Score estimation in the monotone single index model

Fadoua Balabdaoui

Université Paris-Dauphine, PSL Research University,
CNRS, CEREMADE, 75016 Paris, France
and
Seminar für Statistik, ETH Zürich,
8092, Zürich, Schweiz
e-mail: fadoua.balabdaoui@gmail.com

Piet Groeneboom

Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
e-mail: P.Groeneboom@tudelft.nl

Kim Hendrickx

Hasselt University, I-BioStat, Agoralaan, B3590 Diepenbeek, Belgium.
e-mail: kim.hendrickx@uhasselt.be

Abstract: We consider estimation of the regression parameter in the single index model where the link function ψ is monotone. For this model it has been proposed to estimate the link function non-parametrically by the monotone least square estimate $\hat{\psi}_{n\boldsymbol{\alpha}}$ for a fixed regression parameter $\boldsymbol{\alpha}$ and to estimate the regression parameter by minimizing the sum of squared deviations $\sum_i \{Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i)\}^2$ over $\boldsymbol{\alpha}$, where Y_i are the observations and $\boldsymbol{X}_i$ the corresponding covariates. Although it is natural to propose this least squares procedure, it is still unknown whether it will produce $\sqrt{n}$-consistent estimates of $\boldsymbol{\alpha}$. We show that the latter property will hold if we solve a score equation corresponding to this minimization problem. We also compare our method with other methods such as Han's maximum rank correlation estimate, which has been proved to be $\sqrt{n}$-consistent.

AMS 2000 subject classifications: 62G05, 62G20, 62H12.
Keywords and phrases: monotone link functions, nonparametric least squares estimates, semi-parametric model, single index regression model.

1. Introduction

Single index models are flexible models used in regression analysis of the type $\mathbb{E}(Y|\boldsymbol{X}) = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X})$, where ψ_0 is an unknown link function and $\boldsymbol{\alpha}_0$ is an unknown regression parameter. By lowering the dimensionality of the classical linear regression problem, determined by the number of covariates, to a univariate $\boldsymbol{\alpha}_0^T \boldsymbol{X}$ index, single index models do not suffer from the "curse of dimensionality". They also provide an advantage over the generalized linear regression models by overcoming the risk of misspecifying the link function ψ_0. To ensure identifiability of the single index model, one typically assumes that the Euclidean norm $\|\boldsymbol{\alpha}_0\|$ equals one with the first non-zero element of $\boldsymbol{\alpha}_0$ being positive.

Several estimation approaches have been considered in the literature of single index models. These methods can be classified into two groups: M-estimators and direct estimators. In the first approach, one considers a non-parametric regression estimate for the infinite dimensional link function ψ_0 and then estimates $\boldsymbol{\alpha}_0$ by minimizing a certain criterion function, where ψ_0 is replaced by its estimate. Examples of this type are the semi-parametric least squares estimators of [13] and [10] and the pseudo-maximum likelihood estimator of [5], all using kernel regression estimates for the unknown link functions. An example of an M-estimator that does not depend on an estimate of the link function ψ_0 in Han's maximum rank correlation estimator [9].

Direct estimators, such as the average derivative estimator of [11] or the slicing regression method proposed in [6], avoid solving an optimization problem and are often computationally more attractive than M-estimators.

In this paper we focus on estimating the regression parameter $\boldsymbol{\alpha}_0$ under the constraint that ψ_0 is monotone. Shape constrained inference arises naturally in a variety of fields. For example in economics where a concavity restriction is assumed in utility theory to indicate the exhibition of risk conversion in economic behavior.

Convex optimization problems also appear frequently and often allow for straightforward computation and optimization.

The single index model with convex link has been studied in [14] where the authors consider estimation of a penalized least squares estimator using smoothing splines and prove $\sqrt{n}$-consistency and asymptotic normality of their proposed estimator. [1] considered a global least squares estimator for the pair $(\boldsymbol{\alpha}_0, \psi_0)$ under monotonicity of the function ψ_0. They derived an $n^{-1/3}$ convergence rate, but the asymptotic limiting distribution for their estimator of $\boldsymbol{\alpha}_0$ has not been derived. A conjecture is made in [16] that this rate is too slow. In this paper, we will give simulation results on the asymptotic variance of the least squares estimator and investigate its rate of convergence numerically.

Recently, [7] developed several score estimators for the current status linear regression model $Y = \boldsymbol{\beta}_0^T \boldsymbol{Z} + \varepsilon$, where the distribution function F_0 of ε is left unspecified. Instead of observing the response Y a censoring variable T and censoring indicator $\Delta = 1_{Y \leq T}$ are observed. This model is a special case of the monotone single index model and can be formulated as $\mathbb{E}(\Delta | T, \boldsymbol{Z}) = F_0(T - \boldsymbol{\beta}_0^T \boldsymbol{Z}) = F_0(\boldsymbol{\alpha}_0^T \boldsymbol{X})$ where $\boldsymbol{\alpha}_0 = (1, -\boldsymbol{\beta}_0)^T$ and $\boldsymbol{X} = (T, \boldsymbol{Z})^T$. Their estimators are obtained by the root of a score function involving the maximum likelihood estimator (MLE) of the distribution function for fixed $\boldsymbol{\beta}$. The authors prove $\sqrt{n}$-consistency and asymptotic normality of their estimators and show that under certain smoothness assumptions the limiting variance of a score estimator is arbitrarily close to the efficient variance. Their result is remarkable since it is the first time in the current status regression model that a $\sqrt{n}$-consistent estimate for $\boldsymbol{\beta}_0$ is proposed based on the MLE for F_0 which only converges at $n^{-1/3}$-rate to the true distribution function F_0.

