Ann Inst Stat Math (2018) 70:1115–1146
https://doi.org/10.1007/s10463-017-0618-9

General rank-based estimation for regression single index models

Huybrechts F. Bindele[1] · Ash Abebe[2] · Karlene N. Meyer[3]

Received: 7 December 2016 / Revised: 5 August 2017 / Published online: 20 September 2017

Abstract This study considers rank estimation of the regression coefficients of the single index regression model. Conditions needed for the consistency and asymptotic normality of the proposed estimator are established. Monte Carlo simulation experiments demonstrate the robustness and efficiency of the proposed estimator compared to the semiparametric least squares estimator. A real-life example illustrates that the rank regression procedure effectively corrects model nonlinearity even in the presence of outliers in the response space.

Keywords Single index · Rank-based objective function · Strong consistency · Asymptotic normality · Nonparametric kernel estimation

1 Introduction

The single index model (SIM) has gained widespread popularity in many areas of research such as finance, economics, epidemiology, medical science and ecology. One reason for its popularity is that it searches for a single linear combination of p-

✉ Huybrechts F. Bindele
hbindele@southalabama.edu

Ash Abebe
abebeas@auburn.edu

Karlene N. Meyer
knm6@georgetown.edu

1 Department of Mathematics and Statistics, University of South Alabama, 411 University Blvd. N, ILB 316, Mobile, AL 36688-0002, USA

2 Department of Mathematics and Statistics, Auburn University, Auburn, AL 36849, USA

3 Department of Mathematics and Statistics, Georgetown University, Washington, DC 20057, USA

covariates $\mathbf{X}$ that can capture most information about the relation between the response Y and covariate $\mathbf{X}$, thereby avoiding the "curse of dimensionality." It is defined as

$$Y_i = g_0(\mathbf{X}_i^\tau \boldsymbol{\beta}_0) + \varepsilon_i, \qquad 1 \le i \le n \ , \tag{1}$$

where $\boldsymbol{\beta}_0 \in \mathscr{B} \subset \mathbb{R}^p$ is a vector of parameters, $\mathbf{X}_i \in \mathscr{X}$, where $\mathscr{X}$ is a sub-vector space of $\mathbb{R}^p$, are independent (but not necessarily identically distributed) p-variable random covariate vectors, the function g_0 is an unknown function and the model errors ε_i are independent with conditional mean zero given the covariates. We will assume that $\mathscr{B}$ is compact and $E(Y|\mathbf{X}) = E(Y|\mathbf{X}^\tau\boldsymbol{\beta}_0) = E(g_0(\mathbf{X}^\tau\boldsymbol{\beta}_0)|\mathbf{X}^\tau\boldsymbol{\beta}_0)$; that is, if we set $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, we have $m(\mathbf{X}) = g_{\boldsymbol{\beta}_0}(\mathbf{X}^\tau\boldsymbol{\beta}_0)$ with $g_{\boldsymbol{\beta}}(t) = E(g_0(\mathbf{X}^\tau\boldsymbol{\beta}_0)|\mathbf{X}^\tau\boldsymbol{\beta} = t)$. We are interested in direct inference, especially in the estimation of the true parameter $\boldsymbol{\beta}_0$ in model (1). For reasons of identifiability, we set the first element of $\boldsymbol{\beta}_0$ to be 1 (Delecroix et al. 2006). When $g_{\boldsymbol{\beta}_0}(\cdot)$ is assumed to be the inverse of a known link function, this model belongs to the class of the well-known generalized linear model (McCullagh and Nelder 1989). Because of the flexibility and the interpretability that the SIM offers, this estimation problem has experienced rapid development in both theory and methodology in the past few years.

Since both $\boldsymbol{\beta}_0$ and $g_{\boldsymbol{\beta}_0}(\cdot)$ are unknown, they need to be estimated. We typically estimate $g_{\boldsymbol{\beta}_0}(\cdot)$ nonparametrically via kernel estimation and $\boldsymbol{\beta}_0$ via a minimization of some objective function. Several authors, including Powell et al. (1989), Härdle and Stoker (1989), Klein and Spady (1993), Ichimura (1993), Härdle et al. (1993), Sherman (1994), Horowitz and Härdle (1996), Hristache et al. (2001), Xia et al. (2002), Yu and Ruppert (2002), Yin and Cook (2005) and (Delecroix et al. 2006), have proposed a variety of approaches to solving this estimation problem. The commonly applied methods in the literature include the average derivative estimation (ADE) method and the semiparametric least squares (SLS) estimation method. ADE estimates the expected value of the weighted gradient of the regression function $g_{\boldsymbol{\beta}}$, which is proportional to $\boldsymbol{\beta}_0$. This method leads to $\sqrt{n}$-consistent estimator (Härdle and Tsybakov 1993). The SLS estimation method obtains the estimator of $\boldsymbol{\beta}_0$ by minimizing the Euclidean norm of the residuals $Y - E(Y|\mathbf{X}^\tau\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$. The function $g_{\boldsymbol{\beta}}(t)$ is estimated using nonparametric estimation such as Nadaraya–Watson (Ichimura 1993) or local linear approximation (Carroll et al. 1997). Since ADE cannot be used to estimate $\boldsymbol{\beta}_0$ when the gradient of the regression function is zero, Xia (2006) proposed an estimator based on the integration of the outer product of gradients using a local linear approximation.

While ADE and SLS are asymptotically efficient when dealing with models that have controlled designs and normal error distribution, they become inefficient when we have uncontrolled designs (high-leverage points) or non-normal (heavy-tailed or contaminated) error distributions. To overcome the latter issue, efforts have been devoted to constructing robust estimators. Han (1987) proposed an estimator based on maximization of the rank correlation between the observed values and the values fitted by the model. On the other hand, Delecroix et al. (2006) considered the M-estimation procedure to down-weigh large residuals (possibly from heavy tails and outliers) as an alternative to the SLS estimation. More recently, Feng et al. (2012) obtained an estimator based on the Wilcoxon rank-based procedure (Hettmansperger and McKean

2011). Although the estimator proposed by Feng et al. (2012) is robust against outliers in response space, their method lacks the generality and adaptivity of the general rank regression (Jaeckel 1972; Hettmansperger and McKean 2011). Moreover, the use of the gradient, based on the derivative with respect to $\mathbf{x}$ in both Xia (2006) and Feng et al. (2012), might reduce the applicability of their approaches since the function $\mathbf{x} \mapsto E(Y|\mathbf{X} = \mathbf{x})$ has to be assumed differentiable as a function of $\mathbf{x}$. However, it is quite common for $\mathbf{x} \mapsto E(Y|\mathbf{X} = \mathbf{x})$ to be a discontinuous function of $\mathbf{x}$. For example, $\mathbf{x}$ could be comprised of both continuous and categorical predictors as in treatment comparison studies including common ANCOVA type designs.

To overcome the deficiencies in the approaches of Xia (2006) and Feng et al. (2012) and similar to the M-estimation approach by Delecroix et al. (2006), we propose a general rank-based estimation approach, which is based on minimizing (2) or (4) given below. Our proposed procedure is also geared toward addressing robustness issues pertaining to outliers in the response space. Note that the objective function defined by (2) or (4) includes the Wilcoxon rank-based objective function that was used by Feng et al. (2012) which we find by taking a linear score generating function φ (Hettmansperger and McKean 2011, p. 82). Rather than using the outer product of gradients, we propose estimating $\boldsymbol{\beta}_0$ via minimizing (4), which in turn is defined on the basis of the leave-one-out Nadaraya–Watson estimator of $g_{\boldsymbol{\beta}_0}(\cdot)$. As pointed out in Delecroix et al. (2006), any technique that uses a nonparametric component in the optimization procedure must take bandwidth selection into account. As in Delecroix et al. (2006), this can be done by either using a separate minimization procedure over (4), where the starting bandwidth is chosen to belong in $\mathscr{F}_n := \{h : c_1 n^{-\alpha_1} < h < c_2 n^{-\alpha_2}\}$, for some constants $c_1, c_2 > 0$ and $1/8 < \alpha_1 < \alpha_2 < 1/4$, or using a direct joint estimation by minimizing (4) with respect to both $\boldsymbol{\beta}$ and h. Because of higher-order asymptotic expansions of the rank estimator, finding the optimal bandwidth is a complicated task and is left for future work.

We conclude this introduction by stating how the paper is organized. Section 2 provides the estimation procedures of both $g_{\boldsymbol{\beta}_0}(\cdot)$ and $\boldsymbol{\beta}_0$. The asymptotic properties (consistency and $\sqrt{n}$-asymptotic normality) of the rank estimator are studied in Sects. 3 and 4. To demonstrate the appeal of the rank estimator for dealing with heavy-tailed or contaminated error distribution settings, a simulation study is provided in Sect. 5. The same section also contains an analysis of a real-world dataset demonstrating the use of the proposed procedure. Proofs of theoretical results are given in Appendix.

2 Estimation

Let $\mu_{\boldsymbol{\beta}_0}(t)$ be the probability density (mass) of $\mathbf{X}^\tau \boldsymbol{\beta}_0$, and for K being a kernel function with bandwidth $h \equiv h_n$ satisfying $h_n \to 0$ *as* $n \to \infty$, let $\widehat{\mu}_{\tilde{\boldsymbol{\beta}}_n^0}(t) = (nh_n)^{-1} \sum_{k=1}^n K((t - \mathbf{X}_k^\tau \tilde{\boldsymbol{\beta}}_n^0)/h_n)$, where $\tilde{\boldsymbol{\beta}}_n^0$ is a $\sqrt{n}$-preliminary estimator of $\boldsymbol{\beta}_0$. For some positive constants c_0 and c such that $0 < c_0 \le c$, set

$$\Gamma = \{\mathbf{x} : \mu_{\boldsymbol{\beta}_0}(\mathbf{x}^\tau \boldsymbol{\beta}_0) \ge c\} \quad \text{and} \quad \Gamma_n = \{\mathbf{x} : \widehat{\mu}_{\tilde{\boldsymbol{\beta}}_n^0}\left(\mathbf{x}^\tau \tilde{\boldsymbol{\beta}}_n^0\right) \ge c\}.$$

Also, let $\mathscr{H}_n = [n^{-(1/2-\varepsilon)}, n^{-\varepsilon}]$, for ε such that $0 < \varepsilon < 1/2$. Clearly, $\mathscr{F}_n \subset \mathscr{H}_n$. Thus, while all the theoretical results established in this paper are based on $\mathscr{H}_n$, there are also valid on $\mathscr{F}_n$ (Delecroix et al. 2006).

Now, define the residuals $z_i(\boldsymbol{\beta}) = Y_i - g_{\boldsymbol{\beta}}(\mathbf{X}_i^\tau \boldsymbol{\beta}),\ 1 \le i \le n$, and consider the following rank objective function introduced by (Jaeckel 1972)

$$D_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\varphi\Big(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\Big)z_i(\boldsymbol{\beta}), \tag{2}$$

where $R(z_i(\boldsymbol{\beta}))$ is the rank of $z_i(\boldsymbol{\beta})$ among $z_1(\boldsymbol{\beta}), \ldots, z_n(\boldsymbol{\beta})$, and $\varphi : (0, 1) \to \mathbb{R}$ is a bounded nondecreasing score function. Before dealing with the estimation of true parameter $\boldsymbol{\beta}_0$, let us consider a nonparametric estimation of $g_{\boldsymbol{\beta}}(\cdot)$. To that end, define weights as

$$W_{nj}^i(t) = \frac{K\Big(\frac{t-\mathbf{X}_j^\tau\boldsymbol{\beta}}{h_n}\Big)}{\sum_{k\neq i}^n K\Big(\frac{t-\mathbf{X}_k^\tau\boldsymbol{\beta}}{h_n}\Big)}. \tag{3}$$

and let $\widehat{g}^i_{\boldsymbol{\beta},h}(t) = \sum_{j\neq i}^n W^i_{nj}(t)Y_j$ be the leave-one-out Nadaraya–Watson estimator of $g_{\boldsymbol{\beta}}(t)$, $t \in \mathscr{A}$, where $\mathscr{A} := \{\mathbf{x}^\tau\boldsymbol{\beta} : \mathbf{x} \in \mathscr{X},\ \boldsymbol{\beta} \in \mathscr{B}\}$. The function $\widehat{g}^i_{\boldsymbol{\beta},h}(\cdot)$ being a known nonlinear function of $\boldsymbol{\beta}$, we use it to define a rank dispersion analogous to (2) as

$$\widetilde{D}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I_{\Gamma_n}(\mathbf{X}_i)\varphi\Big(\frac{R(\nu_{ni}(\boldsymbol{\beta}))}{n+1}\Big)\nu_{ni}(\boldsymbol{\beta}), \tag{4}$$

where $\nu_{in}(\boldsymbol{\beta}) = Y_i - \widehat{g}^i_{\boldsymbol{\beta},h}(\mathbf{X}_i^\tau\boldsymbol{\beta})$. Based on (2) and (4), let $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$ be the rank estimators of $\boldsymbol{\beta}_0$ defined as

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathscr{B}}{\text{Argmin}}\, D_n(\boldsymbol{\beta}) \quad \text{and} \quad \widetilde{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathscr{B}}{\text{Argmin}}\, \widetilde{D}_n(\boldsymbol{\beta}).$$

Theorem 1 provides the asymptotic equivalence between the two objective functions given in Eqs. (2) and (4). The proof relies on Lemma 3 given in Appendix. The following assumptions are necessary and sufficient in the theoretical development of the paper.

Assumptions

(I_1) φ is a nondecreasing, bounded and twice continuously differentiable score function with bounded derivatives, defined on (0, 1), satisfying:

$$\int_0^1 \varphi(u)\mathrm{d}u = 0 \quad \text{and} \quad \int_0^1 \varphi^2(u)\mathrm{d}u = 1.$$

(I_2) As in (Delecroix et al. 2006), write $g_{\boldsymbol{\beta}}(t)$ as $g_{\boldsymbol{\beta}}(t) = E(Y|\mathbf{X}^\tau\boldsymbol{\beta} = t) =: r_{\boldsymbol{\beta}}(t)/\mu_{\boldsymbol{\beta}}(t)$. We assume that

(i) For every $\boldsymbol{\beta} \in \mathscr{B}$, $\mu_{\boldsymbol{\beta}}(\cdot)$ is the density of $\mathbf{X}^\tau \boldsymbol{\beta}$ with respect to the Lebesgue measure on $\mathbb{R}$, and there exists some constant c_0 such that $\{t : \mu_{\boldsymbol{\beta}}(t) = c\}$ is finite, for any c satisfying $0 < c_0 \leq c$.

(ii) The functions $(\boldsymbol{\beta}, t) \mapsto \mu_{\boldsymbol{\beta}}(t)$, $(\boldsymbol{\beta}, t) \mapsto r_{\boldsymbol{\beta}}(t)$ and $(\boldsymbol{\beta}, t) \mapsto g_{\boldsymbol{\beta}}(t)$ satisfy the so-called L-condition. Just considering $\mu_{\boldsymbol{\beta}}(t)$, the L-condition states that for all $t_1, t_2 \in \Lambda$, with Λ being a compact subset of $\mathbb{R}$, and $\forall\ \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathscr{B}$, there exist $C > 0$ independent of $(\boldsymbol{\beta}, t)$ and $a \in (0, 1]$ such that

$$|\mu_{\boldsymbol{\beta}_1}(t_1) - \mu_{\boldsymbol{\beta}_2}(t_2)| \leq C \|(\boldsymbol{\beta}_1, t_1) - (\boldsymbol{\beta}_2, t_2)\|^a.$$

(iii) For any $t \in \mathbb{R}$, $\boldsymbol{\beta} \to g_{\boldsymbol{\beta}}(t)$ is three time continuously differentiable, and there exists a function $J : \mathscr{X} \to \mathbb{R}^+$ not necessarily the same such that $|\nabla^r_{\boldsymbol{\beta}}[g_{\boldsymbol{\beta}}(\mathbf{x}^\tau \boldsymbol{\beta})]| \leq J(\mathbf{x})$ with $0 < E[J^p(\mathbf{X})] < \infty$, for $0 < p \leq 4$ and for all $\boldsymbol{\beta} \in \mathscr{B}$, where $r = 0, 1, 2, 3$, and $\nabla^r_{\boldsymbol{\beta}}$ is a differential operator of order r.

