\infty = O(n^{-r\nu})$, for $1/(2r+2) < \nu < 1/(2r)$. Denote $\boldsymbol{\theta}_n = (\boldsymbol{\beta}, g_n)$ and let $S_{n,1} = (P_n - P)V_{\boldsymbol{\beta}, g_n, \widehat{\kappa}}$ and $S_{n,2} = L(\boldsymbol{\theta}_n, \widehat{\kappa}) - L(\boldsymbol{\theta}_0, \kappa_0)$. Note that $S_{n,1} \leq A_n$, so that from (B.1), we get that $S_{n,1} \xrightarrow{a.s.} 0$. On the other hand, if we write $S_{n,2} = \sum_{j=1}^2 S_{n,2}^{(j)}$ where $S_{n,2}^{(1)} = L(\boldsymbol{\theta}_n, \widehat{\kappa}) - L(\boldsymbol{\theta}_n, \kappa_0)$ and $S_{n,2}^{(2)} = L(\boldsymbol{\theta}_n, \kappa_0) - L(\boldsymbol{\theta}_0, \kappa_0)$, the continuity of ρ together with the fact that $\|g_n - \eta\|_\infty \to 0$ and the dominated convergence theorem entail that $S_{n,2}^{(2)} \to 0$, while the continuity and boundedness of ρ together with the consistency of $\widehat{\kappa}$ leads to $S_{n,2}^{(1)} = o_{\text{a.s.}}(1)$. Hence, $S_{n,j} = o_{\text{a.s.}}(1)$ for $j = 1, 2$.

Using that $\widehat{\boldsymbol{\theta}}$ minimizes L_n over $\mathbb{R}^p \times \mathcal{M}_n(\mathcal{T}_n, \ell)$ we obtain that

$$A_{n,2} = L_n(\widehat{\boldsymbol{\theta}}, \widehat{\kappa}) - L(\boldsymbol{\theta}_0, \kappa_0) \leq L_n(\boldsymbol{\theta}_n, \widehat{\kappa}) - L(\boldsymbol{\theta}_0, \kappa_0) = S_{n,1} + S_{n,2}\,. \tag{B.3}$$

Hence, from (B.2) and (B.3) and using that $A_{n,j} = o_{\text{a.s.}}(1)$, for $j = 1, 3$ and $S_{n,j} = o_{\text{a.s.}}(1)$, for $j = 1, 2$, we conclude that

$$0 \leq L(\widehat{\boldsymbol{\theta}}, \kappa_0) - L(\boldsymbol{\theta}_0, \kappa_0) = \sum_{j=1}^3 A_{n,j} \leq o_{\text{a.s.}}(1)$$

so that $L(\widehat{\boldsymbol{\theta}}, \kappa_0) \to L(\boldsymbol{\theta}_0, \kappa_0)$. The fact that $\inf_{\pi(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) > \epsilon} L(\widetilde{\boldsymbol{\theta}}, \kappa_0) > L(\boldsymbol{\theta}_0, \kappa_0)$ entails that $\pi(\widehat{\theta}, \theta) \xrightarrow{a.s.} 0$, concluding the proof. $\square$

9.2 Proof of Theorem 3.2

To prove Theorem 3.2 under both sets of assumptions, we will state the common steps at the beginning and we then continue the proof when **C5*** or **C5**** hold.

We denote $\Theta_n = \mathbb{R}^p \times \mathcal{M}_n(\mathcal{T}_n, \ell) \cap \{\boldsymbol{\theta} = (\boldsymbol{\beta}, g) \in \Theta : \pi(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < \epsilon_0\}$, where $\Theta = \mathbb{R}^p \times \mathcal{G}$. Note that, except for a null probability set, $\widehat{\boldsymbol{\theta}} \in \Theta_n$, for n large enough. As in the proof of Theorem 3.1, let $g_n \in \mathcal{M}_n(\mathcal{T}_n, \ell)$ with $\ell \geq r+2$, $g_n(t) = \boldsymbol{\lambda}_n^{\mathrm{T}} \mathbf{B}(t)$, be such that $\|g_n - \eta_0\|_\infty = O(n^{-r\nu})$, for $1/(2r+2) < \nu < 1/(2r)$ and denote $\boldsymbol{\theta}_{0,n} = (\boldsymbol{\beta}_0, g_n)$.