We consider extending the estimators of [7] to the more general single index regression problem and propose two different score equations involving the least squares estimator (LSE) $\hat{\psi}_{n\boldsymbol{\alpha}}$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i - \psi(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \Big\}^2, \tag{1.1}$$

over all monotone increasing functions ψ for fixed $\boldsymbol{\alpha}$. We establish an $n^{-1/3} \log n$-rate for the estimator $\hat{\psi}_{n\boldsymbol{\alpha}}$ and propose a single index score estimator of $\boldsymbol{\alpha}_0$ that converges at the parametric rate $n^{-1/2}$ to the true regression parameter $\boldsymbol{\alpha}_0$.

2. The single-index model with monotone link

Consider the following regression model

$$Y = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) + \varepsilon, \tag{2.1}$$

where Y is a one-dimensional random variable, $\boldsymbol{X} = (X_1, \ldots, X_d)'$ is a d-dimensional random vector with distribution G and ε is a one-dimensional random variable such that $\mathbb{E}[\varepsilon | \boldsymbol{X}] = 0$ G-almost surely. The function ψ_0 is a monotone link function in $\mathcal{M}$, where $\mathcal{M}$ is the set of monotone increasing functions defined on $\mathbb{R}$ and $\boldsymbol{\alpha}_0$ is a vector of regression parameters belonging to the $d-1$ dimensional sphere $\mathcal{S}_{d-1} := \{\boldsymbol{\alpha} \in \mathbb{R}^d : \|\boldsymbol{\alpha}\| = 1\}$, where $\| \cdot \|$ denotes the Euclidean norm in $\mathbb{R}^d$.

3. The least squares estimator (LSE) $\hat{\psi}_{n\boldsymbol{\alpha}}$

Let $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ denote n random variables which are i.i.d. like $(\boldsymbol{X}, Y)$ in (2.1), i.e. $\mathbb{E}(Y | \boldsymbol{X}) = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X})$ G-almost surely and consider the sum of squared errors

$$S_n(\psi, \boldsymbol{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i - \psi(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \Big\}^2,$$

which can be computed for any pair $(\psi, \boldsymbol{\alpha}) \in \mathcal{M} \times \mathcal{S}_{d-1}$. For a fixed $\boldsymbol{\alpha}$, order the values $\boldsymbol{\alpha}^T \boldsymbol{X}_1, \ldots, \boldsymbol{\alpha}^T \boldsymbol{X}_n$ in increasing order and arrange $Y_1, \ldots, Y_n$ accordingly. Then, well-known results from monotone regression theory imply that the functional $\psi \mapsto S_n(\psi, \boldsymbol{\alpha})$ is minimized by the left derivative of the greatest convex minorant of the cumulative sum diagram

$$\Big\{ (0,0), \Big(i, \sum_{j=1}^{i} Y_i^{\alpha} \Big), i = 1, \ldots, n \Big\},$$

where Y_i^α corresponds to the ith order statistic of the $\boldsymbol{\alpha}^T\boldsymbol{X}_i$. See for example Theorem 1.1 in [3].

By strict convexity of $\psi \mapsto S_n(\psi, \boldsymbol{\alpha})$, the minimizer is unique at the distinct projections. We denote by $\hat{\psi}_{n\boldsymbol{\alpha}}$ the monotone function which takes the values of this minimizer at the distinct projections and is a stepwise and right-continuous function outside the set of those projections.

We first list below the assumptions needed to prove the asymptotic results stated in the remainder of the paper.

Assumptions A1-A6

A1. The space $\mathcal{X}$ is convex, with a nonempty interior. There exists also $R > 0$ such that $\mathcal{X} \subset \mathcal{B}(0, R)$.

A2. There exists $K_0 > 0$ such that the true link function ψ_0 satisfies $|\psi_0(u)| \leq K_0$ for all u in $\{\boldsymbol{\alpha}^T\boldsymbol{x}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\alpha} \in \mathcal{S}_{d-1}\}$.

A3. There exists $\delta_0 > 0$ such that the function $u \mapsto \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]$ is monotone increasing on $\mathcal{I}_{\boldsymbol{\alpha}} := \{\boldsymbol{\alpha}^T\boldsymbol{x}, \boldsymbol{x} \in \mathcal{X}\}$ for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0) := \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| \leq \delta_0\}$.