(I_3) $K(\cdot)$ is a three times differentiable, symmetric, positive and compactly supported kernel function with bandwidth $h_n \in \mathscr{H}_n$. Moreover, $K'(\cdot)$ is of bounded variation.

(I_4) $\sup_{\mathbf{x}} E[|Y|^q|\mathbf{X} = \mathbf{x}] < \infty$, for some $q \geq 2$ and $0 < E\{\exp(\lambda\|\mathbf{X}\|)\} < \infty$, for some $\lambda > 0$.

(I_5) $\boldsymbol{\beta}_0 \in Int(\mathscr{B})$ and for fixed n, $\boldsymbol{\beta}_{0,n}$ is the unique minimizer of $E[D_n(\boldsymbol{\beta})]$ such that $\boldsymbol{\beta}_0 = \lim_{n\to\infty} \boldsymbol{\beta}_{0,n}$.

(I_6) $n^{-1} \sum_{i=1}^n I_\Gamma(\mathbf{X}_i) \nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)]\{\nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)]\}^\tau \to \Sigma$ a.s., where Σ is positive definite matrix.

Remark 1 Assumption (I_1) is a common assumption in the rank estimation literature; see Hettmansperger and McKean (2011). Assumptions (I_2)–(I_4) ensure the strong consistency of the leave-one-out Nadaraya–Watson estimator of $g_0(\cdot)$ (Delecroix et al. 2003, 2006). (I_5) ensures the strong consistency of the proposed estimator, as established in Theorem 3. This assumption is assumed in most of the regression problems in a more stronger form; that is, $\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta}\in\mathscr{B}}{\text{Argmin}}\, E\{D_n(\boldsymbol{\beta})\}$ (Delecroix et al. 2006). Together with the previous assumptions, (I_6) is used to show the asymptotic normality of the gradient function. Also, (I_6) is a common regression assumption usually imposed on the design matrix and reduces to $n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\tau \to \Sigma = E\{\mathbf{X}\mathbf{X}^\tau\}$ that is imposed for the linear model: $g_{\boldsymbol{\beta}_0}(\mathbf{X}^\tau \boldsymbol{\beta}_0) = \mathbf{X}^\tau \boldsymbol{\beta}_0$ (Hettmansperger and McKean 2011).

Theorem 1 *Under assumptions* (I_1)–(I_4), $\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathscr{B}, h\in\mathscr{H}_n} |\widetilde{D}_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta})| = 0$ *a.s.*

Remark 2 Note that the functions $I_\Gamma(\mathbf{X})$ and $I_{\Gamma_n}(\mathbf{X})$ in Eqs. (2) and (4) are trimming devices introduced to keep the Nadaraya–Watson estimator away from zero (Delecroix et al. 2003, 2006). The existence of either $\widehat{\boldsymbol{\beta}}_n$ or $\widetilde{\boldsymbol{\beta}}_n$ is ensured by the continuity of the two objective functions over $\mathscr{B}$ compact. $\widehat{\boldsymbol{\beta}}_n$ is defined as a minimizer of an objective function that depends on an unknown function g_0; hence, although the dispersion function uses independent residuals, it cannot be computed. On the other hand, we can compute $\widetilde{\boldsymbol{\beta}}_n$, but the objective function involves dependent residuals as $\widehat{g}^i_{\boldsymbol{\beta},h}(\cdot), \ldots, \widehat{g}^i_{\boldsymbol{\beta},h}(\cdot)$ are not independent. As it will be demonstrated in Appendix, the two estimators possess the same asymptotic distribution. So for simplicity, we can use $\widetilde{\boldsymbol{\beta}}_n$ for asymptotic inference in practice.

3 Consistency

Let $(\Omega, \mathscr{F}, P)$ be a complete probability space. For $i = 1, \ldots, n$, assume that $\varepsilon_i(\boldsymbol{\beta}_0)$ are i.i.d. with common distribution at the true $\boldsymbol{\beta}_0$, and for $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, $z_i(\boldsymbol{\beta})$ are independent with distribution $F_i \equiv F_{i,\boldsymbol{\beta}}$. The following theorem gives the stochastic equicontinuity of the sequence of rank objective functions $\{D_n(\boldsymbol{\beta})\}_{n\geq 1}$ under the above assumptions.

Theorem 2 *Under (I_1), (I_2) and (I_4), and for every $\boldsymbol{\beta} \in \mathscr{B}$,*

$$D_n(\boldsymbol{\beta}) - E\{D_n(\boldsymbol{\beta})\} \to 0 \;\; a.s. \quad \text{as } n \to \infty.$$

Moreover, the sequence $\{D_n(\boldsymbol{\beta})\}_{n\geq 1}$ is stochastically equicontinuous.

This result is used along with Lemma 1 to provide the strong consistency of $\widehat{\boldsymbol{\beta}}_n$ that follows.

Theorem 3 *Under (I_1), (I_2), (I_4) and (I_5), $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ $a.s.$ as $n \to \infty$.*

As mentioned above, the proof of this theorem directly follows from Theorem 2 and the following lemma whose proof of (i) is given in Appendix and that of (ii) can be found in Andrews (1994), Newey and McFadden (1994) and Rao et al. (2014). So, for the sake of brevity, the proof of (ii) will not be provided.

Lemma 1 *Let $\{A_n(\alpha)\}_{n\geq 1}$ be a random objective function defined on a compact space Θ such that $\widehat{\alpha}_n = \underset{\alpha}{\text{Argmin}}\, A_n(\alpha)$ and for fixed n, there is a unique $\alpha_{0,n} \in \Theta$ satisfying $\alpha_{0,n} = \underset{\alpha}{\text{Argmin}}\, E(A_n(\alpha))$, with $E(A_n(\alpha))$ being continuous with respect to α. Assume for $\alpha_0 \in \Theta$, $\alpha_0 = \lim_{n\to\infty} \alpha_{0,n}$.*

- (i) *If $\sup_{\alpha\in\Theta} |A_n(\alpha) - E(A_n(\alpha))| \to 0$ $a.s.$ as $n \to \infty$, then $\widehat{\alpha}_n \to \alpha_0$ $a.s.$ as $n \to \infty$.*
- (ii) *If for every $\alpha \in \Theta$, $A_n(\alpha) - E(A_n(\alpha)) \to 0$ $a.s.$ as $n \to \infty$ and $A_n(\alpha)$ is stochastically equicontinuous, then $\widehat{\alpha}_n \to \alpha_0$ $a.s.$ as $n \to \infty$.*

Theorem 4 *Under assumptions (I_1)–(I_4),*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathscr{B}, h\in\mathscr{H}_n} |E(\widetilde{D}_n(\boldsymbol{\beta})) - E(D_n(\boldsymbol{\beta}))| = 0 \;\; a.s.$$

$$\text{and} \;\; \lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathscr{B}, h\in\mathscr{H}_n} |\widetilde{D}_n(\boldsymbol{\beta}) - E(\widetilde{D}_n(\boldsymbol{\beta}))| = 0 \;\; a.s.$$

Theorem 5 *Under the assumptions of Theorem 4 and (I_5), $\boldsymbol{\beta}_{0,n} = \underset{\boldsymbol{\beta}\in\mathscr{B}}{\text{Argmin}}\, E(\widetilde{D}_n(\boldsymbol{\beta}))$.*

These two theorems together with Lemma 1 imply that $\widetilde{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ $a.s.$ as $n \to \infty$.

4 Asymptotic normality

We will denote the gradient and Hessian operators by $\nabla_{\boldsymbol{\beta}} = \left(\partial/\partial\beta_i\right)_i$ and $\nabla^2_{\boldsymbol{\beta}} = \left(\partial^2/\partial\beta_i\partial\beta_j\right)_{ij}$, respectively, for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\tau$, $i, j = 1, \ldots, p$. Let $S_n(\boldsymbol{\beta}) = -\nabla_{\boldsymbol{\beta}} D_n(\boldsymbol{\beta})$. Since $\widehat{\boldsymbol{\beta}}_n$ is a minimizer of $D_n(\boldsymbol{\beta})$, we have $S_n(\widehat{\boldsymbol{\beta}}_n) = 0$. The gradient function is explicitly given as

$$S_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\nabla_{\boldsymbol{\beta}}[g_{\boldsymbol{\beta}}(\mathbf{X}_i^\tau\boldsymbol{\beta})]\varphi\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right).$$

Theorem 6 *Putting* $T_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n I_\Gamma(\mathbf{X}_i)\nabla_{\boldsymbol{\beta}}[g_{\boldsymbol{\beta}}(\mathbf{X}_i^\tau\boldsymbol{\beta})]\varphi\big(F_i(z_i(\boldsymbol{\beta}))\big)$, *we have under assumptions* (I_1), (I_2)*–(iii) and* (I_4) *that*

$$\lim_{n\to\infty}\sup_{\boldsymbol{\beta}\in\mathscr{B}}|S_n(\boldsymbol{\beta}) - T_n(\boldsymbol{\beta})| = 0 \; a.s.$$

This theorem implies that with probability 1, $S_n(\boldsymbol{\beta}) = T_n(\boldsymbol{\beta}) + o(1)$. Then, performing the second-order Taylor expansion of $T_n(\boldsymbol{\beta})$ at the true parameter $\boldsymbol{\beta}_0$ gives

$$T_n(\boldsymbol{\beta}) = T_n(\boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau\nabla_{\boldsymbol{\beta}_0}T_n(\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau\nabla^2_{\boldsymbol{\xi}}T_n(\boldsymbol{\xi})(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\xi} = \lambda\boldsymbol{\beta}_0 + (1-\lambda)\boldsymbol{\beta}$ for some $\lambda \in [0, 1]$. Thus, with probability 1,

$$S_n(\boldsymbol{\beta}) = T_n(\boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau\nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau\nabla^2_{\boldsymbol{\beta}}T_n(\boldsymbol{\xi})(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o(1).$$

From this, we have

$$\begin{aligned} 0 = S_n(\widehat{\boldsymbol{\beta}}_n) &= T_n(\boldsymbol{\beta}_0) + \nabla_{\boldsymbol{\beta}_0}T_n(\boldsymbol{\beta}_0)\cdot(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ &\quad + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\tau\cdot\nabla^2_{\boldsymbol{\xi}_n}T_n(\boldsymbol{\xi}_n)\cdot(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o(1), \end{aligned}$$

where $\boldsymbol{\xi}_n = \lambda\boldsymbol{\beta}_0 + (1-\lambda)\widehat{\boldsymbol{\beta}}_n$. This implies that

$$\left[\nabla_{\boldsymbol{\beta}_0}T_n(\boldsymbol{\beta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\tau\cdot\nabla^2_{\boldsymbol{\xi}_n}T_n(\boldsymbol{\xi}_n)\right](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = -S_n(\boldsymbol{\beta}_0) + o(1).$$

The asymptotic properties of the quantities in brackets above are established in the following theorem.

Theorem 7 *Under assumptions* (I_1)*–*(I_6), *the following hold*:

(a) $\nabla_{\boldsymbol{\beta}_0}T_n(\boldsymbol{\beta}_0) = \nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \to \mathbf{W}$ *a.s., where*

$$\begin{aligned} \mathbf{W} = &-E\{I_\Gamma(\mathbf{X})\nabla_{\boldsymbol{\beta}_0}(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0))[\nabla_{\boldsymbol{\beta}_0}(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau\boldsymbol{\beta}))]^\tau f(\varepsilon)\varphi'(F(\varepsilon))\} \\ &+ E\{I_\Gamma(\mathbf{X})\nabla^2_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0)]\varphi(F(\varepsilon))\}, \end{aligned}$$

where

$$\nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)] = \nabla_{\boldsymbol{\beta}}[g_{\boldsymbol{\beta}}(\mathbf{X}_i^\tau \boldsymbol{\beta})]|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \text{ and,}$$
$$\nabla^2_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)] = \nabla^2_{\boldsymbol{\beta}}[g_{\boldsymbol{\beta}}(\mathbf{X}_i^\tau \boldsymbol{\beta})]|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$$

is a positive definite matrix,

(b) $\nabla^2_{\boldsymbol{\xi}_n} T_n(\boldsymbol{\xi}_n) = \nabla^2_{\boldsymbol{\beta}} T_n(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\xi}_n}$ *is almost surely bounded.*

As consequence to this theorem, we have with probability 1

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = -\mathbf{W}^{-1}\sqrt{n}S_n(\boldsymbol{\beta}_0) + o(1), \tag{5}$$

and thus, the asymptotic normality of the rank estimator is derived from that of $\sqrt{n}S_n(\boldsymbol{\beta}_0)$. Note that if we assume that ε and $\mathbf{X}$ are independent, $\mathbf{W}$ can be expressed using the rank scale parameter as $\mathbf{W} = \gamma_\varphi^{-1}\Sigma$ similar to the linear model case, where

$$\Sigma = E\{I_\Gamma(\mathbf{X})\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\big)[\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\big)]^\tau\},$$

$\gamma_\varphi^{-1} = \int_0^1 \varphi(u)\varphi_f(u)\mathrm{d}u$, where $\varphi_f(u) = f'(F^{-1}(u))/f(F^{-1}(u))$, with f being the density of $\varepsilon(\boldsymbol{\beta}_0)$. The following two theorems give the asymptotic normality of $S_n(\boldsymbol{\beta}_0)$ and $\widehat{\boldsymbol{\beta}}_n$.

Theorem 8 *Under assumptions* (I_1)–(I_6), $\sqrt{n}S_n(\boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \Sigma)$ *as* $n \to \infty$.

Theorem 9 *Under assumptions* (I_1)–(I_6), $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N\big(\mathbf{0}, \mathbf{W}^{-1}\Sigma\mathbf{W}^{-1}\big)$. *If we restrict ourselves to the case where* $\mathbf{X}$ *and* ε *are independent,* $\mathbf{W}^{-1}\Sigma\mathbf{W}^{-1} = \gamma_\varphi^2\Sigma^{-1}$.

The proof of Theorem 9 is directly obtained by combining Eq. (5) and Theorem 8.

Now, define $M_n(\boldsymbol{\beta})$ and $\widetilde{M}_n(\boldsymbol{\beta})$ by

$$M_n(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau[\nabla_{\boldsymbol{\beta}_0} T_n(\boldsymbol{\beta}_0)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) - (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau S_n(\boldsymbol{\beta}_0) + D_n(\boldsymbol{\beta}_0)$$

and

$$\widetilde{M}_n(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau[\nabla_{\boldsymbol{\beta}_0} \widetilde{T}_n(\boldsymbol{\beta}_0)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) - (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau \widetilde{S}_n(\boldsymbol{\beta}_0) + \widetilde{D}_n(\boldsymbol{\beta}_0),$$

where $\widetilde{S}_n(\boldsymbol{\beta}) = -\nabla\widetilde{D}_n(\boldsymbol{\beta})$. This provides the following result, known as the asymptotic quadraticity, proved by Jaeckel (1972). Let $\mathscr{B}_n := \{\boldsymbol{\beta} \in \mathscr{B} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le c/\sqrt{n}\}$, for some positive constant c, by just taking $d_n = O(1/\sqrt{n})$ in Lemma 3.