In order to get the convergence rate of our estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\eta})$ we will apply Theorem 3.4.1 of van der Vaart and Wellner (1996). For that purpose, following the notation in that Theorem, denote as $M(\boldsymbol{\theta}) = -L(\boldsymbol{\theta}, \widehat{\kappa})$ and $\mathbb{M}_n(\boldsymbol{\theta}) = -L_n(\boldsymbol{\theta}, \widehat{\kappa})$ and for $\boldsymbol{\theta} \in \Theta_n$, denote $d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \pi_{\mathbb{P}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$. Note that the function M is random, due to the nuisance parameter estimator $\widehat{\kappa}$. Let $\delta_n = A\|\eta_0 - g_n\|_{\mathcal{F}}$, where $A = 4\sqrt{(C_0/\|w\|_\infty + A_0)/C_0}$ with $A_0 = \|w\|_\infty \|\chi\|_\infty / 2$ and C_0 given in **C8**.

Using that $|(L_n(\boldsymbol{\theta},\widehat{\kappa}) - L(\boldsymbol{\theta},\widehat{\kappa})) - (L_n(\boldsymbol{\theta}_{0,n},\widehat{\kappa}) - L(\boldsymbol{\theta}_{0,n},\widehat{\kappa}))| = |(\mathbb{M}_n - M)(\boldsymbol{\theta}) - (\mathbb{M}_n - M)(\boldsymbol{\theta}_{0,n})|$, to make use of Theorem 3.4.1 of van der Vaart and Wellner (1996), we have to show that there exists a function ϕ_n such that $\phi_n(\delta)/\delta^\nu$ is decreasing on (δ_n,∞) for some $\nu<2$ and that for any $\delta>\delta_n$,

$$\sup_{\boldsymbol{\theta}\in\Theta_{n,\delta}} L(\boldsymbol{\theta}_{0,n},\widehat{\kappa}) - L(\boldsymbol{\theta},\widehat{\kappa}) = \sup_{\boldsymbol{\theta}\in\Theta_{n,\delta}} M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}_{0,n}) \quad \lesssim \quad -\delta^2 \tag{B.4}$$

$$\mathbb{E}^* \sup_{\boldsymbol{\theta}\in\Theta_{n,\delta}} \sqrt{n}\,|(L_n(\boldsymbol{\theta},\widehat{\kappa}) - L(\boldsymbol{\theta},\widehat{\kappa})) - (L_n(\boldsymbol{\theta}_{0,n},\widehat{\kappa}) - L(\boldsymbol{\theta}_{0,n},\widehat{\kappa}))| \quad \lesssim \quad \phi_n(\delta) \tag{B.5}$$

$$d_n(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}_{0,n}) \quad \xrightarrow{p} \quad 0 \tag{B.6}$$

where the symbol $\lesssim$ means *less or equal up to a constant*, $\mathbb{E}^*$ stands for the outer expectation and $\Theta_{n,\delta} = \{\boldsymbol{\theta}\in\Theta_n : \delta/2 < d_n(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) \le \delta\}$.