A4. Let a_0 and b_0 denote the infimum and supremum of the interval $\mathcal{I}_{\boldsymbol{\alpha}_0} = \big\{\boldsymbol{\alpha}_0^T\boldsymbol{x},\ \boldsymbol{x} \in \mathcal{X}\big\}$. Then, the true link function ψ_0 is continuously differentiable on $(a_0 - \delta_0 R, b_0 + \delta_0 R)$, where R is the same radius of assumption A1 above, and there exists $C > 0$ such that $\psi_0' \geq C$ on $(a_0 - \delta_0 R, b_0 + \delta_0 R)$.

A5. The distribution of $\boldsymbol{X}$ admits a density g, which is differentiable on $\mathcal{X}$. Also, there exist positive constants $\underline{c}_0$, $\bar{c}_0$, $\underline{c}_1$ and $\bar{c}_1$ such that $\underline{c}_0 \leq g \leq \bar{c}_0$ and $\underline{c}_1 \leq \partial g/\partial x_i \leq \bar{c}_1$ on $\mathcal{X}$ for all $i = 1, \ldots, d$.

A6. There exist $a_0, M_0 > 0$ such that $\mathbb{E}\Big[|Y|^m\ |\boldsymbol{X} = \boldsymbol{x}\Big] \leq m! M_0^{m-2} a_0$ for all integers $m \geq 2$ and $\boldsymbol{x} \in \mathcal{X}$ G-almost surely.

Assumption A1 ensures that the support of the linear predictor $\boldsymbol{\alpha}^T\boldsymbol{X}$ is an interval for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. Assumption A3 is made to enable deriving the explicit limit of the LSE $\hat{\psi}_{n\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. We will show, in Lemma 3.6 in the supplemental material, the plausibility of this Assumption A3 by proving that for $\boldsymbol{\alpha}$ in a neighborhood of $\boldsymbol{\alpha}_0$ the derivative of the function $u \mapsto \psi_{\boldsymbol{\alpha}} := \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]$ is indeed strictly positive if the derivative of the true link function stays away from zero. Assumption A6 is needed to show that $\max_{1 \leq i \leq n} |Y_i| = O_p(\log n)$. As noted in [1], such an assumption is satisfied if the conditional distribution of $Y|\boldsymbol{X} = \boldsymbol{x}$ belongs to an exponential family.

In Figure 1 we compare the true link function ψ_0 with the function $\psi_{\boldsymbol{\alpha}}$ for the model $E(Y|\boldsymbol{X}) = \psi_0(\alpha_{01}X_1 + \alpha_{02}X_1)$, where $X_1, X_2 \overset{i.i.d}{\sim} U[0,1]$, $\psi_0(\boldsymbol{x}) = x^3$ and $\alpha_{01} = \alpha_{02} = 1/\sqrt{2}$ for $\alpha_1 = 1/2, \alpha_2 = \sqrt{3}/2$. Figure 1 illustrates the monotonicity of the function $\psi_{\boldsymbol{\alpha}}$ introduced in Assumption $A3$.

We have the following results.

Proposition 3.1. *Suppose that Assumptions A1-A2 hold and let the function $\psi_{\boldsymbol{\alpha}}$ be defined by*

$$\psi_{\boldsymbol{\alpha}}(u) = \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]. \tag{3.1}$$

Then, the functional $L_{\boldsymbol{\alpha}}$ given by,

$$\psi \mapsto L_{\boldsymbol{\alpha}}(\psi) := \int_{\mathcal{X}} \Big(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)^2 g(\boldsymbol{x}) d\boldsymbol{x}, \tag{3.2}$$

admits a minimizer $\hat{\psi}^{\boldsymbol{\alpha}}$, over the set of monotone increasing functions defined on $\mathbb{R}$, denoted by $\mathcal{M}$, such that $\hat{\psi}^{\boldsymbol{\alpha}}$ is uniquely given by the function in (3.1) on $\mathcal{I}_{\boldsymbol{\alpha}} = \Big\{\boldsymbol{\alpha}^T\boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\Big\}$.

Proposition 3.2. *Under Assumptions A1-A6, we have,*

$$\sup_{\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)} \int \Big\{\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big\}^2\ dG(\boldsymbol{x}) = O_p\left((\log n)^2 n^{-2/3}\right).$$

The proofs of Proposition 3.1 and Proposition 3.2 are given in the Appendix.