Theorem 10 *Under* (I_1)–(I_6), $\forall\varepsilon > 0$, *we have*

$$\lim_{n\to\infty} P_{\boldsymbol{\beta}_0}\Big[\sup_{\boldsymbol{\beta}\in\mathscr{B}_n} |D_n(\boldsymbol{\beta}) - M_n(\boldsymbol{\beta})| > \varepsilon\Big] = 0 \quad a.s., \tag{6}$$

and

$$\lim_{n\to\infty} P_{\boldsymbol{\beta}_0}\Big[\sup_{\boldsymbol{\beta}\in\mathscr{B}_n, h\in\mathscr{H}_n} |\widetilde{D}_n(\boldsymbol{\beta}) - \widetilde{M}_n(\boldsymbol{\beta})| > \varepsilon\Big] = 0 \quad a.s. \tag{7}$$

Moreover,

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathscr{B}_n, h\in\mathscr{H}_n} |\widetilde{M}_n(\boldsymbol{\beta}) - M_n(\boldsymbol{\beta})| = 0. \quad a.s. \tag{8}$$

This theorem gives rise to the following lemma whose proof can be constructed by mimicking those given in Ichimura (1993) for the semiparametric least squares approach or in Delecroix et al. (2006) for M-estimation with some minor modifications. We only provide the proof of equation (9) in Appendix, and those of Eqs. (10) and (11) could obtained following similar arguments.

Lemma 2 *Under* (I_1)–(I_6), *and for any* $\varepsilon > 0$, *the following results are obtained in a straight forward manner*:

$$\lim_{n\to\infty} P\Big[\big\|\sqrt{n}\big(S_n(\boldsymbol{\beta}_0) - \widetilde{S}_n(\boldsymbol{\beta}_0)\big)\big\| > \varepsilon\Big] = 0, \tag{9}$$

$$\lim_{n\to\infty} P\Big\{\sup_{\boldsymbol{\beta}\in\mathscr{B}_n, h\in\mathscr{H}_n} \big\|\nabla_{\boldsymbol{\beta}} T_n(\boldsymbol{\beta}) - \nabla_{\boldsymbol{\beta}} \widetilde{T}_n(\boldsymbol{\beta})\big\| > \varepsilon\Big\} = 0, \tag{10}$$

and

$$\lim_{n\to\infty} P\Big\{\sup_{\boldsymbol{\beta}\in\mathscr{B}_n, h\in\mathscr{H}_n} \big\|\nabla^2_{\boldsymbol{\beta}} T_n(\boldsymbol{\beta}) - \nabla^2_{\boldsymbol{\beta}} \widetilde{T}_n(\boldsymbol{\beta})\big\| > \varepsilon\Big\} = 0 \quad a.s. \tag{11}$$

Remark 3 Using a similar argument to the Taylor expansion that leads to Eq. (5), one can easily show that

$$\sqrt{n}\big(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\big) = -\widetilde{\mathbf{W}}_n^{-1}\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0) + o_p(1), \tag{12}$$

where

$$\widetilde{\mathbf{W}}_n = -\frac{1}{n}\sum_{i=1}^{n} I_{\Gamma_n}(\mathbf{X}_i)\nabla_{\boldsymbol{\beta}_0}\big(\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0)\big)[\nabla_{\boldsymbol{\beta}_0}\big(\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0)\big)]^\tau f_\nu(\nu_{in}(\boldsymbol{\beta}_0))$$
$$\varphi'(F_\nu(\nu_{in}(\boldsymbol{\beta}_0))) + \frac{1}{n}\sum_{i=1}^{n} I_{\Gamma_n}(\mathbf{X}_i)\nabla^2_{\boldsymbol{\beta}_0}[\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0)]\varphi(F_\nu(\nu_{in}(\boldsymbol{\beta}_0)))$$

with F_ν being the cumulative distribution function of $\nu_{in}(\boldsymbol{\beta}_0)$, and f_ν the corresponding density. Also, from Eq. (10) and Theorem 7(a), and applying the dominated convergence theorem, we get for any $\varepsilon > 0$ and $h \in \mathscr{H}_n$,

$$\lim_{n\to\infty} P_{\boldsymbol{\beta}_0}\Big\{\big\|\widetilde{\mathbf{W}}_n - \mathbf{W}\big\| > \varepsilon\Big\} = 0. \tag{13}$$

This result is not surprising as it is also obtained by Newey (2004) who considered a generalized methods of moments (GMM) estimator and pointed out that the estimator obtained based on $\widehat{g}^i_{\boldsymbol{\beta}_0,h}(t)$ has the same asymptotic covariance matrix as the one based on $g_{\boldsymbol{\beta}_0}(t)$, $t \in \mathscr{A}_0$. Thus, both $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$ have similar asymptotic relative efficiency. For small sample sizes, however, one would expect $\widehat{\boldsymbol{\beta}}_n$ to be more efficient if $g_{\boldsymbol{\beta}_0}(\cdot)$ was assumed to be known, as the estimation of $g_{\boldsymbol{\beta}_0}(\cdot)$ might introduce some loss in efficiency.

Theorem 11 *Under assumptions* (I_1)–(I_6), $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N\big(\mathbf{0}, \mathbf{W}^{-1}\Sigma\mathbf{W}^{-1}\big)$. *Again, if we restrict ourselves to the case where* $\mathbf{X}$ *and* ε *are independent,* $\mathbf{W}^{-1}\Sigma\mathbf{W}^{-1} = \gamma_\varphi^2\Sigma^{-1}$.

Theorems 9 and 11 show that $\widetilde{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_n$ have the same asymptotic distribution.

5 Simulation and real data studies

5.1 Simulation

To illustrate the performance of the proposed method, a simulation study is conducted using the R software environment (R Development Core Team 2009). Three scenarios are considered:

Scenario 1: To demonstrate the robustness of the proposed procedure independent of the choice of the link function $g_{\boldsymbol{\beta}}(\cdot)$, the response variable is generated according to Eq. (1) with a variety of link functions, $g_{\boldsymbol{\beta}}(t)$, and different distributions for the model errors (ε). The error distributions studied are the contaminated normal distribution $\mathscr{C}\mathscr{N}(\varepsilon) = (1-\varepsilon)N(0,1) + \varepsilon N(0, 2^2)$ with different rates of contamination $\varepsilon = 0, 0.01, 0.05, 0.1, 0.15, 0.25$, the t and Chi-square distributions with different degrees of freedom ($df = 5, 15, 25, 35, 45$), and the Laplace distribution with different sample sizes ($n = 15, 35, 55, 75, 95, 115$). Except for the Laplace distribution case that has different sample sizes, the sample size considered is $n = 50$. The link functions used in the simulation are the identity link function defined by $g_{\boldsymbol{\beta}}(t) = t$, the inverse link function defined by $g_{\boldsymbol{\beta}}(t) = 1/t$, the logistic link function defined by $g_{\boldsymbol{\beta}}(t) = e^t/(1+e^t)$, and the sine link function $g_{\boldsymbol{\beta}}(t) = (\sin(2\pi t^2))^{1/3}$ with $t = \mathbf{x}^\tau\boldsymbol{\beta} = \mathrm{x}_1\beta_0 + \mathrm{x}_2\beta_1$ where the true $\boldsymbol{\beta}$ is set at $\boldsymbol{\beta}_0 = (1, \sqrt{1/3})^\tau$, $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2)$, with $\mathrm{x}_i \sim U_{(1,2)}$, for $i = 1, 2$.

Scenario 2: We considered the same error distributions as in Scenario 1 but with an increased sample of $n = 200$, for the $\mathscr{C}\mathscr{N}(\varepsilon)$, t and χ^2 distributions, and $n = 250, 350, 450$, for the Laplace distribution. Also, $t = \mathbf{x}^\tau\boldsymbol{\beta} = \mathrm{x}_1\beta_0 + \mathrm{x}_2\beta_1 + \beta_3\mathrm{x}_3 + \beta_4\mathrm{x}_4$ where the true $\boldsymbol{\beta}$ is set at $\boldsymbol{\beta}_0 = (1, \sqrt{3}/3, 0, 0)$ and $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4)$, with $\mathrm{x}_i \sim U_{(1,2)}$ for $i = 1, 2$ and $\mathrm{x}_i \sim U_{(2,5)}$, for $i = 3, 4$.

Scenario 3: We considered the same model error distributions as in Scenario 1 but we set the link function as $g_{\boldsymbol{\beta}}(t) = \sin\big(\pi(t-a)/(c-a)\big)$, where $a = \sqrt{3}/2 - 1.645/\sqrt{12}$, $c = \sqrt{3}/2 + 1.645/\sqrt{12}$. $t = \mathbf{x}^\tau\boldsymbol{\beta} = \beta_1\mathrm{x}_1 + \beta_2\mathrm{x}_2 + \beta_3\mathrm{x}_3$, where the true $\boldsymbol{\beta}$ is set at $\boldsymbol{\beta}_0 = (1, \sqrt{3}/3, \sqrt{3}/3)^\tau$ for the true parameter, and $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3)$ a matrix formed by the column vectors. To evaluate the performance of the procedures in the presence of categorical predictors, we took $\mathrm{x}_1 \sim U_{(0,1)}$, $\mathrm{x}_2 \sim \mathrm{Binomial}(n, 0.4)$ and $\mathrm{x}_3 \sim \mathrm{Binomial}(n, 0.7)$ with $n = 100$. The model is the same as the one used in (Liu et al. 2013), with the exception of the manner in which predictors are generated.

Table 1 Relative efficiencies (REs) of the rank estimator of $\boldsymbol{\beta}_0 = (1, \sqrt{3}/3)$ against the proportion of contamination (ε) for the contaminated normal distribution, against the sample sizes (n) for the Laplace distribution, and against the degrees of freedom (df) for the t and the Chi-square distributions

$g_{\boldsymbol{\beta}}(t)$				t		$\frac{e^t}{1+e^t}$		$1/t$		$\sin^{\frac{1}{3}}(2\pi t^2)$	
	n	ε	df	$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_0$	
$\mathscr{CN}(\varepsilon)$	50	0.00		1.165	1.157	1.057	1.050	1.050	1.133	0.926	0.916
		0.01		1.192	1.165	1.083	1.086	1.098	1.150	1.012	0.962
		0.05		1.197	1.200	1.096	1.090	1.099	1.268	1.012	1.072
		0.10		1.205	1.214	1.117	1.206	1.104	1.285	1.017	1.178
		0.15		1.231	1.275	1.143	1.227	1.106	1.350	1.051	1.235
		0.25		1.253	1.396	1.152	1.243	1.122	1.473	1.085	1.312
t_{df}	50		5	1.234	1.270	1.114	1.229	1.165	1.164	1.077	1.131
			15	1.176	1.180	1.105	1.227	1.121	1.137	1.027	1.052
			25	1.141	1.157	1.103	1.169	1.100	1.124	1.019	1.003
			35	1.125	1.064	1.091	1.122	1.096	1.049	1.004	0.996
			45	1.084	1.064	1.081	1.046	1.052	1.018	0.990	0.991
χ^2_{df}	50		5	1.216	2.006	1.160	2.194	1.251	2.023	1.179	1.727
			15	1.181	1.899	1.155	1.873	1.193	1.842	1.174	1.676
			25	1.174	1.850	1.142	1.736	1.186	1.767	1.159	1.536
			35	1.163	1.619	1.131	1.708	1.186	1.688	1.157	1.513
			45	1.162	1.329	1.117	1.393	1.157	1.552	1.135	1.206
Laplace	15			1.157	1.221	1.082	1.065	1.059	1.209	1.023	1.073
	35			1.258	1.231	1.109	1.277	1.095	1.231	1.033	1.125
	55			1.350	1.368	1.152	1.328	1.099	1.305	1.065	1.319
	75			1.504	1.379	1.165	1.338	1.191	1.323	1.152	1.350
	95			1.633	1.485	1.186	1.371	1.200	1.337	1.175	1.464
	115			1.949	1.593	1.208	1.574	1.202	1.379	1.188	1.467

In all three scenarios, the regression estimates were obtained using the nonlinear minimization routine implemented in the function `optim` contained in the "`stats`" package of R, where the score function in (2) was set to be the Wilcoxon defined as $\varphi(u) = \sqrt{12}\,(u - 0.5)$, and the kernel in (3) was taken to be the Epanechnikov kernel $K(u) = 0.75(1 - u^2)I(|u| \leq 1)$. Also the nonparametric estimator of $g_{\boldsymbol{\beta}}(\cdot)$ was obtained using the "`npksum`" function in the "`np`" package of R. As this estimation involves a bandwidth selection, we considered a joint minimization of $\widetilde{D}_n(\beta, h)$ with the starting value of $h \in \{h : c_1 n^{-\alpha_1} < h < c_2 n^{-\alpha_2}\}$, for some $c_1, c_2 > 0$, $1/8 < \alpha_1 < \alpha_2 < 1/4$.

From 1000 simulations, relative efficiencies (REs) of the rank estimator with respect to the SLS estimator were calculated by taking the ratio of the mean squared error of the SLS estimates to that of the rank estimates. These are displayed in Tables 1, 2 and Fig. 1. While the threshold set at 1 represents equal performance for the two estimators, the dashed line below 1 set at $3/\pi \approx 0.95$ is the theoretical asymptotic

Table 2 Relative efficiencies (REs) of the rank estimator of $\boldsymbol{\beta}_0 = (1, \sqrt{3}/3, 0, 0)$ against the proportion of contamination (ε) for the contaminated normal distribution, against the sample sizes (n) for the Laplace distribution and against the degrees of freedom (df) for the t and the Chi-square distributions

	n	ε	df	$g_{\boldsymbol{\beta}}(t) = t$ $\boldsymbol{\beta}_0$				$g_{\boldsymbol{\beta}}(t) = e^t/(1+e^t)$ $\boldsymbol{\beta}_0$				$g_{\boldsymbol{\beta}}(t) = 1/t$ $\boldsymbol{\beta}_0$				$g_{\boldsymbol{\beta}}(t) = (\sin(2\pi t^2))^{1/3}$ $\boldsymbol{\beta}_0$			
$\mathscr{CN}(\varepsilon)$	200	0.00		1.021	0.977	1.043	0.986	1.051	1.052	1.145	1.178	1.082	1.040	1.121	1.113	1.010	0.987	1.094	1.016
		0.01		1.075	1.020	1.066	0.993	1.131	1.107	1.159	1.182	1.129	1.076	1.130	1.151	1.022	0.999	1.208	1.083
		0.05		1.119	1.034	1.069	1.174	1.153	1.107	1.171	1.230	1.130	1.089	1.162	1.178	1.023	1.005	1.271	1.086
		0.10		1.148	1.071	1.123	1.177	1.240	1.111	1.184	1.243	1.166	1.119	1.191	1.239	1.030	1.010	1.309	1.169
		0.15		1.167	1.097	1.314	1.178	1.246	1.128	1.196	1.246	1.246	1.121	1.217	1.257	1.032	1.028	1.319	1.172
		0.25		1.291	1.146	1.457	1.341	1.327	1.149	1.234	1.249	1.285	1.125	1.318	1.289	1.050	1.032	1.610	1.287
t_{df}	200		5	1.193	1.138	1.299	1.422	1.153	1.130	1.286	1.192	1.388	1.104	1.263	1.213	1.081	1.049	1.914	1.798
			15	1.087	1.034	1.098	1.253	1.153	1.125	1.196	1.180	1.105	1.094	1.193	1.179	1.065	1.041	1.793	1.715
			25	1.029	1.003	1.070	1.046	1.144	1.104	1.183	1.180	1.079	1.081	1.178	1.140	1.050	1.035	1.773	1.578
			35	0.988	0.998	1.063	1.039	1.123	1.101	1.180	1.167	1.058	1.058	1.165	1.131	1.042	1.025	1.720	1.486
			45	0.978	0.989	1.019	1.026	1.118	1.092	1.171	1.157	1.028	1.058	1.151	1.103	1.033	1.007	1.704	1.425
χ^2_{df}	200		5	1.399	1.099	1.160	1.210	1.256	1.102	1.175	1.116	1.332	1.120	1.109	1.099	1.075	1.033	1.438	2.249
			15	1.221	1.067	1.033	1.069	1.219	1.099	1.040	1.005	1.198	1.102	1.025	1.086	1.065	1.026	1.129	1.840
			25	1.208	1.056	1.020	1.018	1.210	1.071	1.022	0.997	1.182	1.074	1.009	1.023	1.062	1.020	1.090	1.788
			35	1.192	1.050	1.000	1.002	1.204	1.060	0.999	0.982	1.162	1.056	1.009	0.996	1.059	1.007	1.074	1.768
			45	1.166	1.035	0.965	0.996	1.198	1.039	0.979	0.972	1.158	1.043	0.975	0.990	1.053	0.986	1.043	1.659
Laplace	250			1.592	1.280	1.807	1.741	1.265	1.094	1.244	1.236	1.359	1.132	1.269	1.228	1.049	1.041	1.604	1.478
	350			1.606	1.371	1.905	1.787	1.308	1.096	1.275	1.311	1.433	1.135	1.289	1.305	1.067	1.050	2.240	2.249
	450			1.704	1.375	1.964	1.790	1.344	1.170	1.284	1.316	1.442	1.169	1.399	1.363	1.072	1.064	3.693	3.895