Assumption **C8** and the fact that $\widehat{\kappa} \xrightarrow{a.s.} \kappa_0$ entails that, except for a null probability set, for any $\boldsymbol{\theta}\in\Theta_n$, $L(\boldsymbol{\theta},\widehat{\kappa}) - L(\boldsymbol{\theta}_0,\widehat{\kappa}) \ge C_0\,\pi^2_{\mathbb{P}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)$. On the other hand, using **C6**, we get that

$$\begin{aligned}
0 \le L(\boldsymbol{\theta}_{0,n},a) - L(\boldsymbol{\theta}_0,a) &= \mathbb{E}\left\{\mathbb{E}\left[w(\mathbf{x})\Psi(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(t),a)\,(g_n(t)-\eta_0(t))\,|(\mathbf{x},t)\right]\right\}\\
&+ \frac{1}{2}\,\mathbb{E}\left[w(\mathbf{x})\,\chi(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\widetilde{\eta}(t),a)\,(g_n(t)-\eta_0(t))^2\right]\\
&= \frac{1}{2}\,\mathbb{E}\left[w(\mathbf{x})\,\chi(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\widetilde{\eta}(t),a)\,(g_n(t)-\eta_0(t))^2\right]\\
&\le \frac{1}{2}\,\|w\|_\infty\|\chi\|_\infty\mathbb{E}\,(g_n(t)-\eta_0(t))^2 = A_0\,\|g_n-\eta_0\|_2^2 \le A_0\,\|g_n-\eta_0\|^2_{\mathcal{F}} = O(n^{-2r\nu})\,,
\end{aligned}$$

where $A_0 = \|w\|_\infty\|\chi\|_\infty/2$ and $\widetilde{\eta}(t)$ is an intermediate value between $\eta_0(t)$ and $g_n(t)$. Thus, using that $d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) \le 2d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) + 2d_n^2(\boldsymbol{\theta}_{0,n},\boldsymbol{\theta}_0) \le 2d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) + 2\|w\|_\infty\,\|g_n-\eta_0\|_2^2 \le 2d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) + 2\|w\|_\infty\,\|g_n-\eta_0\|^2_{\mathcal{F}}$ and that $\delta/2 < d_n(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n})$ we obtain that

$$\begin{aligned}
L(\boldsymbol{\theta},\widehat{\kappa}) - L(\boldsymbol{\theta}_{0,n},\widehat{\kappa}) &\ge C_0\,d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) - A_0\,\|g_n-\eta_0\|^2_{\mathcal{F}} \ge \frac{C_0}{2} d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) - \left(\frac{C_0}{\|w\|_\infty} + A_0\right)\|g_n-\eta_0\|^2_{\mathcal{F}}\\
&\ge \frac{C_0}{8}\delta^2 - \frac{1}{A^2}\left(\frac{C_0}{\|w\|_\infty} + A_0\right)\delta_n^2 = \frac{C_0}{8}\delta^2 - \frac{C_0}{16}\delta_n^2 \ge \frac{C_0}{16}\delta^2\,,
\end{aligned}$$

concluding the proof of (B.4).

We have now to find $\phi_n(\delta)$ such that $\phi_n(\delta)/\delta$ is decreasing in δ and (B.5) holds. Note that from the consistency of $\widehat{\kappa}$, we have that, with probability one for n large enough

$$\begin{aligned}
\sqrt{n}\,|(L_n(\boldsymbol{\theta},\widehat{\kappa}) - L(\boldsymbol{\theta},\widehat{\kappa})) \quad &- (L_n(\boldsymbol{\theta}_{0,n},\widehat{\kappa}) - L(\boldsymbol{\theta}_{0,n},\widehat{\kappa}))| \le \\
&\sup_{a\in\mathcal{V}} \sqrt{n}\,|(L_n(\boldsymbol{\theta},a) - L(\boldsymbol{\theta},a)) - (L_n(\boldsymbol{\theta}_{0,n},a) - L(\boldsymbol{\theta}_{0,n},a))|\,.
\end{aligned}$$