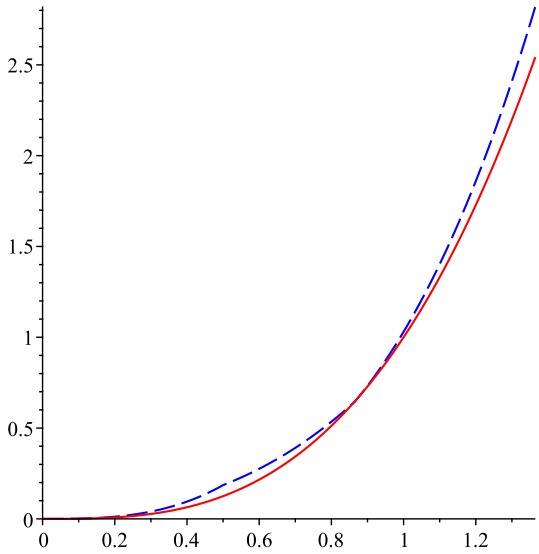

Fig 1: The real ψ_0 (red, solid) and the function $\psi_{\boldsymbol{\alpha}}$ (blue, dashed) for $\psi_0(\boldsymbol{x}) = x^3$, $\alpha_{01} = \alpha_{02} = 1/\sqrt{2}$ and $\alpha_1 = 1/2, \alpha_2 = \sqrt{3}/2$, with $X_1, X_2 \overset{i.i.d}{\sim} U[0,1]$.

4. The score estimator on the unit sphere

Consider the problem of minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}^2 \tag{4.1}$$

over all $\boldsymbol{\alpha} \in \mathcal{S}_{d-1}$, where $\hat{\psi}_{n\boldsymbol{\alpha}}$ is the LSE of $\psi_{\boldsymbol{\alpha}}$.

Let $\mathbb{S}$ be a local parametrization mapping $\mathbb{R}^{d-1}$ to the sphere $\mathcal{S}_{d-1}$, i.e., for each $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ on the sphere $\mathcal{S}_{d-1}$, there exists a unique vector $\boldsymbol{\beta} \in \mathbb{R}^{d-1}$ such that

$$\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta}).$$

The minimization problem given in (4.1) is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}_i\right)\right\}^2 \tag{4.2}$$

over all $\boldsymbol{\beta}$ where $\hat{\psi}_{n\boldsymbol{\alpha}}$ is the LSE of $\psi_{\boldsymbol{\alpha}}$ with $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$. Analogously to the treatment of the score approach in the current status regression model proposed by [7], we consider the derivative of (4.2) w.r.t. $\boldsymbol{\beta}$, where we ignore the non-differentiability of the LSE $\hat{\psi}_{n\boldsymbol{\alpha}}$. This leads to the set of equations,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T\boldsymbol{X}_i\left\{Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}_i\right)\right\} = \mathbf{0} \tag{4.3}$$

where $\boldsymbol{J}_{\mathbb{S}}$ is the Jacobian of the map $\mathbb{S}$ and where $\mathbf{0} \in \mathbb{R}^{d-1}$ is the vector of zeros. Just as in the analogous case of the "simple score equation" in [7], we cannot hope to solve equation (4.3) exactly. Instead, we define the solution in terms of a "zero-crossing" of the above equation. The following definition is taken from [7].

Definition 4.1 (zero-crossing)**.** We say that $\boldsymbol{\beta}_*$ is a crossing of zero of a real-valued function $\zeta : \mathcal{A} \mapsto \mathbb{R} : \boldsymbol{\beta} \mapsto \zeta(\boldsymbol{\beta})$ if each open neighborhood of $\boldsymbol{\beta}_*$ contains points $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{A}$ such that $\bar{\zeta}(\boldsymbol{\beta}_1)\bar{\zeta}(\boldsymbol{\beta}_2) \leq 0$, where $\bar{\zeta}$ is the closure of the image of the function (so contains its limit points). We say that an m-dimensional function $\zeta : \mathcal{A} \mapsto \mathbb{R}^m : \boldsymbol{\beta} \mapsto \zeta(\boldsymbol{\beta}) = (\zeta_1(\boldsymbol{\beta}), \ldots \zeta_m(\boldsymbol{\beta}))'$ has a crossing of zero at a point $\boldsymbol{\beta}_*$, if $\boldsymbol{\beta}_*$ is a crossing of zero of each component $\zeta_j : \mathcal{A} \mapsto \mathbb{R}, j = 1 \ldots, m$.

Our index score estimator $\hat{\boldsymbol{\alpha}}_n$ is defined by,

$$\hat{\boldsymbol{\alpha}}_n := \mathbb{S}(\hat{\boldsymbol{\beta}}_n), \tag{4.4}$$

where $\hat{\boldsymbol{\beta}}_n$ is a zero crossing of the function

$$\phi_n(\boldsymbol{\beta}) := \int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T \boldsymbol{x} \left\{ y - \hat{\psi}_{n\boldsymbol{\alpha}} \left(\mathbb{S}(\boldsymbol{\beta})^T \boldsymbol{x} \right) \right\} d\mathbb{P}_n(\boldsymbol{x}, y), \tag{4.5}$$

and $\mathbb{P}_n$ denotes the empirical probability measure of $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$. The probability measure of $(\boldsymbol{X}, Y)$ will be denoted by P_0 in the remainder of the paper.

In addition to Assumptions A1-A6 above, the following assumptions will also be made.

Assumptions A7-A9

A7. For all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ such that $\mathbb{S}(\boldsymbol{\beta}) \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$, the random variable

$$\mathrm{Cov}\Big[(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})^T \boldsymbol{X}, \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)\big) |\, \mathbb{S}(\boldsymbol{\beta})^T \boldsymbol{X} \Big]$$

is not equal to 0 almost surely.