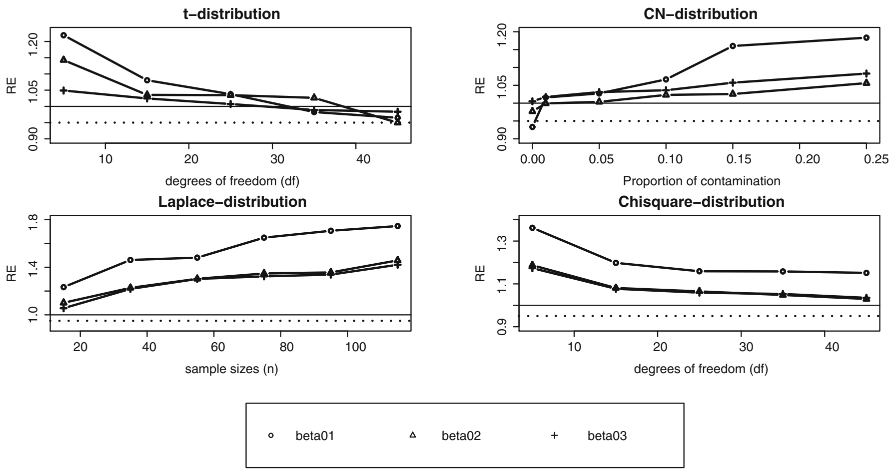

Fig. 1 Relative efficiencies (REs) of the rank estimators of the three true parameters against the proportion of contamination for the contaminated normal distribution, against the sample sizes for the Laplace distribution, and against the degrees of freedom for the t and the Chi-square distributions

relative efficiency (ARE) at the standard normal distribution. This is also the theoretical ARE for t and Chi-square distributions as $df \to \infty$. For the Laplace distribution, the theoretical ARE is 1.57.

Tables 1 and 2 clearly show that the rank estimator is generally more efficient than the SLS estimator. As expected, the REs increase as the rate of contamination increases for the $\mathscr{CN}(\varepsilon)$ distribution case while they decrease as df increases for the t and Chi-square distribution cases. For the $\mathscr{CN}$ case with $\varepsilon = 0$ (the standard normal distribution), the RE is close to the ARE of 0.95 as expected. The same is true for increasing df of t and Chi-square distributions. For the Laplace distribution, the REs increase as the sample size (n) increases approaching the theoretical ARE of 1.57. One also notes that the four considered link functions perform similarly, in line with our stipulation above that the efficiency comparison would be independent of the choice of link function.

Figure 1 displays REs of the rank estimator with respect to the SLS estimator under Scenario 3. Our observations remain the same as in Scenarios 1 and 2.

5.2 Real data example

We considered the county demographic information (CDI) data from the Geospatial and Statistical Data Center of the University of Virginia provided in Kutner et al. (2004). The data consist of information on 440 of the most populous counties in the USA pertaining to the years 1990 and 1992. We considered the response Y to be the logarithm of the number of active physicians and the predictors to be the total population, total personal income, percent of population over 65, number of hospital beds, land area and total number of serious crimes. Los Angeles County, CA, and Cook County, IL, represented obvious outliers in the Y direction with disproportionately high number of physicians. The predictor matrix was centered and scaled to have

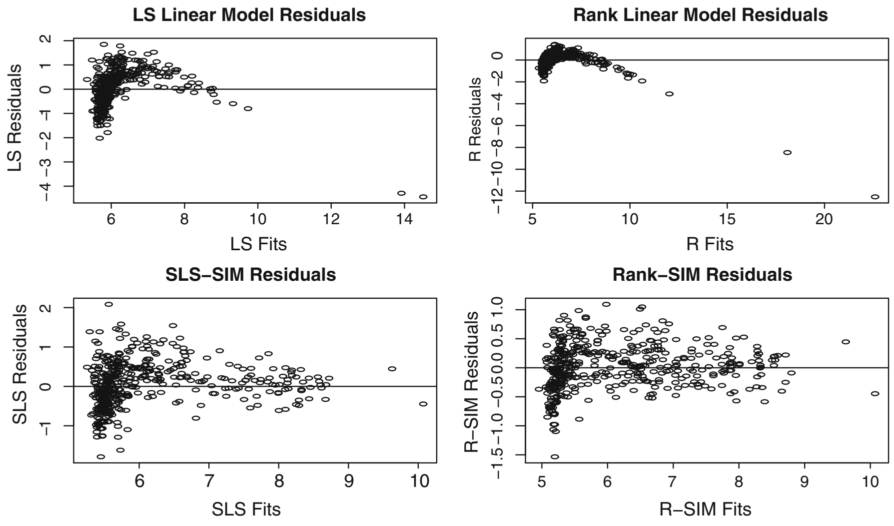

Fig. 2 Residual plots for the original CDI data

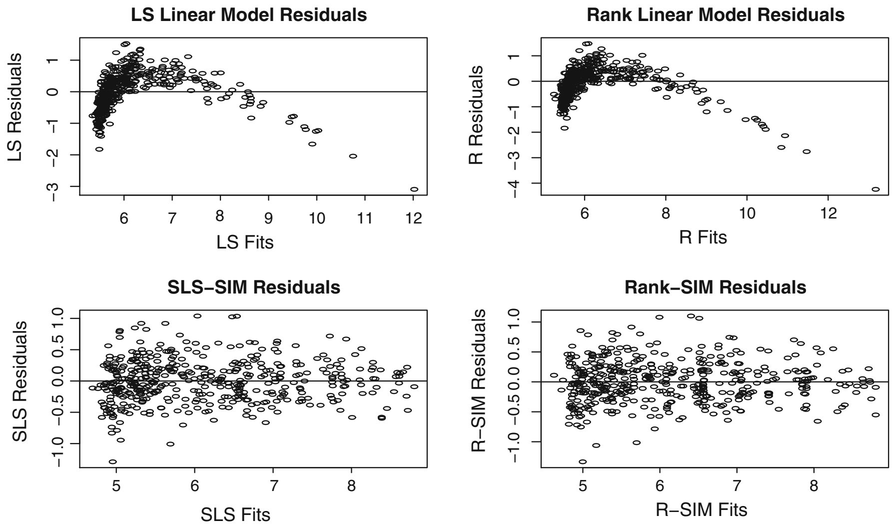

Fig. 3 Residual plots for the CDI data with outliers removed

mean zero and variance 1. We first fitted linear regression models using least squares (`lm` in R) and Wilcoxon rank regression (`Rfit` in R) including all the data. We also fitted both the semiparametric least squares and rank single index models for the data. This process was repeated after removing the two outliers. Residual plots from these fits are given in Figs. 2 and 3, where the top panels represent least squares and rank linear regression residual plots while the bottom panes give the semiparametric least squares and rank single index residual plots.

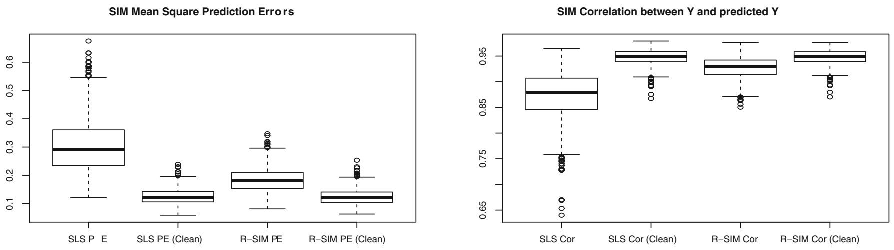

Fig. 4 Cross-validation results for the CDI data with and without outliers

Table 3 Mean prediction error, median prediction error and the correction between Y and $\widehat{Y}$ from the 500 replications

	Mean prediction error	Median prediction error	Correlation coefficient
SLS-outlier	0.310	0.290	0.869
Rank-SIM outlier	0.184	0.180	0.926
SLS-clean	0.128	0.125	0.946
Rank-SIM clean	0.127	0.124	0.947

The top panels of Figs. 2 and 3 clearly indicate that a linear model fit is not adequate. The appropriate model appears to be nonlinear. The bottom panels of the figures show that the residuals from the single index models have no nonlinearity indicating that the models capture the underlying nonlinear pattern of relationship between Y and the predictors. Thus, single index models clearly provide a superior analysis approach for these data compared to linear models.

To evaluate the predictive performance of these single index models, we performed a cross-validation study. The data were randomly split into 10 parts. Nine are used as training set to fit the model while one was used as the testing set. We performed 500 replications and calculated the mean square prediction error by computing $\texttt{mean}(Y - \widehat{Y})^2$ and $\texttt{median}(Y - \widehat{Y})^2$ on the testing set. We also computed a simple Pearson correlation of Y and $\widehat{Y}$ on the testing set to evaluate the linearity of the fit. The results are displayed in Fig. 4 and Table 3.

For the original data, the semiparametric least squares fit single index model gave higher prediction error and correspondingly lower correlation between true and predicted responses compared to the rank fit single index model. This indicates that the existence of outliers reduces the ability of the semiparametric least squares procedure to capture the correct underlying functional relationship. On the other hand, both the semiparametric least squares and rank single index fits provided low prediction errors in the absence of gross outliers. From the high correlation between the predicted and true responses, we can infer that they also capture the functional relationship effectively when the obvious outliers are removed from the data. The change in fit was much more pronounced for the semiparametric least squares fit indicating its sensitivity to outlying observations. These same observations are made considering the results in Table 3, as it is observed that in the presence of outliers (SLS-Outlier, Rank-SIM Outlier) in the response space, the SLS provides higher average (median) prediction error compared to rank-based approach, which confirm the robustness of the rank-based estimator when dealing with outliers in the response space. For the same scenario, based on the estimated correlation coefficients, the rank-based provides a better fit than the SLS. However, the two approaches show similar performance, once the outliers are removed (SLS-Clean, Rank-SIM Clean).

Conclusion

The rank regression procedure proposed in this paper provides a robust and efficient alternative to the least squares method for the fitting of single index models. One should note, however, that robustness is only in the direction of the response. As in least squares case, high-leverage points would still unduly affect the performance of the proposed method. For studies with data containing high-leverage points, one may consider weighted versions of the rank objective function such as the GR weighted Wilcoxon (Naranjo and Hettmansperger 1994), HBR weighted Wilcoxon (Chang et al. 1999; Abebe and McKean 2013) or the general weighted signed-rank (Bindele and Abebe 2012).

Appendix

This appendix contains proofs of the theoretical main results together with a key Lemma due to Delecroix et al. (2006) that ensures the uniform strong consistency of the leave-one-out Nadaraya–Watson estimator. For details regarding the proof of this lemma, readers are referred to the aforementioned paper.

Lemma 3 *Let $\mathscr{B}_n =: \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq d_n\}$, where d_n is some sequence decreasing to zero. Then,*

(a) *if $\delta > 0$, we have,*

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}_n, h\in\mathscr{H}_n} \left| I_{\{\mathbf{x}:\widehat{\mu}^i_{\boldsymbol{\beta},h}(\mathbf{x}^\tau\boldsymbol{\beta})\geq c\}}(\mathbf{X}_i) - I_\Gamma(\mathbf{X}_i) \right| \leq I_{\Gamma^\delta}(\mathbf{X}_i) + I_{(\delta,\infty)}(Z_n),$$

where $\Gamma^\delta = \{\mathbf{x} : |\mu_{\boldsymbol{\beta}_0,h}(\mathbf{x}^\tau\boldsymbol{\beta}_0) - c| \leq \delta\}$ and

$$Z_n = \max_{1\leq i\leq n} \sup_{\boldsymbol{\beta}\in\mathscr{B}_n, h\in\mathscr{H}_n} \left| \widehat{\mu}^i_{\boldsymbol{\beta},h}(\mathbf{X}_i^\tau\boldsymbol{\beta}) - \mu_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0) \right|.$$

(b) *Assume $d_n = o(1/\sqrt{n})$, and there exists a sequence $\delta_n \to 0$ such that $\delta_n/n^{-a\varepsilon} \to \infty$ and $\delta_n[d_n\sqrt{n}]^{-a\varepsilon} \to \infty$, for some $a > 0$, then $I_{(\delta_n,\infty)}(Z_n) = o_p(n^{-\alpha})$, for all $\alpha > 0$. Moreover, together with assumptions (I_2)–(I_4), assuming that $E(|Y|^2) < \infty$, we have*

$$\max_{1\leq i\leq n} \sup_{\boldsymbol{\beta}\in\mathscr{B}, h\in\mathscr{H}_n} |\widehat{g}^i_{\boldsymbol{\beta},h}(\mathbf{X}_i\boldsymbol{\beta}) - g_{\boldsymbol{\beta}}(\mathbf{X}_i\boldsymbol{\beta})| I_\Gamma(\mathbf{X}_i) \to 0 \quad a.s. \text{ as } n \to \infty,$$

and

$$\max_{1\le i\le n}\ \sup_{\boldsymbol\beta\in\mathscr{B},h\in\mathscr{H}_n} |\nabla_{\boldsymbol\beta}[\widehat{g}^{i}_{\boldsymbol\beta,h}(\mathbf{X}_i\boldsymbol\beta)] - \nabla_{\boldsymbol\beta}[g_{\boldsymbol\beta}(\mathbf{X}_i\boldsymbol\beta)]| I_{\Gamma}(\mathbf{X}_i) \to 0 \quad a.s.\ as\ n\to\infty.$$