Define the class of functions

$$\mathcal{F}_{n,\delta} = \{V_{\boldsymbol{\theta},a} - V_{\boldsymbol{\theta}_{0,n},a} : \frac{\delta}{2} \le d_n(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) \le \delta\,, \boldsymbol{\theta}\in\Theta_n\,,\, a\in\mathcal{V}\} = \{V_{\boldsymbol{\theta},a} - V_{\boldsymbol{\theta}_{0,n},a} : \boldsymbol{\theta}\in\Theta_{n,\delta}\,,\, a\in\mathcal{V}\}\,,$$

with $V_{\boldsymbol{\theta},a} = \rho\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g(t), a\right) w(\mathbf{x})$, for $\boldsymbol{\theta} = (\boldsymbol{\beta}, g)$. The inequality (B.5) involves an empirical process indexed by $\mathcal{F}_{n,\delta}$, since

$$\mathbb{E}^* \sup_{\boldsymbol{\theta}\in\Theta_{n,\delta}} \sqrt{n}\,|(L_n(\boldsymbol{\theta},\widehat{\kappa}) - L(\boldsymbol{\theta},\widehat{\kappa})) - (L_n(\boldsymbol{\theta}_{0,n},\widehat{\kappa}) - L(\boldsymbol{\theta}_{0,n},\widehat{\kappa}))| \leq \mathbb{E}^* \sup_{f\in\mathcal{F}_{n,\delta}} \sqrt{n}|(P_n - P)f|\,.$$

For any $f \in \mathcal{F}_{n,\delta}$ we have that $\|f\|_\infty \leq A_1 = 2\|\rho\|_\infty \|w\|_\infty$. Furthermore, if $A_2 = \|\psi\|_\infty \|w\|_\infty$ using that

$$|V_{\boldsymbol{\theta},a} - V_{\boldsymbol{\theta}_{0,n},a}| \leq \|\psi\|_\infty w(\mathbf{x})|\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + g(t) - g_n(t)|\,,$$

and the fact that $\pi_{\mathbb{P}}(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) = d_n(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) \leq \delta$, we get that

$$Pf^2 \leq \|\psi\|_\infty \mathbb{E}\left(w^2(\mathbf{x})\left[\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + g(t) - g_n(t)\right]^2\right) \leq A_2\, \pi^2_{\mathbb{P}}(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) \leq A_2\, \delta^2\,.$$

Lemma 3.4.2 van der Vaart and Wellner (1996) leads to

$$\mathbb{E}^* \sup_{f\in\mathcal{F}_{n,\delta}} \sqrt{n}|(P_n - P)f| \leq J_{[\,]}\left(A_2^{1/2}\delta, \mathcal{F}_{n,\delta}, L_2(P)\right)\left(1 + A_1 \frac{J_{[\,]}(A_2^{1/2}\,\delta, \mathcal{F}_{n,\delta}, L_2(P))}{A_2\delta^2\,\sqrt{n}}\right)\,,$$

where $J_{[\,]}(\delta,\mathcal{F},L_2(P)) = \int_0^\delta \sqrt{1 + \log N_{[\,]}(\epsilon,\mathcal{F},L_2(P))}d\epsilon$ is the bracketing integral.

a) Assume now that $\mathbf{C5}^\star$ holds and note that for any $\boldsymbol{\theta} = (\boldsymbol{\beta}, g) \in \Theta_{n,\delta}$, g can be written as $g = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}$ for some $\boldsymbol{\lambda} \in \mathcal{L}_{k_n}$, so

$$d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_{0,n}) \;=\; \mathbb{E}\left(w(\mathbf{x})\left[\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_n)^{\mathrm{T}}\mathbf{B}(t)\right]^2\right)\,.$$

Hence, $\mathcal{F}_{n,\delta} \subset \mathcal{G}_{n,c,\boldsymbol{\lambda}_n}$ with $c = \delta$ and the bound given in $\mathbf{C5}^\star$ leads to

$$N_{[\,]}\left(\epsilon, \mathcal{F}_{n,\delta}, L_2(P)\right) \leq C_2 \left(\frac{\delta}{\epsilon}\right)^{k_n+p+1}\,.$$