A8. The functions $J_{\mathbb{S}}^{ij}(\boldsymbol{\beta})$, where $J_{\mathbb{S}}^{ij}(\boldsymbol{\beta})$ denotes the $i \times j$ entry of $J_{\mathbb{S}}(\boldsymbol{\beta})$ for $i = 1, \ldots, d$ and $j = 1, \ldots, d-1$ are $d-1$ times continuously differentiable on $\mathcal{C} := \{\boldsymbol{\beta} \in \mathbb{R}^{d-1} : \mathbb{S}(\boldsymbol{\beta}) \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)\}$ and there exists $M > 0$ satisfying

$$\max_{k_\cdot \leq d-1} \sup_{\boldsymbol{\beta} \in \mathcal{C}} |D^k J_{\mathbb{S}}^{ij}(\boldsymbol{\beta})| \leq M \tag{4.6}$$

where $k = (k_1, \ldots, k_d)$ with k_j an integer $\in \{0, \ldots, d-1\}$, $k_\cdot = \sum_{i=1}^{d-1} k_i$ and

$$D^k s(\boldsymbol{\beta}) \equiv \frac{\partial^{k_\cdot} s(\boldsymbol{\beta})}{\partial \beta_{k_1} \ldots \partial \beta_{k_d}}.$$

We also assume that $\mathcal{C}$ is a convex and bounded set in $\mathbb{R}^{d-1}$ with a nonempty interior.

A9. $(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \, \mathbb{E}\Big[\psi_0'(\boldsymbol{\alpha}_0^T \boldsymbol{X}) \, \mathrm{Cov}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X}) \Big] \, (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))$ is non-singular.

Theorem 4.1. *Let Assumptions A1-A9 be satisfied. Let also $\hat{\boldsymbol{\alpha}}_n$ be defined by (4.4). Then:*

(i) [Existence of a root] *For all large n, a crossing of zero $\hat{\boldsymbol{\beta}}_n$ of $\phi_n(\boldsymbol{\beta})$ exists with probability tending to one.*

(ii) [Consistency]

$$\hat{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0, \qquad n \to \infty.$$

(iii) [Asymptotic normality] *Define the matrices,*

$$\boldsymbol{A} \stackrel{def}{=} \mathbb{E}\Big[\psi_0'(\boldsymbol{\alpha}_0^T \boldsymbol{X}) \, \mathrm{Cov}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X}) \Big], \tag{4.7}$$

and

$$\boldsymbol{\Sigma} \stackrel{def}{=} \mathbb{E}\left[\left\{ Y - \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) \right\}^2 \left\{ \boldsymbol{X} - \mathbb{E}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X}) \right\} \left\{ \boldsymbol{X} - \mathbb{E}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X}) \right\}^T \right], \tag{4.8}$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \to_d N_d\left(\boldsymbol{0}, \boldsymbol{A}^- \boldsymbol{\Sigma} \boldsymbol{A}^- \right),$$

where $\boldsymbol{A}^-$ is the Moore-Penrose inverse of $\boldsymbol{A}$.

Remark 4.1. Note that $\boldsymbol{\alpha}_0^T \boldsymbol{A} = \boldsymbol{0}$ and that the normal distribution $N_d(\boldsymbol{0}, \boldsymbol{A}^- \boldsymbol{\Sigma} \boldsymbol{A}^-)$ is concentrated on the $(d-1)$-dimensional subspace, orthogonal to $\boldsymbol{\alpha}_0$ and is therefore degenerate, as is also clear from its covariance matrix $\boldsymbol{A}^- \boldsymbol{\Sigma} \boldsymbol{A}^-$, which is a matrix of rank $d-1$.

4.1. The asymptotic relation

To obtain the asymptotic normality result of the score estimator $\hat{\boldsymbol{\alpha}}_n$ given in Theorem 4.1, we shall prove the following asymptotic relationship for $\hat{\boldsymbol{\beta}}_n$:

$$\boldsymbol{B}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = \int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\} d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, y) + o_p\left(n^{-1/2} + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right). \tag{4.9}$$

where

$$\boldsymbol{B} = (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\, \mathbb{E}\Big[\psi_0'(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\,\mathrm{Cov}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\Big]\,(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)) = (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\,\boldsymbol{A}\,\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0), \tag{4.10}$$

in $\mathbb{R}^{(d-1)\times(d-1)}$. We assume in Assumption A9 that $\boldsymbol{B}$ is invertible so that

$$\begin{aligned}
&\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \qquad (4.11)\\
&\quad= \sqrt{n}\boldsymbol{B}^{-1}\int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\} d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, y)\\
&\qquad + o_p\left(1 + \sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right)\\
&\quad\to_d N(\mathbf{0}, \boldsymbol{\Pi}). \qquad (4.12)
\end{aligned}$$

where

$$\boldsymbol{\Pi} = \boldsymbol{B}^{-1}\,(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\,\boldsymbol{\Sigma}\,\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\,\boldsymbol{B}^{-1} \in \mathbb{R}^{(d-1)\times(d-1)}. \tag{4.13}$$