Proof of Lemma 1 (i) By definition, we have $A_n(\hat\alpha_n) \le A_n(\alpha_{0,n})$ and $E(A_n(\alpha_{0,n})) \le E(A_n(\hat\alpha_n))$. These inequalities give $A_n(\hat\alpha_n) - E(A_n(\hat\alpha_n)) \le A_n(\hat\alpha_n) - E(A_n(\alpha_{0,n})) \le A_n(\alpha_{0,n}) - E(A_n(\alpha_{0,n}))$. Thus,

$$\begin{aligned}|A_n(\hat\alpha_n) - E(A_n(\alpha_{0,n}))| &\le \max\{|A_n(\hat\alpha_n) - E(A_n(\hat\alpha_n))|, |A_n(\alpha_{0,n}) - E(A_n(\alpha_{0,n}))|\}\\ &\le \sup_{\alpha\in\Theta} |A_n(\alpha) - E(A_n(\alpha))|.\end{aligned}$$

Since $\alpha_{0,n}$ is unique for any fixed n, $\alpha_{0,n} \to \alpha_0$ and $\sup_{\alpha\in\Theta} |A_n(\alpha) - E(A_n(\alpha))| \to 0$ $a.s.$ as $n\to\infty$, we have $\hat\alpha_n \to \alpha_0$ $a.s.$ as $n\to\infty$. □

Proof of Lemma 2 We provide the proof of equation (9) and those of Eqs. (10) and (11) could be obtained using similar arguments. By Chebyshev's inequality, we have, for any $\varepsilon > 0$,

$$P_{\boldsymbol\beta_0}\left(\sqrt{n}\|\widetilde{S}_n(\boldsymbol\beta_0) - S_n(\boldsymbol\beta_0)\| > \varepsilon\right) \le \frac{1}{\varepsilon^2} E\left\{n\left\|\widetilde{S}_n(\boldsymbol\beta_0) - S_n(\boldsymbol\beta_0)\right\|^2\right\}.$$

Setting $a_{ni}(\boldsymbol\beta_0) = R(v_{ni}(\boldsymbol\beta_0))/(n+1)$, $b_{ni}(\boldsymbol\beta_0) = R(z_i(\boldsymbol\beta_0))/(n+1)$, let us introduce the following notation: $\psi_i(\boldsymbol\beta_0) = \varphi(a_{ni}(\boldsymbol\beta_0)) - \varphi(b_{ni}(\boldsymbol\beta_0))$ and $U_i(\boldsymbol\beta_0) = I_{\Gamma_n}(\mathbf{X}_i)\nabla_{\boldsymbol\beta_0}[\widehat{g}^{i}_{\boldsymbol\beta_0,h}(\mathbf{X}_i^{\tau}\boldsymbol\beta_0)] - I_{\Gamma}(\mathbf{X}_i)\nabla_{\boldsymbol\beta_0}[g_{\boldsymbol\beta_0}(\mathbf{X}_i^{\tau}\boldsymbol\beta_0)]$.

$$\begin{aligned}&E\left\{n\left\|\widetilde{S}_n(\boldsymbol\beta_0) - S_n(\boldsymbol\beta_0)\right\|^2\right\}\\ &= \frac{1}{n\varepsilon^2} E\left[\left(\sum_{i=1}^{n}\left\{I_{\Gamma_n}(\mathbf{X}_i)\nabla_{\boldsymbol\beta_0}[\widehat{g}^{i}_{\boldsymbol\beta_0,h}(\mathbf{X}_i^{\tau}\boldsymbol\beta_0)]\varphi\left(a_{ni}(\boldsymbol\beta_0)\right)\right.\right.\right.\\ &\quad\left.\left.\left. - I_{\Gamma}(\mathbf{X}_i)\nabla_{\boldsymbol\beta_0}[g_{\boldsymbol\beta_0}(\mathbf{X}_i^{\tau}\boldsymbol\beta_0)]\varphi\left(b_{ni}(\boldsymbol\beta_0)\right)\right\}\right)^2\right]\\ &= \frac{1}{\varepsilon^2} E\left[\frac{1}{n}\sum_{i=1}^{n} U_i^2(\boldsymbol\beta_0)\varphi^2\left(a_{ni}(\boldsymbol\beta_0)\right)\right]\\ &\quad + \frac{1}{\varepsilon^2} E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\nabla_{\boldsymbol\beta_0}[g_{\boldsymbol\beta_0}(\mathbf{X}_i^{\tau}\boldsymbol\beta_0)]\right\}^2 \psi_i^2(\boldsymbol\beta_0)\right]\\ &\quad + \frac{2}{\varepsilon^2} E\left[\frac{1}{n}\sum_{i<j}^{n}\left\{\nabla_{\boldsymbol\beta_0}[g_{\boldsymbol\beta_0}(\mathbf{X}_i^{\tau}\boldsymbol\beta_0)]\varphi\left(a_{nj}(\boldsymbol\beta_0)\right)\right\} U_i(\boldsymbol\beta_0)\psi_j(\boldsymbol\beta_0)\right]\\ &= J_{1n} + J_{2n} + J_{3n}.\end{aligned}$$

We now show that $J_{in} \to 0$ as $n \to \infty$, for $i = 1, 2, 3$. Indeed, from the boundedness of φ, there exists a positive constant L such that $|\varphi(u)| \le L$, for all $u \in (0, 1)$. Also

$$|\psi_i(\boldsymbol{\beta}_0)| \le |\varphi(a_{ni}(\boldsymbol{\beta}_0)) - \varphi(F_\nu(\nu_{ni}(\boldsymbol{\beta}_0)))| \\ + |\varphi(F_\nu(\nu_{ni}(\boldsymbol{\beta}_0))) - \varphi(F(z_i(\boldsymbol{\beta}_0)))| + |\varphi(F(z_i(\boldsymbol{\beta}_0))) - \varphi(b_{ni}(\boldsymbol{\beta}_0))|.$$

For $i = 1, \ldots, n$, $F_\nu(\nu_{ni}(\boldsymbol{\beta}_0))$ and $F(z_i(\boldsymbol{\beta}_0))$ are independent uniformly distributed in (0, 1) random variables. Following Chapter 6 of Hájek et al. (1999), it is obtained that $\nu_{ni}(\boldsymbol{\beta}_0) - F_\nu(\nu_{ni}(\boldsymbol{\beta}_0)) \to 0$ $a.s.$ and $b_{ni}(\boldsymbol{\beta}_0) - F(z_i(\boldsymbol{\beta}_0)) \to 0$ $a.s.$, for each i. Thus, by continuity of φ and by Lemma 3, we have $\varphi(a_{ni}(\boldsymbol{\beta}_0)) - \varphi(F_\nu(\nu_{ni}(\boldsymbol{\beta}_0))) \to 0$ $a.s.$ and $\varphi(F(z_i(\boldsymbol{\beta}_0))) - \varphi(b_{ni}(\boldsymbol{\beta}_0)) \to 0$ $a.s.$, for each i. Also, by Lemma 3, we have $\nu_{ni}(\boldsymbol{\beta}_0) - z_i(\boldsymbol{\beta}_0) \to 0$ $a.s.$, from which by the continuity of the probability measure and the continuity of φ, we have $\varphi(F_\nu(\nu_{ni}(\boldsymbol{\beta}_0))) - \varphi(F(z_i(\boldsymbol{\beta}_0))) \to 0$ $a.s.$, for each i. On the other hand,

$$\|U_i(\boldsymbol{\beta}_0)\| \le |I_{\Gamma_n}(\mathbf{X}_i) - I_\Gamma(\mathbf{X}_i)| \|\nabla_{\boldsymbol{\beta}_0}[\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\| + \|\nabla_{\boldsymbol{\beta}_0}[\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)] \\ - \nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)]\| I_\Gamma(\mathbf{X}_i).$$

For $\mathbf{X}_i \in \Gamma$ and for all $\varepsilon > 0$, there exists $N > 0$ such that for all $n \ge N$,

$$\|\nabla_{\boldsymbol{\beta}_0}[\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\| < \|\nabla_{\boldsymbol{\beta}_0}[g(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)]\| + \varepsilon \le J(\mathbf{X}_i) + \varepsilon.$$

ε being arbitrary, letting $\varepsilon \to 0$, we have $\|\nabla_{\boldsymbol{\beta}_0}[\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\| \le J(\mathbf{X}_i) < \infty$ $a.s.$, as J is integrable. Thus, by Lemma 3, $|I_{\Gamma_n}(\mathbf{X}_i) - I_\Gamma(\mathbf{X}_i)| \|\nabla_{\boldsymbol{\beta}_0}[\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\| \to 0$ $a.s.$ and $\|\nabla_{\boldsymbol{\beta}_0}[\widehat{g}^i_{\boldsymbol{\beta}_0,h}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)] - \nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)]\| I_\Gamma(\mathbf{X}_i) \to 0$ $a.s.$, for all i. Therefore, $\|U_i(\boldsymbol{\beta}_0)\| \to 0$ $a.s.$, for all i. Then

$$\|J_{1n}\| \le \frac{L^2}{\varepsilon^2} E\left(\max_{1\le i\le n} \|U_i(\boldsymbol{\beta}_0)\|^2\right) \to 0 \ a.s.,$$

by applying the dominated convergence theorem together with Lemma 3. Next, using Cauchy–Schwarz inequality, we have

$$\begin{aligned}\| J_{2n} \| &\le \frac{1}{\varepsilon^2} E\left[\frac{1}{n}\sum_{i=1}^n J^2(\mathbf{X}_i)\, |\psi_i(\boldsymbol{\beta}_0)|^2\right] \\ &\le \frac{1}{\varepsilon^2} E\left[\left(\frac{1}{n}\sum_{i=1}^n J^4(\mathbf{X}_i)\right)^{1/2} \left(\max_{1\le i\le n} |\psi_i(\boldsymbol{\beta}_0)|^4\right)^{1/2}\right].\end{aligned}$$

By the strong law of large numbers (SLLN), $n^{-1}\sum_{i=1}^n J^4(\mathbf{X}_i) \to E\{J^4(\mathbf{X})\} < \infty$ $a.s.$ Also, from the above discussion, $\max_{1\le i\le n} |\psi_i(\boldsymbol{\beta}_0)|^4 \to 0$ $a.s.$ Thus, applying the dominated convergence theorem once again, we have $J_{2n} \to 0$ $a.s.$ Moreover, using the simple inequality $ab \leq (a^2+b^2)/2$ together with Cauchy–Schwarz inequality, we have

$$\begin{aligned}\|J_{3n}\| &\leq \frac{2L}{\varepsilon^2} E\left[\frac{1}{n}\sum_{i<j}^{n} J(\mathbf{X}_i)\,\|U_i(\boldsymbol{\beta}_0)\|\,|\psi_i(\boldsymbol{\beta}_0)|\right] \\ &\leq \frac{L}{\varepsilon^2} E\left[\frac{1}{n}\sum_{i=1}^{n} J^2(\mathbf{X}_i)\,\|U_i(\boldsymbol{\beta}_0)\|^2\right] + \frac{L}{\varepsilon^2} E\left[\frac{1}{n}\sum_{j=1}^{n} |\psi_i(\boldsymbol{\beta}_0)|^2\right] \\ &\leq \frac{L}{\varepsilon^2} E\left[\left(\frac{1}{n}\sum_{i=1}^{n} J^4(\mathbf{X}_i)\right)^{1/2}\left(\max_{1\leq i\leq n}\|U_i(\boldsymbol{\beta}_0)\|^4\right)^{1/2}\right] \\ &\quad + \frac{L}{\varepsilon^2} E\left[\max_{1\leq j\leq n}|\psi_i(\boldsymbol{\beta}_0)|^2\right].\end{aligned}$$

By Lemma 3, $\max_{1\leq i\leq n}\|U_i(\boldsymbol{\beta}_0)\|^4 \to 0\ a.s.$, and again, by the SLLN, $\frac{1}{n}\sum_{i=1}^{n} J^4(\mathbf{X}_i)$ converges almost surely to $E\{J^4(\mathbf{X})\} < \infty$. Also, as before, $\max_{1\leq i\leq n}|\psi_i(\boldsymbol{\beta}_0)|^2 \to 0\ a.s.$ To this end, once again, a direct application of the dominated convergence theorem gives $J_{3n} \to 0\ a.s.$ and consequently, $\lim_{n\to\infty} P_{\boldsymbol{\beta}_0}\left(\sqrt{n}\|\widetilde{S}_n(\boldsymbol{\beta}_0) - S_n(\boldsymbol{\beta}_0)\| > \varepsilon\right) = 0.$ □

Proof of Theorem 1 In this proof, we take L to be an arbitrary positive constant not necessarily the same, and as in the proof of Lemma 2, set $b_{ni}(\boldsymbol{\beta}) = R(\nu_{ni}(\boldsymbol{\beta}))/(n+1)$ and $a_{ni}(\boldsymbol{\beta}) = R(z_i(\boldsymbol{\beta}))/(n+1)$. By definition of $\widetilde{D}_n(\boldsymbol{\beta})$ and $D_n(\boldsymbol{\beta})$, we have

$$\begin{aligned}\widetilde{D}_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[I_{\Gamma_n}(\mathbf{X}_i) - I_{\Gamma}(\mathbf{X}_i)\right]\varphi(b_{ni}(\boldsymbol{\beta}))\nu_{ni}(\boldsymbol{\beta}) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} I_{\Gamma}(\mathbf{X}_i)\left[\varphi(b_{ni}(\boldsymbol{\beta}))\nu_{ni}(\boldsymbol{\beta}) - \varphi(a_{ni}(\boldsymbol{\beta}))z_i(\boldsymbol{\beta})\right].\end{aligned} \tag{14}$$

Considering the first term to the right-hand side of Eq. (14), we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\left[I_{\Gamma_n}(\mathbf{X}_i) - I_{\Gamma}(\mathbf{X}_i)\right]\varphi(b_{ni}(\boldsymbol{\beta}))\nu_{ni}(\boldsymbol{\beta})\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i) - I_{\Gamma}(\mathbf{X}_i)\right|\,|\varphi(b_{ni}(\boldsymbol{\beta}))|\,|\nu_{ni}(\boldsymbol{\beta})| \\ &\leq \frac{L}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i) - I_{\Gamma}(\mathbf{X}_i)\right|\,|\nu_{ni}(\boldsymbol{\beta})| \\ &\leq \frac{L}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i) - I_{\Gamma}(\mathbf{X}_i)\right|\,|Y_i|\end{aligned}$$

$$+\frac{L}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right||\widehat{g}^{i}_{\boldsymbol{\beta},h}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})-g_{\boldsymbol{\beta}}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})|$$
$$+\frac{L}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right||g_{\boldsymbol{\beta}}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})|,$$

where L is the bound of φ, as φ is assumed bounded by assumption (I_1). By Cauchy–Schwarz inequality, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right||Y_i|\le\left(\frac{1}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}|Y_i|^2\right)^{1/2}.$$

The strong law of large numbers gives $n^{-1}\sum_{i=1}^{n}|Y_i|^2\to E[|Y|^2]<\infty\ a.s.$ On the other hand, we have,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right|^2&\le\max_{1\le i\le n}\sup_{h\in\mathscr{H}_n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right|^2\\&\le\max_{1\le i\le n}\sup_{\boldsymbol{\beta}\in\mathscr{B}_n,h\in\mathscr{H}_n}\left|I_{\{\mathbf{x}:\widehat{\mu}^{i}_{\boldsymbol{\beta},h}(\mathbf{x}^{\tau}\boldsymbol{\beta})\ge c\}}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right|^2\\&\to 0\quad a.s.,\end{aligned}$$

as $n\to\infty$, by Lemma 3. Similarly,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right||\widehat{g}^{i}_{\boldsymbol{\beta},h}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})-g_{\boldsymbol{\beta}}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})|\\&\le\left(\frac{1}{n}\sum_{i=1}^{n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}|\widehat{g}^{i}_{\boldsymbol{\beta},h}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})-g_{\boldsymbol{\beta}}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})|^2I_{\Gamma}(\mathbf{X}_i)\right)^{1/2}\\&\le\left(\max_{1\le i\le n}\sup_{\boldsymbol{\beta}\in\mathscr{B}_n,h\in\mathscr{H}_n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right|^2\right)^{1/2}\\&\quad\times\left(\max_{1\le i\le n}\sup_{\boldsymbol{\beta}\in\mathscr{B}_n,h\in\mathscr{H}_n}|\widehat{g}^{i}_{\boldsymbol{\beta},h}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})-g_{\boldsymbol{\beta}}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})|^2I_{\Gamma}(\mathbf{X}_i)\right)^{1/2}.\end{aligned}$$