This implies that

$$J_{[\,]}(A_2^{1/2}\delta, \mathcal{F}_{n,\delta}, L_2(P)) \lesssim \delta\sqrt{k_n + p + 1}\,.$$

If we denote $q_n = k_n + p + 1$ we obtain that for some constant A_3 independent of n and δ,

$$\mathbb{E}^* \sup_{\boldsymbol{\theta}\in\Theta_{n,\delta}} |\mathbb{G}_n V_{\boldsymbol{\theta}_{0,n},\kappa_0} - \mathbb{G}_n V_{\boldsymbol{\theta},\kappa_0}| \leq A_3\left[\delta\, q_n^{1/2} + \frac{q_n}{\sqrt{n}}\right]\,.$$

Choosing

$$\phi_n(\delta) = \delta\, q_n^{1/2} + \frac{q_n}{\sqrt{n}}\,,$$

we have that $\phi_n(\delta)/\delta$ is decreasing in δ, concluding the proof of (B.5). The fact that $\pi(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) \xrightarrow{a.s.} 0$, entails that $\pi_{\mathbb{P}}(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) \xrightarrow{a.s.} 0$ which together with $\pi_{\mathbb{P}}(\boldsymbol{\theta}_{0,n},\boldsymbol{\theta}_0) \to 0$, leads to (B.6).

Let $\gamma_n = O(n^{\min(r\nu,(1-\nu)/2)})$, then $\gamma_n \lesssim \delta_n^{-1}$, where $\delta_n = A\|\eta_0 - g_n\|_{\mathcal{F}} = O(n^{-r\nu})$. We have to show that $\gamma_n^2 \phi_n\left(1/\gamma_n\right) \lesssim \sqrt{n}$. Note that

$$\gamma_n^2 \phi_n \left(\frac{1}{\gamma_n}\right) = \gamma_n q_n^{1/2} + \gamma_n^2 \frac{q_n}{\sqrt{n}} = \sqrt{n}\, a_n(1+a_n)\ ,$$

where $a_n = \gamma_n q_n^{1/2}/\sqrt{n}$. Hence, to derive that $\gamma_n^2\phi_n\left(1/\gamma_n\right) \lesssim \sqrt{n}$, it is enough to show that $a_n = O(1)$, which follows easily since $k_n = O(n^{\nu})$ and $\gamma_n = O(n^{\varsigma})$ with $\varsigma = \min(r\nu, (1-\nu)/2)$.

Finally, the condition $\mathbb{M}_n(\widehat{\boldsymbol{\theta}}) \geq \mathbb{M}_n(\boldsymbol{\theta}_{0,n}) - O_{\mathbb{P}}(\gamma_n^{-2})$ required by Theorem 3.4.1 of van der Vaart and Wellner (1996) is trivially fulfilled because $\widehat{\boldsymbol{\theta}}_n$ minimizes $L_n(\boldsymbol{\theta}, \widehat{\kappa})$. Hence, we get that $\gamma_n^2 d_n^2(\boldsymbol{\theta}_{0,n}, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$.

On the other hand, $d_n(\boldsymbol{\theta}_{0,n}, \boldsymbol{\theta}_0) \leq \|w\|_{\infty}^{1/2} \|g_n - \eta_0\|_{\infty} = O(n^{-r\nu}) \leq \gamma_n$, which together with $\gamma_n^2 d_n^2(\boldsymbol{\theta}_{0,n}, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$ and the triangular inequality leads to $\gamma_n^2 d_n^2(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$, concluding the proof.

b) We will assume now that **C5**** holds. Therefore, using that any $f \in \mathcal{F}_{n,\delta}$ can be written as $f = f_1 - f_2$ with $f_j \in \mathcal{F}_{n,\epsilon_0}^{\star}$ and the bound given in **C5****, we get that

$$N_{[\,]}\left(\epsilon, \mathcal{F}_{n,\delta}, L_2(P)\right) \leq C_2^2 \frac{1}{\epsilon^{2(k_n+p+1)}}\,.$$