The limit distribution of the single index score estimator $\hat{\boldsymbol{\alpha}}_n$ defined in (4.4) now follows by an application of the delta-method and we conclude that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) = \sqrt{n}\left(\mathbb{S}\big(\hat{\boldsymbol{\beta}}_n\big) - \mathbb{S}\left(\boldsymbol{\beta}_0\right)\right) &= J_{\mathbb{S}}(\boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right)\\
&\to_d N_d\left(\mathbf{0}, J_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{\Pi}\,(J_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\right) = N_d\left(\mathbf{0}, \boldsymbol{A}^-\boldsymbol{\Sigma}\boldsymbol{A}^-\right),
\end{aligned}$$

where the last equality follows from the following lemma.

Lemma 4.1. *Let the matrix $\boldsymbol{A}$ be defined by (4.7) and let $\boldsymbol{A}^-$ be the Moore-Penrose inverse of $\boldsymbol{A}$. Then*

$$\boldsymbol{A}^- = \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\,\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T = \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{B}^{-1}\,(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T.$$

The proof of Lemma 4.1 is given in Supplement C of [2]. An example of the mapping $\mathbb{S}$ and corresponding matrix $\boldsymbol{J}_{\mathbb{S}}$ is given in Section 6.

Remark 4.2. *For each map $\mathbb{S}$ and each parameter vector $\boldsymbol{\beta}$, we have*

$$(\mathbb{S}(\boldsymbol{\beta}))^T\,\mathbb{S}(\boldsymbol{\beta}) = 1.$$

Taking derivatives w.r.t. $\boldsymbol{\beta}$, we get

$$(\mathbb{S}(\boldsymbol{\beta}))^T\,J_{\mathbb{S}}(\boldsymbol{\beta}) = \mathbf{0}^T,$$

so that the columns of $J_{\mathbb{S}}(\boldsymbol{\beta})$ belong to the space

$$\{\boldsymbol{\alpha}\}^\perp \equiv \{\mathbb{S}(\boldsymbol{\beta})\}^\perp \equiv \left\{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{\alpha}^T z = 0\right\} \equiv \left\{\boldsymbol{z} \in \mathbb{R}^d : (\mathbb{S}(\boldsymbol{\beta}))^T\, z = 0\right\}.$$

By Lemma 4.1 it is now easy to see that also $\boldsymbol{\alpha}_0^T\boldsymbol{A}^- = \mathbf{0}$. It is shown in Lemma 1 of [14] that it is possible to construct a set of "local parametrization matrices" $H_{\boldsymbol{\alpha}}$ for each $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ with $\|\boldsymbol{\alpha}\| = 1$ satisfying

$$\boldsymbol{\alpha}^T H_{\boldsymbol{\alpha}} = \mathbf{0}^T \quad \text{and} \quad (H_{\boldsymbol{\alpha}})^T\,H_{\boldsymbol{\alpha}} = \boldsymbol{I}_{d-1}$$

Their matrix $(H_{\boldsymbol{\alpha}})^T$ corresponds to the Moore-Penrose pseudo-inverse of the matrix $H_{\boldsymbol{\alpha}}$ and is the analogue of our matrix $(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T$ in the proof of asymptotic normality of their estimator. We however show that the orthormality assumption is not needed in the proofs.

5. The efficient score estimator

In this section we extend the score approach of Section 4 by incorporating an estimate of the derivative of the link function ψ_0 to obtain an efficient estimator of $\boldsymbol{\alpha}_0$. Let $\hat{\psi}_{n\boldsymbol{\alpha}}$ denote again the LSE of $\psi_{\boldsymbol{\alpha}}$ defined in Section 3 and define the estimate $\tilde{\psi}'_{nh,\boldsymbol{\alpha}}$ by

$$\tilde{\psi}'_{nh,\boldsymbol{\alpha}}(u) = \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\hat{\psi}_{n\boldsymbol{\alpha}}(x),$$

where h is a chosen bandwidth. Here $d\hat{\psi}_{n\boldsymbol{\alpha}}$ represents the jumps of the discrete function $\hat{\psi}_{n\boldsymbol{\alpha}}$ and K is one of the usual symmetric twice differentiable kernels with compact support $[-1,1]$, used in density estimation. The estimator $\tilde{\boldsymbol{\alpha}}_n$ is given by

$$\tilde{\boldsymbol{\alpha}}_n := \mathbb{S}(\tilde{\boldsymbol{\beta}}_n), \tag{5.1}$$

where $\tilde{\boldsymbol{\beta}}_n$ is a zero crossing of ξ_{nh} (see Definition 4.1) defined by

$$\xi_{nh}(\boldsymbol{\beta}) := \int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T \boldsymbol{x}\, \tilde{\psi}'_{nh,\boldsymbol{\alpha}}\big(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\big)\left\{y - \hat{\psi}_{n\boldsymbol{\alpha}}\big(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\big)\right\} d\mathbb{P}_n(\boldsymbol{x},y). \tag{5.2}$$

The function ξ_{nh} is inspired by representing the sum of squares

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}^2,$$

in a local coordinate system with $d-1$ unknown parameters $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{d-1})^T$ followed by differentiation of the re-parametrized sum of squares w.r.t. $\boldsymbol{\beta}$ where we also consider differentiation of the function $\psi_{\boldsymbol{\alpha}}$.