Again, by Lemma 3,

$$\max_{1\le i\le n}\sup_{\boldsymbol{\beta}\in\mathscr{B}_n,h\in\mathscr{H}_n}\left|I_{\Gamma_n}(\mathbf{X}_i)-I_{\Gamma}(\mathbf{X}_i)\right|^2\to 0\ a.s.$$

and

$$\max_{1\le i\le n}\sup_{\boldsymbol{\beta}\in\mathscr{B}_n,h\in\mathscr{H}_n}|\widehat{g}^{i}_{\boldsymbol{\beta},h}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})-g_{\boldsymbol{\beta}}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})|^2I_{\Gamma}(\mathbf{X}_i)\to 0\ a.s.$$

Thus, $n^{-1}\sum_{i=1}^n \left|I_{\Gamma_n}(\mathbf{X}_i) - I_\Gamma(\mathbf{X}_i)\right| |\widehat{g}^i_{\boldsymbol\beta,h}(\mathbf{X}_i^\tau\boldsymbol\beta) - g_{\boldsymbol\beta}(\mathbf{X}_i^\tau\boldsymbol\beta)| \to 0\ a.s.$ Moreover,

$$\frac{1}{n}\sum_{i=1}^n \left|I_{\Gamma_n}(\mathbf{X}_i) - I_\Gamma(\mathbf{X}_i)\right| |g_{\boldsymbol\beta}(\mathbf{X}_i^\tau\boldsymbol\beta)| \le \left(\frac{1}{n}\sum_{i=1}^n \left|I_{\Gamma_n}(\mathbf{X}_i) - I_\Gamma(\mathbf{X}_i)\right|^2\right)^{1/2} \times \left(\frac{1}{n}\sum_{i=1}^n |g_{\boldsymbol\beta}(\mathbf{X}_i^\tau\boldsymbol\beta)|^2\right)^{1/2}.$$

Following the same argument as above, we have $n^{-1}\sum_{i=1}^n \left|I_{\Gamma_n}(\mathbf{X}_i) - I_\Gamma(\mathbf{X}_i)\right|^2 \to 0\ a.s.$, and a direct application of the strong of large numbers gives

$$\frac{1}{n}\sum_{i=1}^n |g_{\boldsymbol\beta}(\mathbf{X}_i^\tau\boldsymbol\beta)|^2 \to E\{|g_{\boldsymbol\beta}(\mathbf{X}^\tau\boldsymbol\beta)|^2\} \le E\{J^2(\mathbf{X})\} < \infty\ a.s.,$$

by assumption $(I_3) -$ (iii).

When it comes to the second term on the right-hand side of Eq. (14), it can be further decomposed as follows

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n I_\Gamma(\mathbf{X}_i)\left[\varphi(b_{ni}(\boldsymbol\beta))\nu_{ni}(\boldsymbol\beta) - \varphi(a_{ni}(\boldsymbol\beta))z_i(\boldsymbol\beta)\right]\\
&= \frac{1}{n}\sum_{i=1}^n I_\Gamma(\mathbf{X}_i)\varphi(b_{ni}(\boldsymbol\beta))[\nu_{ni}(\boldsymbol\beta) - z_i(\boldsymbol\beta)]\\
&\quad + \frac{1}{n}\sum_{i=1}^n I_\Gamma(\mathbf{X}_i)\left[\varphi(b_{ni}(\boldsymbol\beta)) - \varphi(a_{ni}(\boldsymbol\beta))\right] z_i(\boldsymbol\beta).
\end{aligned}$$

Considering the first term to the right-hand side of this equation, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n I_\Gamma(\mathbf{X}_i)\varphi(b_{ni}(\boldsymbol\beta))[\nu_{ni}(\boldsymbol\beta) - z_i(\boldsymbol\beta)]\right| \le \frac{L}{n}\sum_{i=1}^n |\widehat{g}^i_{\boldsymbol\beta,h}(\mathbf{X}_i^\tau\boldsymbol\beta) - g_{\boldsymbol\beta}(\mathbf{X}_i^\tau\boldsymbol\beta)| I_\Gamma(\mathbf{X}_i)\\
&\le \max_{1\le i\le n}\ \sup_{\boldsymbol\beta\in\mathscr{B}_n, h\in\mathscr{H}_n} |\widehat{g}^i_{\boldsymbol\beta,h}(\mathbf{X}_i^\tau\boldsymbol\beta) - g_{\boldsymbol\beta}(\mathbf{X}_i^\tau\boldsymbol\beta)| I_\Gamma(\mathbf{X}_i)
\end{aligned}$$

which converges to 0 a.s. by Lemma 3. Now, let's set $F_{i\nu}(s) = P(\nu_{in}(\boldsymbol\beta) \le s)$ and $F_i(s) = P(z_i(\boldsymbol\beta) \le s)$. Then,

$$\begin{aligned}
\varphi(b_{ni}(\boldsymbol\beta)) - \varphi(a_{ni}(\boldsymbol\beta)) &= [\varphi(b_{ni}(\boldsymbol\beta)) - \varphi(F_{i\nu}(\nu_{in}(\boldsymbol\beta)))]\\
&\quad + [\varphi(F_{i\nu}(\nu_{in}(\boldsymbol\beta))) - \varphi(F_i(z_i(\boldsymbol\beta)))]\\
&\quad + [\varphi(F_i(z_i(\boldsymbol\beta))) - \varphi(a_{ni}(\boldsymbol\beta))].
\end{aligned}$$

As in the proof of Lemma 2, since for $i = 1, \ldots, n$ and for all $\boldsymbol\beta \in \mathscr{B}$, $F_{i\nu}(\nu_{in}(\boldsymbol\beta))$ and $F_i(z_i(\boldsymbol\beta))$ are independent uniformly distributed random variables on (0, 1), following

Hájek et al. (1999), we have, $b_{ni}(\boldsymbol{\beta})-F_{i\nu}(\nu_{in}(\boldsymbol{\beta})) \to 0\ a.s.$ and $a_{ni}(\boldsymbol{\beta})-F_i(z_i(\boldsymbol{\beta})) \to 0\ a.s.$, for each i. Applying the generalized continuous mapping theorem (Whitt 2011), we have $\varphi(b_{ni}(\boldsymbol{\beta})) - \varphi(F_{i\nu}(\nu_{in}(\boldsymbol{\beta}))) \to 0\ a.s.$ and $\varphi(F_i(z_i(\boldsymbol{\beta}))) - \varphi(a_{ni}(\boldsymbol{\beta})) \to 0\ a.s.$, for each i and for all $\boldsymbol{\beta} \in \mathscr{B}$. Also, since $\nu_{in}(\boldsymbol{\beta}) - z_i(\boldsymbol{\beta}) \to 0\ a.s.$, by the continuity of the probability measure and the continuity of φ, we have $\varphi(F_{i\nu}(\nu_{in}(\boldsymbol{\beta})))-\varphi(F_i(z_i(\boldsymbol{\beta}))) \to 0\ a.s.$, for each i and for all $\boldsymbol{\beta} \in \mathscr{B}$. Thus,

$$
\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\,[\varphi(b_{ni}(\boldsymbol{\beta})) - \varphi(a_{ni}(\boldsymbol{\beta}))]\,z_i(\boldsymbol{\beta})\Big| \\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n} |\varphi(b_{ni}(\boldsymbol{\beta})) - \varphi(a_{ni}(\boldsymbol{\beta}))|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} |z_i(\boldsymbol{\beta})|^2\right)^{1/2}.
\end{aligned}
$$

From this, we have

$$
n^{-1}\sum_{i=1}^{n} |\varphi(b_{ni}(\boldsymbol{\beta})) - \varphi(a_{ni}(\boldsymbol{\beta}))|^2 \leq \max_{1\leq i\leq n} \sup_{\boldsymbol{\beta}\in\mathscr{B}} |\varphi(b_{ni}(\boldsymbol{\beta})) - \varphi(a_{ni}(\boldsymbol{\beta}))|^2,
$$

which converges almost surely to zero. Furthermore,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} |z_i(\boldsymbol{\beta})|^2 &\leq \frac{1}{n}\sum_{i=1}^{n} (|Y_i| + |J(\mathbf{X}_i)|)^2 \leq \frac{1}{n}\sum_{i=1}^{n} |Y_i|^2 + \frac{1}{n}\sum_{i=1}^{n} |J(\mathbf{X}_i)|^2 \\
&\quad + 2\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} J^2(\mathbf{X}_i)\right)^{1/2} := J_{4n}. \qquad (15)
\end{aligned}
$$

By the strong law of large numbers, the entire expression on the right-hand side of this inequality converges a.s. to $E\{|Y|^2\} + E\{J^2(\mathbf{X})\} + 2\left(E\{|Y|^2\}E\{J^2(\mathbf{X})\}\right)^{1/2} < \infty$, by assumptions (I_2)–(iii) and (I_4). Thus,

$$
\sup_{\boldsymbol{\beta}\in\mathscr{B}} \left|\frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\,[\varphi(b_{ni}(\boldsymbol{\beta})) - \varphi(a_{ni}(\boldsymbol{\beta}))]\,z_i(\boldsymbol{\beta})\right| \to 0\ a.s.
$$

Now, combining all these facts, we have $\sup_{\boldsymbol{\beta}\in\mathscr{B}, h\in\mathscr{H}_n} |\widetilde{D}_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta})| \to 0\ a.s.$ □

Proof of Theorem 2 Note that φ has a bounded first derivative. So, $\varphi \in Lip(1)$. Moreover, by (I_2)–(iii) and (I_4), we have $\mathrm{Var}(z_i(\boldsymbol{\beta})) < \infty$, for all i and $\boldsymbol{\beta} \in \mathscr{B}$. Then

$$
\sum_{i=1}^{n} \frac{\mathrm{Var}(z_i(\boldsymbol{\beta}))}{n^2} \leq \frac{\sigma^2_{\max}(\boldsymbol{\beta})}{n} = O(1/n),
$$

where $\sigma^2_{\max}(\boldsymbol{\beta}) = \max\{\mathrm{Var}(z_1(\boldsymbol{\beta})), \ldots, \mathrm{Var}(z_n(\boldsymbol{\beta}))\}$. Setting $\alpha_n = 1/n$ and $\beta = 1$ in the theorem of (Xiang 1995), we find that for every $\boldsymbol{\beta} \in \mathscr{B}$, $D_n(\boldsymbol{\beta}) - E\{D_n(\boldsymbol{\beta})\} \to 0$ $a.s.$

To complete the proof, we have to show that $\{D_n(\boldsymbol{\beta})\}_{n\geq 1}$ is stochastically equicontinuous. To that end, taking $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathscr{B}$, we have

$$\begin{aligned} &D_n(\boldsymbol{\beta}_1) - D_n(\boldsymbol{\beta}_2) \\ &= \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\left[\varphi\left(\frac{R(z_i(\boldsymbol{\beta}_1))}{n+1}\right) z_i(\boldsymbol{\beta}_1) - \varphi\left(\frac{R(z_i(\boldsymbol{\beta}_2))}{n+1}\right) z_i(\boldsymbol{\beta}_2)\right]. \end{aligned}$$

As in the proof of Theorem 1, set $a_{ni}(\boldsymbol{\beta}) = R(z_i(\boldsymbol{\beta}))/(n+1)$. Then,

$$\begin{aligned} D_n(\boldsymbol{\beta}_1) - D_n(\boldsymbol{\beta}_2) &= \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\left[\varphi\left(a_{ni}(\boldsymbol{\beta}_1)\right) z_i(\boldsymbol{\beta}_1) - \varphi\left(a_{ni}(\boldsymbol{\beta}_2)\right) z_i(\boldsymbol{\beta}_2)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\varphi\left(a_{ni}(\boldsymbol{\beta}_1)\right)\left[z_i(\boldsymbol{\beta}_1) - z_i(\boldsymbol{\beta}_2)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\left[\varphi\left(a_{ni}(\boldsymbol{\beta}_1)\right) - \varphi\{F_i(z_i(\boldsymbol{\beta}_1))\}\right] z_i(\boldsymbol{\beta}_2) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\left[\varphi\{F_i(z_i(\boldsymbol{\beta}_1))\} - \varphi\{F_i(z_i(\boldsymbol{\beta}_2))\}\right] z_i(\boldsymbol{\beta}_2) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\left[\varphi\{F_i(z_i(\boldsymbol{\beta}_2))\} - \varphi\left(a_{ni}(\boldsymbol{\beta}_2)\right)\right] z_i(\boldsymbol{\beta}_2). \end{aligned}$$

Note that $z_i(\boldsymbol{\beta}_1) - z_i(\boldsymbol{\beta}_2) = g_{\boldsymbol{\beta}_1}(\mathbf{X}_i^\tau\boldsymbol{\beta}_1) - g_{\boldsymbol{\beta}_2}(\mathbf{X}_i^\tau\boldsymbol{\beta}_2)$. Since $g_{\boldsymbol{\beta}}(\cdot)$ is differentiable with respect to $\boldsymbol{\beta}$, applying the mean value theorem on the function $g_{\boldsymbol{\beta}}(\mathbf{X}^\tau\boldsymbol{\beta})$, there exists $\boldsymbol{\xi} = \lambda\boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2$ for some $\lambda \in (0, 1)$ such that

$$g_{\boldsymbol{\beta}_1}(\mathbf{X}_i^\tau\boldsymbol{\beta}_1) - g_{\boldsymbol{\beta}_2}(\mathbf{X}_i^\tau\boldsymbol{\beta}_2) = \nabla_{\boldsymbol{\xi}}[g_{\boldsymbol{\xi}}(\mathbf{X}_i^\tau\boldsymbol{\xi})](\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2).$$

Then, by assumption (I_2)–(iii) we have

$$|g_{\boldsymbol{\beta}_1}(\mathbf{X}_i^\tau\boldsymbol{\beta}_1) - g_{\boldsymbol{\beta}_2}(\mathbf{X}_i^\tau\boldsymbol{\beta}_2)| = |\nabla_{\boldsymbol{\xi}}[g_{\boldsymbol{\xi}}(\mathbf{X}_i^\tau\boldsymbol{\xi})](\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)| \leq J(\mathbf{X}_i)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|.$$

Furthermore, set $h_i(\boldsymbol{\beta}) = \varphi\{F_i(z_i(\boldsymbol{\beta}))\} = \varphi\{F_i(Y_i - g_{\boldsymbol{\beta}}(\mathbf{X}_i^\tau\boldsymbol{\beta}))\}$, where F_i is a cumulative distribution function of $z_i(\boldsymbol{\beta})$, and therefore almost surely differentiable. So by the mean value theorem, there exists $\eta = \lambda\boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2$ for $\lambda \in (0, 1)$ such that $h_i(\boldsymbol{\beta}_1) - h_i(\boldsymbol{\beta}_2) = h_i'(\eta)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$, with $h_i'(\eta) = -\nabla_\eta[g_\eta(\mathbf{X}_i^\tau\eta)]f_i(z_i(\eta))\varphi'\{F_i(z_i(\eta))\}$ and $f_i(t) = dF_i(t)/dt$. It is worth pointing out that f_i being a density is almost surely bounded. Thus, by assumption $(I_2) - iii)$ again together with the boundedness of φ', we have $\|h_i'(\eta)\| \leq MJ(\mathbf{X}_i)$ $a.s.$, where M is such that $|f_i(z_i(\eta))\varphi'\{F_i(z_i(\eta))\}| \leq M$ $a.s.$ On the other hand, for