This implies that

$$J_{[\,]}(A_2^{1/2}\delta, \mathcal{F}_{n,\delta}, L_2(P)) \lesssim \delta \log\left(\frac{1}{\delta}\right)\sqrt{k_n + p + 1}\,.$$

If we denote $q_n = k_n + p + 1$ we obtain

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta_{n,\delta}} |\mathbb{G}_n V_{\boldsymbol{\theta}_{0,n},\kappa_0} - \mathbb{G}_n V_{\boldsymbol{\theta},\kappa_0}| \leq A\left(q_n^{1/2}\delta \log\left(\frac{1}{\delta}\right) + n^{-1/2} q_n \left[\log\left(\frac{1}{\delta}\right)\right]^2\right)\,.$$

Choosing

$$\phi_n(\delta) = q_n^{1/2}\delta\log\left(\frac{1}{\delta}\right) + n^{-1/2} q_n \left[\log\left(\frac{1}{\delta}\right)\right]^2\,,$$

we have that $\phi_n(\delta)/\delta$ is decreasing in δ.

Therefore, from Theorem 3.4.1 of van der Vaart and Wellner (1996), we conclude that $\gamma_n^2 d_n^2(\boldsymbol{\theta}_{0,n}, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$, where γ_n is any sequence satisfying $\gamma_n \lesssim \delta_n^{-1}$ with $\delta_n = \pi(\boldsymbol{\theta}_0, \boldsymbol{\theta}_{0,n}) = O(n^{-r\nu})$ and $\gamma_n^2 \phi_n\left(1/\gamma_n\right) \leq \sqrt{n}$. The first condition, entails that $\gamma_n \leq O(n^{r\nu})$. The second one, implies that

$$\gamma_n^2 \left(q_n^{1/2}\gamma_n^{-1}\log(\gamma_n) + q_n n^{-1/2}[\log(\gamma_n)]^2\right) \leq n^{1/2}\,,$$

so using that $k_n = O(n^{\nu})$ we get that $\gamma_n \log(\gamma_n) \leq O(n^{(1-\nu)/2})$. Finally, the condition $\mathbb{M}_n(\widehat{\boldsymbol{\theta}}) \geq \mathbb{M}_n(\theta_0) - O_{\mathbb{P}}(r_n^{-2})$ required by Theorem 3.4.1 of van der Vaart and Wellner (1996) is trivially fulfilled because $\widehat{\boldsymbol{\theta}}_n$ minimizes $L_n(\boldsymbol{\theta}, \widehat{\kappa})$.

On the other hand, $d_n(\boldsymbol{\theta}_{0,n}, \boldsymbol{\theta}_0) \leq \|w\|_\infty^{1/2} \|g_n - \eta_0\|_\infty = O(n^{-r\nu}) \leq \gamma_n$, which together with $\gamma_n^2 d_n^2(\boldsymbol{\theta}_{0,n}, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$ and the triangular inequality leads to $\gamma_n^2 d_n^2(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}) = O_{\mathbb{P}}(1)$. $\square$

Acknowledgements. This research was partially supported by Grants PIP 112-201101-00742 from CONICET, PICT 2014-0351 from ANPCYT and 20020130100279BA and 20020120200244BA from the Universidad de Buenos Aires at Buenos Aires, Argentina.

References

Álvarez, E. and Yohai, J. (2012). $M-$estimators for isotonic regression. *J. Statist. Plann. Inf.*, **142**, 2241-2284.

Bianco, A. and Boente, G. (2004). Robust estimators in semiparametric partly linear regression models. *J. Statist. Planning and Inference* **122**, 229-252.