We make the following additional assumptions for establishing the weak convergence of $\tilde{\boldsymbol{\beta}}_n$

Assumptions A10-A11

A10. The function $\psi_{\boldsymbol{\alpha}}$ is two times continuously differentiable on $\mathcal{I}_{\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha}$.

A11. $(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\,\mathbb{E}\Big[\psi'_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\,\mathrm{Cov}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\Big]\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ is non-singular.

Theorem 5.1. *Let Assumptions A1-A8, A10-A11 be satisfied. Let $\tilde{\boldsymbol{\alpha}}_n$ be defined by (5.1) and suppose $h \asymp n^{-1/7}$. Then:*

(i) [Existence of a root] *For all large n, a crossing of zero $\tilde{\boldsymbol{\beta}}_n$ of $\xi_{nh}(\boldsymbol{\beta})$ exists with probability tending to one.*

(ii) [Consistency]

$$\tilde{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0, \qquad n\to\infty.$$

(iii) [Asymptotic normality] *Define the matrices,*

$$\tilde{\boldsymbol{A}} \overset{def}{=} \mathbb{E}\Big[\psi'_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\,\mathrm{Cov}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\Big], \tag{5.3}$$

and

$$\tilde{\boldsymbol{\Sigma}} \overset{def}{=} \mathbb{E}\left[\left\{Y-\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^2\psi'_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T\right], \tag{5.4}$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \to_d N_d\left(\boldsymbol{0}, \tilde{\boldsymbol{A}}^-\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{A}}^-\right),$$

where $\tilde{\boldsymbol{A}}^-$ is the Moore-Penrose inverse of $\tilde{\boldsymbol{A}}$.

Remark 5.1. The asymptotic variance of the estimator $\tilde{\boldsymbol{\alpha}}_n$ is similar to that obtained for the "efficient" estimates proposed in [20] and in [14]. The efficient score function for the semi-parametric single index model is

$$\tilde{\ell}_{\boldsymbol{\alpha}_0,\psi_0}(\boldsymbol{x},y) = \frac{y-\psi\left(\boldsymbol{\alpha}_0^T\boldsymbol{x}\right)}{\sigma^2(\boldsymbol{x})}\psi'\left(\boldsymbol{\alpha}_0^T\boldsymbol{x}\right)\left\{\boldsymbol{x} - \frac{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X}=\boldsymbol{\alpha}_0^T\boldsymbol{x}\right\}}{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})|\boldsymbol{\alpha}_0^T\boldsymbol{X}=\boldsymbol{\alpha}_0^T\boldsymbol{x}\right\}}\right\}.$$

More details on the efficiency calculations can be found in e.g. [18], chapter 25 for a general description of the efficient score functions and in [5] or [14] for the efficient score in the single index model.

In a homoscedastic model with $\text{var}(Y|\boldsymbol{X}=\boldsymbol{x}) = \sigma^2$, where σ^2 is independent of covariates $\boldsymbol{x}$, the asymptotic variance equals $\sigma^2\tilde{\boldsymbol{A}}^-$ which is the same as the inverse of $\mathbb{E}(\tilde{\ell}_{\boldsymbol{\alpha}_0,\psi_0}(\boldsymbol{X},Y)\tilde{\ell}_{\boldsymbol{\alpha}_0,\psi_0}(\boldsymbol{X},Y)^T)$. This indeed shows that our estimate defined in (5.1) is efficient in the homoscedastic model. As also explained in Remark 2 of [14], our estimator has also a high relative efficiency with respect to the optimal semi parametric efficiency bound if the constant variance assumption provides a good approximation to the truth.

The asymptotic variance is obtained similarly to the derivations of the asymptotic limiting distribution for the simple score estimator as shown in Section 4.1. First the asymptotic variance is expressed in terms of the parametrization $\mathbb{S}$ as in (4.13) and next, similar to Lemma 4.1, equivalence to the expression $\tilde{\boldsymbol{A}}^-\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{A}}^-$ given in Theorem 5.1 is proved.