$i = 1, \ldots, n$, $F_i(z_i(\boldsymbol{\beta}))$ being independent uniformly distributed in the interval (0, 1), for all $\boldsymbol{\beta} \in \mathscr{B}$, as in Theorem 1, following Hájek et al. (1999) again, it is obtained that $a_{ni}(\boldsymbol{\beta}) - F_i(z_i(\boldsymbol{\beta})) \to 0$ $a.s.$, for all $\boldsymbol{\beta} \in \mathscr{B}$ and for each i. By continuity of φ, we have $\varphi(a_{ni}(\boldsymbol{\beta})) - \varphi\{F_i(z_i(\boldsymbol{\beta}))\} \to 0$ $a.s.$, for all $\boldsymbol{\beta} \in \mathscr{B}$ and for each i. Thus,

$$\max_{1\le i\le n} |\varphi(a_{ni}(\boldsymbol{\beta})) - \varphi\{F_i(z_i(\boldsymbol{\beta}))\}| \to 0 \; a.s.,$$

for all $\boldsymbol{\beta} \in \mathscr{B}$. Now

$$\left| \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\varphi(a_{ni}(\boldsymbol{\beta}_1))\,[z_i(\boldsymbol{\beta}_1) - z_i(\boldsymbol{\beta}_2)] \right| \le \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \frac{L}{n}\sum_{i=1}^{n} J(\mathbf{X}_i),$$

where L is such that $|\varphi(t)| \le L$, for all $t \in (0, 1)$. Also, with probability 1, we have

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\,[\varphi\{F_i(z_i(\boldsymbol{\beta}_1)) - \varphi\{F_i(z_i(\boldsymbol{\beta}_2))\}]\,z_i(\boldsymbol{\beta}_2) \right| \\
&\le \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \frac{M}{n}\sum_{i=1}^{n} J(\mathbf{X}_i)|z_i(\boldsymbol{\beta}_2)| \\
&\le \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| M \left(\frac{1}{n}\sum_{i=1}^{n} J^2(\mathbf{X}_i) \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} |z_i(\boldsymbol{\beta}_2)|^2 \right)^{1/2} \\
&\le \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| M \left(\frac{1}{n}\sum_{i=1}^{n} J^2(\mathbf{X}_i) \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} [|Y_i| + J(\mathbf{X}_i)]^2 \right)^{1/2}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\,[\varphi(a_{ni}(\boldsymbol{\beta}_1)) - \varphi\{F_i(z_i(\boldsymbol{\beta}_1))\}]\,z_i(\boldsymbol{\beta}_2) \right| \\
&\le \frac{1}{n}\sum_{i=1}^{n} |\varphi(a_{ni}(\boldsymbol{\beta}_1)) - \varphi\{F_i(z_i(\boldsymbol{\beta}_1))\}|\,|z_i(\boldsymbol{\beta}_2)| \\
&\le \left(\max_{1\le i\le n} |\varphi(a_{ni}(\boldsymbol{\beta}_1)) - \varphi\{F_i(z_i(\boldsymbol{\beta}_1))\}|^2 \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} [|Y_i| + J(\mathbf{X}_i)]^2 \right)^{1/2} \to 0 \; a.s.
\end{aligned}$$

as $\max_{1\le i\le n} |\varphi(a_{ni}(\boldsymbol{\beta}_1)) - \varphi\{F_i(z_i(\boldsymbol{\beta}_1))\}|^2 \to 0$ $a.s.$ and

$$\frac{1}{n}\sum_{i=1}^{n} (|Y_i| + |J(\mathbf{X}_i)|)^2 \le J_{4n},$$

where J_{4n}, defined in Eq. (15), converges almost surely to a finite quantity by the strong law of large numbers under assumptions (I_2)–(iii) and (I_4). Similarly,

$$\left| \frac{1}{n} \sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i) \left[\varphi\{F_i(z_i(\boldsymbol{\beta}_2))\} - \varphi\left(a_{ni}(\boldsymbol{\beta}_2)\right)\right] z_i(\boldsymbol{\beta}_2) \right|$$

converges almost surely to zero. Hence, with probability 1, we have

$$|D_n(\boldsymbol{\beta}_1) - D_n(\boldsymbol{\beta}_2)| \le B_n \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|,$$

where

$$B_n =: \frac{L}{n} \sum_{i=1}^{n} J(\mathbf{X}_i) + \times M \left(\frac{1}{n} \sum_{i=1}^{n} J^2(\mathbf{X}_i) \right)^{1/2} J_{4n}^{1/2} + o(1).$$

For n large enough, B_n does not depend on $\boldsymbol{\beta}$. From the fact that all terms in the definition of B_n converge almost surely to a finite quantity, so does B_n. Therefore, $\{D_n(\boldsymbol{\beta})\}_{n \ge 1}$ is stochastically equicontinuous (Rao et al. 2014). □

Proof of Theorem 4 Note that by Jensen inequality,

$$\sup_{\boldsymbol{\beta} \in \mathscr{B}, h \in \mathscr{H}_n} |E(\widetilde{D}_n(\boldsymbol{\beta})) - E(D_n(\boldsymbol{\beta}))| \le E\Big(\sup_{\boldsymbol{\beta} \in \mathscr{B}, h \in \mathscr{H}_n} |\widetilde{D}_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta})| \Big). \quad (16)$$

Thus, together with Theorem 1, applying the dominated convergence theorem to the right-hand side of this inequality, we obtain the result. On the other hand,

$$\begin{aligned} \widetilde{D}_n(\boldsymbol{\beta}) - E(\widetilde{D}_n(\boldsymbol{\beta})) = \widetilde{D}_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta}) + D_n(\boldsymbol{\beta}) - E(D_n(\boldsymbol{\beta})) + E(D_n(\boldsymbol{\beta})) \\ - E(\widetilde{D}_n(\boldsymbol{\beta})). \end{aligned}$$

Thus,

$$\begin{aligned} \sup_{\boldsymbol{\beta} \in \mathscr{B}, h \in \mathscr{H}_n} |\widetilde{D}_n(\boldsymbol{\beta}) - E(\widetilde{D}_n(\boldsymbol{\beta}))| \le & \sup_{\boldsymbol{\beta} \in \mathscr{B}, h \in \mathscr{H}_n} |\widetilde{D}_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta})| \\ & + \sup_{\boldsymbol{\beta} \in \mathscr{B}} |D_n(\boldsymbol{\beta}) - E(D_n(\boldsymbol{\beta}))| \\ & + \sup_{\boldsymbol{\beta} \in \mathscr{B}, h \in \mathscr{H}_n} |E(D_n(\boldsymbol{\beta})) - E(\widetilde{D}_n(\boldsymbol{\beta}))|. \end{aligned} \quad (17)$$

From Theorems 1, 2 and Eq. (16), the terms to the right-hand side of Eq. (17) converge to zero with probability 1. □

Proof of Theorem 5 By assumption (I_6), $\boldsymbol{\beta}_{0,n} = \underset{\boldsymbol{\beta}}{\text{Argmin}}\, E(D_n(\boldsymbol{\beta}))$ which implies that

$$E(D_n(\boldsymbol{\beta}_{0,n})) \leq E(D_n(\boldsymbol{\beta})),$$

for all $\boldsymbol{\beta} \in \mathscr{B}$. On the other hand, by Theorem 4, we have

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathscr{B}, h\in\mathscr{H}_n} |E(\widetilde{D}_n(\boldsymbol{\beta})) - E(D_n(\boldsymbol{\beta}))| = 0.$$

Thus, $\forall\, \varepsilon > 0$, there exists $N > 0$ such that for all $n \geq N$, $|E(\widetilde{D}_n(\boldsymbol{\beta})) - E(D_n(\boldsymbol{\beta}))| < \varepsilon/2$ for all $\boldsymbol{\beta} \in \mathscr{B}$. This implies that

$$-\varepsilon/2 + E(D_n(\boldsymbol{\beta}_{0,n})) < E(\widetilde{D}_n(\boldsymbol{\beta})). \tag{18}$$

Also, for all $n \geq N$, $|E(D_n(\boldsymbol{\beta}_{0,n})) - E(\widetilde{D}_n(\boldsymbol{\beta}_{0,n}))| < \varepsilon/2$. Thus, we have

$$-\varepsilon/2 + E\left(\widetilde{D}_n\left(\boldsymbol{\beta}_{0,n}\right)\right) < E\left(D_n\left(\boldsymbol{\beta}_{0,n}\right)\right). \tag{19}$$

Equations (19) in (18) gives $-\varepsilon + E(\widetilde{D}_n(\boldsymbol{\beta}_{0,n})) < E(\widetilde{D}_n(\boldsymbol{\beta}))$, for all $\boldsymbol{\beta} \in \mathscr{B}$ and for all $n \geq N$. Now ε being arbitrary, letting $\varepsilon \to 0$, we have $E(\widetilde{D}_n(\boldsymbol{\beta}_{0,n})) \leq E(\widetilde{D}_n(\boldsymbol{\beta}))$, for all $\boldsymbol{\beta} \in \mathscr{B}$ which completes the proof. □

Proof of Theorem 6 Note that

$$S_n(\boldsymbol{\beta}) - T_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I_{\Gamma}(\mathbf{X}_i)\nabla_{\boldsymbol{\beta}}[g_{\boldsymbol{\beta}}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})]\left[\varphi\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right) - \varphi\left(F_i(z_i(\boldsymbol{\beta}))\right)\right].$$

So,

$$\begin{aligned}|S_n(\boldsymbol{\beta}) - T_n(\boldsymbol{\beta})| &\leq \frac{1}{n}\sum_{i=1}^{n} J(\mathbf{X}_i)\left|\varphi\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right) - \varphi\left(F_i(z_i(\boldsymbol{\beta}))\right)\right| \quad \text{by } (I_2) - iii)\\ &\leq \left\{\frac{1}{n}\sum_{i=1}^{n} J^2(\mathbf{X}_i)\right\}^{1/2}\left\{\max_{1\leq i\leq n}\sup_{\boldsymbol{\beta}\in\mathscr{B}}\left|\varphi\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right)\right.\right.\\ &\quad\left.\left.-\varphi\left(F_i(z_i(\boldsymbol{\beta}))\right)\right|^2\right\}^{1/2}.\end{aligned}$$

By continuity of φ and the fact that for $i = 1, \ldots, n$, $F_i(z_i(\boldsymbol{\beta}))$ are independent uniformly distributed in $(0, 1)$, once again following (Hájek et al. 1999), we have $\left|\varphi\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right) - \varphi\left(F_i(z_i(\boldsymbol{\beta}))\right)\right| \to 0\; a.s.$, for all i and $\boldsymbol{\beta} \in \mathscr{B}$. Thus,

$$\max_{1\leq i\leq n}\sup_{\boldsymbol{\beta}\in\mathscr{B}}\left|\varphi\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right) - \varphi\left(F_i(z_i(\boldsymbol{\beta}))\right)\right|^2 \to 0\; a.s.$$

On the other hand, $n^{-1}\sum_{i=1}^{n} J^2(\mathbf{X}_i) \rightarrow E[J^2(\mathbf{X})] < \infty$ *a.s.* Hence, $\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathscr{B}} |S_n(\boldsymbol{\beta}) - T_n(\boldsymbol{\beta})| = 0$ *a.s.* □

Proof of Theorem 7 Note that

$$\nabla_{\boldsymbol{\beta}_0} T_n(\boldsymbol{\beta}_0) = -\frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau\boldsymbol{\beta})]\{\nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0)]\}^\tau f(z_i(\boldsymbol{\beta}_0))$$
$$\varphi'(F(z_i(\boldsymbol{\beta}_0))) + \frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\nabla^2_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}}(\mathbf{X}_i^\tau\boldsymbol{\beta}_0)]\varphi(F(z_i(\boldsymbol{\beta}_0))).$$

A direct application of the strong of large numbers shows that $\nabla_{\boldsymbol{\beta}_0} T_n(\boldsymbol{\beta}_0) \rightarrow \mathbf{W}$ *a.s.* If we assume that $\mathbf{X}$ is independent of ε, we have

$$\mathbf{W} = -E\{I_\Gamma(\mathbf{X})\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}^\tau\boldsymbol{\beta})\big)[\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}^\tau\boldsymbol{\beta}_0)\big)]^\tau\}E\{f(\varepsilon)\varphi'(F(\varepsilon))\}$$
$$+ E\{I_\Gamma(\mathbf{X})\nabla^2_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}}(\mathbf{X}^\tau\boldsymbol{\beta}_0)]\}E\{\varphi(F(\varepsilon))\}.$$

But

$$E\left[f(\varepsilon)\varphi'(F(\varepsilon))\right] = \int_{-\infty}^{\infty} f(\varepsilon)\varphi'(F(\varepsilon))dF(\varepsilon) = -\int_{-\infty}^{\infty} f'(\varepsilon)\varphi(F(\varepsilon))d\varepsilon,$$

from integration by parts, since $f(\varepsilon)\varphi(F(\varepsilon)) \rightarrow 0$ as $\varepsilon \rightarrow \pm\infty$. Now, putting $u = F(\varepsilon)$, we have

$$\int_{-\infty}^{\infty} f'(\varepsilon)\varphi(F(\varepsilon))d\varepsilon = -\int_{0}^{1} \varphi(u)\varphi_f(u)du = -\gamma_\varphi^{-1}.$$

On the other have, by assumption (I_1), $E\left[\varphi\big(F(\varepsilon)\big)\right] = \int_0^1 \varphi(t)dt = 0$. Thus,

$$\mathbf{W} = \gamma_\varphi^{-1} E\{I_\Gamma(\mathbf{X})\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}^\tau\boldsymbol{\beta})\big)[\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}^\tau\boldsymbol{\beta}_0)\big)]^\tau\}.$$

On the other hand, to simplify notation, set $\mathbf{A}_i = \nabla_{\boldsymbol{\xi}}[g_{\boldsymbol{\xi}}(\mathbf{X}_i^\tau\boldsymbol{\xi})]$, $\mathbf{B}_i = \nabla^2_{\boldsymbol{\xi}}[g_{\boldsymbol{\xi}}(\mathbf{X}_i^\tau\boldsymbol{\xi})]$ and $\mathbf{C}_i = \nabla^3_{\boldsymbol{\xi}}[g_{\boldsymbol{\xi}}(\mathbf{X}_i^\tau\boldsymbol{\xi})]$

$$\nabla^2_{\boldsymbol{\xi}} T_n(\boldsymbol{\xi}) = -\frac{3}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\mathbf{B}_i\mathbf{A}_i^\tau f_i(z_i(\boldsymbol{\xi}))\varphi'(F_i(z_i(\boldsymbol{\xi})))$$
$$+\frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\mathbf{C}_i\varphi(F_i(z_i(\boldsymbol{\xi})))$$
$$+\frac{1}{n}\sum_{i=1}^{n} I_\Gamma(\mathbf{X}_i)\mathbf{A}_i\mathbf{A}_i^\tau\mathbf{A}_i f_i'(z_i(\boldsymbol{\xi}))\varphi'(F_i(z_i(\boldsymbol{\xi})))$$

$$+\frac{1}{n}\sum_{i=1}^{n} I_{\Gamma}(\mathbf{X}_i)\mathbf{A}_i\mathbf{A}_i^{\tau}\mathbf{A}_i f_i^2(z_i(\boldsymbol{\xi}))\varphi''(F_i(z_i(\boldsymbol{\xi}))).$$