Bianco, A, Boente, G. and Rodrigues, I. (2013a). Resistant estimators in Poisson and Gamma models with missing responses and an application to outlier detection. *J. Multivar. Anal.*, **114**, 209-226.

Bianco, A, Boente, G. and Rodrigues, I. (2013b) Robust tests in generalized linear models with missing responses. *Comp. Statist. . Data Anal.*, **65**, 80-97.

Bianco, A., García Ben, M. and Yohai, V. (2005). Robust estimation for linear regression with asymmetric errors. *Canad. J. Statist.*, **33**, 511-528.

Bianco, A. and Yohai, V. (1996). Robust estimation in the logistic regression model. *Lecture Notes in Statistics*, **109**, 17-34. Springer–Verlag, New York.

Boente, G., He, X. and Zhou, J. (2006). Robust estimates in generalized partially linear models. *Ann. Statist.*, **34**, 2856-2878.

Boente, G. and Rodríguez, D. (2010). Robust inference in generalized partially linear models. *Comput. Statist. Data Anal.*, **54**, 2942-2966.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.*, **96**, 1022-1030.

Croux, C. and Haesbroeck, G. (2002). Implementing the Bianco and Yohai estimator for logistic regression. *Comp. Statist. Data Anal.*, **44**, 273-295.

Du, J., Sun, Z. and Xie, T. (2013). $M-$estimation for the partially linear regression model under monotonic constraints. *Statist. Prob. Letters*, **83**, 1353-1363.

Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models*. Physica-Verlag.

He, X. and Shi, P. (1996). Bivariate tensor–product $B-$spline in a partly linear model. *J. Multivariate Anal.*, **58**, 162-181.

He, X. and Shi, P. (1998). Monotone B-Spline smoothing. *J. Amer. Statist. Assoc.*, **93**, 643-650.

He, X., Zhu, Z. and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, **89**, 579-590.

Heritier, S., Cantoni, E., Copt, S. and Victoria–Feser, M.P. (2009). *Robust Methods in Biostatistics*. Wiley Series in Probability and Statistics. Wiley.

Huang, J. (2002). A note on estimating a partly linear model under monotonicity constraints. *J. Statist. Plann. Inf.*, **107**, 343-351.

Künsch, H., Stefanski, L. and Carroll, R. (1989). Conditionally unbiased bounded influence estimation in general regression models with applications to generalized linear models. *J. Amer. Statist. Assoc.* **84**, 460-466.

Lu, M., Zhang, Y. and Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika*, **94**, 705-718.

Lu, M. (2010). Spline-based sieve maximum likelihood estimation in the partly linear model under monotonicity constraints. *J. Multivar. Anal.*, **101**, 2528-2542.

Lu, M. (2015). Spline estimation of generalised monotonic regression. *J. Nonpar. Statist.*, **27**, 19-39.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. (2nd ed.) London: Champman and Hall.

Marazzi, A. and Yohai, V. (2004). Adaptively truncated maximum likelihood regression with asymmetric errors. *J. Statist. Plann. Inference*. **122**, 271-291.

Maronna R., Martin D. and Yohai V. (2006). *Robust statistics: Theory and methods* , Wiley, New York.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, **3**, 425–441.

Schumaker,L. (1981). *Spline Functions: Basic Theory*, Wiley, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Shen, X., and Wong, W. H. (1994) Convergence rate of sieve estimates. *Ann. Statist.*, **22**, 580-615.

Stefanski, L., Carroll, R. and Ruppert, D. (1986). Bounded score functions for generalized linear models. *Biometrika* **73**, 413-424.

Sun,Z., Zhang,Z. and Du,J. (2012). Semiparametric analysis of isotonic errors–in–variables regression models with missing response. *Communications in Statistics: Theory and Methods*, **41**, 2034–2060.

Van der Geer, S. (2000). *Empirical Processes in $M-$Estimation*, Cambridge University Press.

van der Vaart, A. (1998). *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer–Verlag, New York.