From this, it can be easily shown that with each term to the right-hand side of this equation is bounded by

$$3Ln^{-1}\sum_{i=1}^{n}\exp\{\lambda\|\mathbf{X}_i\|\}[J(\mathbf{X}_i)+J^2(\mathbf{X}_i)+J^3(\mathbf{X}_i)],$$

which converges almost surely to $3L \times E[\exp\{\lambda\|\mathbf{X}\|\}\{J(\mathbf{X})+J^2(\mathbf{X})+J^3(\mathbf{X})\}] < \infty$, by the strong law of large numbers under (I_2)–(iii) and (I_4). Thus, $\nabla^2_{\boldsymbol{\beta}} T_n(\boldsymbol{\xi})$ is almost surely bounded and the result follows from Theorem 2. □

Proof of Theorem 8 We mimic the proof given in Hettmansperger and McKean (1998) for the linear model. Set

$$T_n(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n} I_{\Gamma}(\mathbf{X}_i)\nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^{\tau}\boldsymbol{\beta})]\varphi[F(\varepsilon_i(\boldsymbol{\beta}_0))].$$

It follows by a routine argument that $\sqrt{n}(S_n(\boldsymbol{\beta}_0) - T_n(\boldsymbol{\beta}_0))$ converges to $\mathbf{0}$ in probability. Hence, the proof will be completed by showing that $\sqrt{n}T_n(\boldsymbol{\beta}_0)$ converges to the intended distribution. Using the Cramér–Wold device (Serfling 1980), let

$$U = n^{-1/2}\sum_{i=1}^{n} I_{\Gamma}(\mathbf{X}_i)\mathbf{a}^{\tau}\nabla_{\boldsymbol{\beta}_0}[g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^{\tau}\boldsymbol{\beta}_0)]\varphi[F(\varepsilon_i(\boldsymbol{\beta}_0))],$$

where $\mathbf{a} \in \mathbb{R}^p$. Since F is the distribution of $\varepsilon(\boldsymbol{\beta}_0)$ and $\int_0^1 \varphi(t)\mathrm{d}t = 0$, we have $E(U) = 0$. Also, since $\int_0^1 \varphi^2(t)\mathrm{d}t = 1$,

$$\mathrm{Var}(U) = \frac{1}{n}\sum_{i=1}^{n} E(I_{\Gamma}(\mathbf{X}_i)\mathbf{a}^{\tau}\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^{\tau}\boldsymbol{\beta}_0)\big)[\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^{\tau}\boldsymbol{\beta}_0)\big)]^{\tau}\mathbf{a} \;\rightarrow\; \mathbf{a}^{\tau}\Sigma\mathbf{a} \quad a.s.$$

Note that U is the sum of independent functions of random variables which are not necessarily identically distributed; hence, we need to establish the limit distribution by the Lindeberg–Feller central limit theorem. To this end, set $\sigma_n^2 = \mathrm{Var}(U)$. Defining A_n by

$$A_n = \frac{1}{\sqrt{n}} I_{\Gamma}(\mathbf{X}_i)\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^{\tau}\boldsymbol{\beta}_0)\big)[\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^{\tau}\boldsymbol{\beta}_0)\big)]^{\tau}\varphi[F(\varepsilon_i(\boldsymbol{\beta}_0))],$$

we need to show that

$$\lim_{n\to\infty}\frac{1}{\sigma_n^2}\sum_{i=1}^{n} E[A_n^2 I\{|A_n| > \varepsilon\sigma_n\}] = 0. \tag{20}$$

By assumption (I_3)–(iii), $\|\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta})\| \le J(\mathbf{X}_i)$ and so,

$$\frac{1}{\sqrt{n}}\left|\mathbf{a}^\tau \nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta})\right| \le \frac{1}{\sqrt{n}} J(\mathbf{X}_i)\|\mathbf{a}\|.$$

$J(\cdot)$ being integrable, is almost surely bounded. Thus, there exists a positive constant c such that $J(\mathbf{X}_i) \le c\ a.s.$, and therefore, $n^{-1/2}|\mathbf{a}^\tau \nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta})| \le n^{-1/2}c\|\mathbf{a}\|\ a.s.$ Hence,

$$\frac{1}{\sqrt{n}}|\mathbf{a}^\tau \nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta})| \to 0\ a.s. \quad as \quad n \to \infty.$$

Set $\lambda_n = n^{-1/2}c\|\mathbf{a}\|$. Then, $\lambda_n \to 0$ as $n \to \infty$, and is independent of i. Since σ_n^2 converges to a positive quantity, the ratio $\sigma_n/\lambda_n \to \infty$ as $n \to \infty$. Now conditioning on $\mathbf{X}_i$, it is easy to see that

$$\begin{aligned} E[A_n^2 I\{|A_n| > \varepsilon\sigma_n\}] \le E\Big[&\varphi^2[F(\varepsilon(\boldsymbol{\beta}_0))]I\Big(\big|\varphi[F(\varepsilon(\boldsymbol{\beta}_0))]\big| > \varepsilon\sigma_n/\lambda_n\Big)\Big] \\ &\times \frac{1}{n}\sum_{i=1}^n E\{I_\Gamma(\mathbf{X})\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\big)[\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\big)]^\tau\}. \end{aligned}$$

In this expression, $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n E\{I_\Gamma(\mathbf{X})\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\big)[\nabla_{\boldsymbol{\beta}_0}\big(g_{\boldsymbol{\beta}_0}(\mathbf{X}_i^\tau \boldsymbol{\beta}_0)\big)]^\tau\}$ $< \infty$ by (I_2)–(iii), (I_4) and (I_6). From the boundedness of φ and applying the dominated convergence theorem, we have

$$E\Big[\varphi^2[F(\varepsilon(\boldsymbol{\beta}_0))]I\Big(\big|\varphi[F(\varepsilon(\boldsymbol{\beta}_0))]\big| > \varepsilon\sigma_n/\lambda_n\Big)\Big] \to 0 \quad as \quad n \to \infty.$$

This shows that the limit in (20) goes to zero as $n \to \infty$. □

Proof of Theorem 10 Recall that from Eq. (2), for any $\mathbf{X}_i \in \Gamma$,

$$D_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \varphi\Big(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\Big)z_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \varphi\Big(\frac{i}{n+1}\Big)z_{(i)}(\boldsymbol{\beta}),$$

where $z_{(1)}(\boldsymbol{\beta}) \le z_{(2)}(\boldsymbol{\beta}) \le \cdots \le z_{(n)}(\boldsymbol{\beta})$. Since $R(t)$ is a step function, it has a finite number of jumps. The set of such jumps is finite and therefore has a zero probability. Since $g_{\boldsymbol{\beta}}(\cdot)$ is assumed to be three times continuously differentiable by (I_2)–(iii), $D_n(\boldsymbol{\beta})$ is almost surely differentiable. From this, taking into account Theorem 6 and expanding $D_n(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$ up to order 2, we have with probability 1,

$$D_n(\boldsymbol{\beta}) = D_n(\boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)S_n(\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau \nabla_{\boldsymbol{\beta}} T_n(\boldsymbol{\xi})(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o(1),$$

where $\boldsymbol{\xi} = \lambda\boldsymbol{\beta}_0 + (1-\lambda)\boldsymbol{\beta}$, for $\lambda \in (0, 1)$. Thus,

$$M_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\tau \nabla_{\boldsymbol{\beta}_0} T_n(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

$$-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\tau}\nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\xi})(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+o(1).$$

From this, we have

$$\begin{aligned}|M_n(\boldsymbol{\beta})-D_n(\boldsymbol{\beta})| &\le \|\nabla_{\boldsymbol{\beta}_0}T_n(\boldsymbol{\beta}_0)\|\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2+\frac{1}{2}\|\nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\xi})\|\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2+o(1)\\ &=\left\{\|\nabla_{\boldsymbol{\beta}_0}T_n(\boldsymbol{\beta}_0)\|+\frac{1}{2}\|\nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\xi})\|\right\}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2+o(1)\\ &\le \frac{3L}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2\frac{1}{n}\sum_{i=1}^{n}[J(\mathbf{X}_i)+J^2(\mathbf{X}_i)]+o(1).\end{aligned}$$

as $\|\nabla_{\boldsymbol{\beta}_0}T_n(\boldsymbol{\beta}_0)\|$ and $\|\nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\xi})\|$ are bounded by $Ln^{-1}\sum_{i=1}^{n}[J(\mathbf{X}_i)+J^2(\mathbf{X}_i)]$. On the other hand, $n^{-1}\sum_{i=1}^{n}[J(\mathbf{X}_i)+J^2(\mathbf{X}_i)]\to E\big[J(\mathbf{X})+J^2(\mathbf{X})\big]<\infty$ *a.s.*, by assumption (I_2)–(iii) and (I_4). Now, for any $\boldsymbol{\beta}\in\mathscr{B}_n$, $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le c/\sqrt{n}$. This implies that

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}_n}|M_n(\boldsymbol{\beta})-D_n(\boldsymbol{\beta})|\le\frac{3c^2L}{2n}\frac{1}{n}\sum_{i=1}^{n}[J(\mathbf{X}_i)+J^2(\mathbf{X}_i)]+o(1).$$

By Markov's inequality, we have for any $\varepsilon>0$ and for n large enough,

$$\begin{aligned}P_{\boldsymbol{\beta}_0}\Big[\sup_{\boldsymbol{\beta}\in\mathscr{B}_n}|D_n(\boldsymbol{\beta})-M_n(\boldsymbol{\beta})|>\varepsilon\Big] &\le\frac{1}{\varepsilon}E\left[\sup_{\boldsymbol{\beta}\in\mathscr{B}_n}|M_n(\boldsymbol{\beta})-D_n(\boldsymbol{\beta})|\right]\\ &\le\frac{3c^2L}{2n\varepsilon}E\left\{\frac{1}{n}\sum_{i=1}^{n}[J(\mathbf{X}_i)+J^2(\mathbf{X}_i)]\right\}.\end{aligned}$$

A direct application of the dominated convergence theorem gives

$$\lim_{n\to\infty}E\left\{\frac{1}{n}\sum_{i=1}^{n}[J(\mathbf{X}_i)+J^2(\mathbf{X}_i)]\right\}\to E\Big\{[J(\mathbf{X})+J^2(\mathbf{X})]\Big\}<\infty.$$

Thus, $\lim_{n\to\infty}P_{\boldsymbol{\beta}_0}\Big[\sup_{\boldsymbol{\beta}\in\mathscr{B}_n}|D_n(\boldsymbol{\beta})-M_n(\boldsymbol{\beta})|>\varepsilon\Big]=0$. The proof of Eq. (8) is obtained similarly, while that of Eq. (7) is obtained by combining Eq. (6) and Theorem 1. □

Proof of Theorem 11 Equation (12) gives $\sqrt{n}\big(\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\big)=-\widetilde{\mathbf{W}}_n^{-1}\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)+o_p(1)$ and by (9) we have $\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)=\sqrt{n}S_n(\boldsymbol{\beta}_0)+o_p(1)$. Moreover, $\widetilde{\mathbf{W}}=\mathbf{W}+o_p(1)$ by (13). Since $\mathbf{W}$ is positive definite, we have $\sqrt{n}\big(\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\big)=-\mathbf{W}^{-1}\sqrt{n}S_n(\boldsymbol{\beta}_0)+o_p(1)$. The result follows by Theorem 8. □

References

Abebe, A., McKean, J. W. (2013). Weighted Wilcoxon estimators in nonlinear regression. *Australian and New Zealand Journal of Statistics*, *55*(4), 401–420.

Andrews, D. W. K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, *62*(1), 43–72.

Bindele, H. F., Abebe, A. (2012). Bounded influence nonlinear signed-rank regression. *Canadian Journal of Statistics*, *40*(1), 172–189.

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.

Chang, W. H., McKean, J. W., Naranjo, J. D., Sheather, S. J. (1999). High-breakdown rank regression. *Journal of the American Statistical Association*, *94*(445), 205–219.

Delecroix, M., Härdle, W., Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, *86*(2), 213–226.

Delecroix, M., Hristache, M., Patilea, V. (2006). On semiparametric-estimation in single-index regression. *Journal of Statistical Planning and Inference*, *136*(3), 730–769.

Feng, L., Zou, C., Wang, Z. (2012). Rank-based inference for the single-index model. *Statistics & Probability Letters*, *82*(3), 535–541.

Hájek, J., Šidák, Z., Sen, P. K. (1999). *Theory of rank tests. Probability and mathematical statistics* (2nd ed.). San Diego, CA: Academic Press, Inc.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, *35*(2), 303–316.

Härdle, W., Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, *84*(408), 986–995.

Härdle, W., Tsybakov, A. B. (1993). How sensitive are average derivatives? *Journal of Econometrics*, *58*(1–2), 31–48.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, *21*(1), 157–178.

Hettmansperger, T. P., McKean, J. W. (1998). *Robust nonparametric statistical methods, volume 5 of Kendall's library of statistics*. London: Edward Arnold.

Hettmansperger, T. P., McKean, J. W. (2011). *Robust nonparametric statistical methods, volume 119 of monographs on statistics and applied probability* (2nd ed.). Boca Raton, FL: CRC Press.

Horowitz, J. L., Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, *91*(436), 1632–1640.

Hristache, M., Juditsky, A., Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, *29*(3), 595–623.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, *58*(1–2), 71–120.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics*, *43*, 1449–1458.

Klein, R. W., Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*(2), 387–421.

Kutner, M., Nachtsheim, C., Neter, J., Li, W. (2004). *Applied Linear Statistical Models with Student CD*. New York: McGraw-Hill/Irwin Companies, Incorporated.

Liu, J., Zhang, R., Zhao, W., Lv, Y. (2013). A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis*, *122*, 226–238.

McCullagh, P., Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

Naranjo, J. D., Hettmansperger, T. P. (1994). Bounded influence rank regression. *Journal of the Royal Statistical Society. Series B*, *56*(1), 209–220.

Newey, W. K. (2004). Efficient semiparametric estimation via moment restrictions. *Econometrica*, *72*(6), 1877–1897.

Newey, W. K., McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Powell, J. L., Stock, J. H., Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, *57*(6), 1403–1430.

R Development Core Team. (2009). *R*: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rao, T. S., Das, S., Boshnakov, G. N. (2014). A frequency domain approach for the estimation of parameters of spatio-temporal stationary random processes. *Journal of Time Series Analysis*, *35*(4), 357–377.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*., Wiley series in probability and mathematical statistics. New York: Wiley.

Sherman, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, *22*(1), 439–459.

Whitt, W. (2011). *Stochastic-process limits*: *An introduction to stochastic-process limits and their application to queues*. New York: Springer.

Xia, Y. (2006). Asymptotic distribution for two estimators of the single-index model. *Econometric Theory*, *22*, 1112–1137.

Xia, Y., Tong, H., Li, W. K., Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *64*(3), 363–410.

Xiang, X. (1995). A strong law of large number for L-statistics in the non-i.d. case. *Communications in Statistics. Theory and Methods*, *24*(7), 1813–1819.

Yin, X., Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika*, *92*(2), 371–384.

Yu, Y., Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*(460), 1042–1054.