6. Computation

In this section we describe how the score estimator $\hat{\boldsymbol{\alpha}}_n$ defined in (4.4) can be obtained using a local coordinate system representing the unit sphere in combination with a pattern search numerical optimization algorithm. An example of such a parameterization is the spherical coordinate system $\mathbb{S}:[0,\pi]^{(d-2)}\times[0,2\pi]\mapsto\mathcal{S}_{d-1}$:

$$\begin{aligned}(\beta_1,\beta_2,\ldots,\beta_{d-1}) \mapsto &(\cos(\beta_1),\sin(\beta_1)\cos(\beta_2),\sin(\beta_1)\sin(\beta_2)\cos(\beta_3),\ldots,\\ &\sin(\beta_1)\ldots\sin(\beta_{d-2})\cos(\beta_{d-1}),\sin(\beta_1)\ldots\sin(\beta_{d-2})\sin(\beta_{d-1}))^T.\end{aligned}$$

The map parameterizing the positive half of the sphere $\mathbb{S}:\{(\beta_1,\beta_2,\ldots,\beta_{d-1})\in[0,1]^{(d-1)}:\|\boldsymbol{\beta}\|\leq 1\}\mapsto\mathcal{S}_{d-1}$:

$$(\beta_1,\beta_2,\ldots,\beta_{d-1})\mapsto\left(\beta_1,\beta_2,\ldots,\beta_{d-1},\sqrt{1-\beta_1^2-\ldots-\beta_{d-1}^2}\right)^T,$$

is another example that can be used provided α_d is positive. Prior knowledge about the position of $\boldsymbol{\alpha}_0$ can be derived from an initial estimate such as the LSE proposed in [1].

We illustrate the set of equations corresponding to (4.3) for dimension $d=3$ using the simulation model studied in Section 8. We consider the model

$$Y=\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})+\varepsilon,\quad ,\psi_0(\boldsymbol{x})=x^3,\quad \alpha_{01}=\alpha_{02}=\alpha_{03}=1/\sqrt{3},\qquad X_1,X_2,X_3\overset{i.i.d}{\sim}U[1,2],\quad \varepsilon\sim N(0,1),$$

where ε is independent of the covariate vector $\boldsymbol{X}=(X_1,X_2,X_3)^T$. For this model, we have

$$\boldsymbol{A}=\frac{17}{15}\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix},\quad \boldsymbol{\Sigma}=\frac{1}{36}\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix}\quad\text{and}\quad \tilde{\boldsymbol{A}}=\tilde{\boldsymbol{\Sigma}}=11.898545\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix},$$

where the matrices $\boldsymbol{A},\boldsymbol{\Sigma},\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{\Sigma}}$ are defined in (4.7), (4.8), (5.3) and (5.4) respectively. Note that the rank of the matrices is equal to $d-1=2$. We consider the parametrization

$$\mathcal{S}_3=\{(\alpha_1,\alpha_2,\alpha_3)=(\cos(\beta_1)\sin(\beta_2),\sin(\beta_1)\sin(\beta_2),\cos(\beta_2)):0\leq\beta_1\leq 2\pi,0\leq\beta_2\leq\pi\}\subset\mathbb{R}^2, \tag{6.1}$$

and we solve the problem

$$\begin{cases} s_1(\beta_1,\beta_2)=\frac{1}{n}\sum_{i=1}^n(-\sin(\beta_1)\sin(\beta_2)X_{i1}+\cos(\beta_1)\sin(\beta_2)X_{i2})\left\{Y_i-\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}=0,\\ s_2(\beta_1,\beta_2)=\frac{1}{n}\sum_{i=1}^n(\cos(\beta_1)\cos(\beta_2)X_{i1}+\sin(\beta_1)\cos(\beta_2)X_{i2}-\sin(\beta_2)X_{i3})\left\{Y_i-\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}=0.\end{cases} \tag{6.2}$$

Note that

$$\mathbb{S}(\boldsymbol{\beta}_0) = (\cos(\beta_{01})\sin(\beta_{02}), \sin(\beta_{01})\sin(\beta_{02}), \cos(\beta_{02}))^T = \left(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}\right)^T,$$

and

$$\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0) = \begin{bmatrix} -\sin(\beta_{01})\sin(\beta_{02}) & \cos(\beta_{01})\cos(\beta_{02}) \\ \cos(\beta_{01})\sin(\beta_{02}) & \sin(\beta_{01})\cos(\beta_{02}) \\ 0 & -\sin(\beta_{02}) \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & -\sqrt{\frac{2}{3}} \end{bmatrix},$$

where $\beta_{01} = \pi/4$ and $\beta_{02} = \arctan(\sqrt{2})$. We also have,

$$\mathbb{S}(\boldsymbol{\beta})^T \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}) = (0, 0)\,, \tag{6.3}$$

for all $\boldsymbol{\beta}$. This implies that the columns of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ are perpendicular to the vector $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$. Note moreover that the columns are linearly independent and hence form a basis for $\{\boldsymbol{\alpha}\}^{\perp}$. The asymptotic variance of $\hat{\boldsymbol{\alpha}}_n$ resp. $\tilde{\boldsymbol{\alpha}}_n$ defined in Theorem 4.1 resp. Theorem 5.1 is equal to,

$$\boldsymbol{A}^{-}\boldsymbol{\Sigma}\boldsymbol{A}^{-} = \frac{25}{2601}\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \quad \text{resp.} \quad \tilde{\boldsymbol{A}}^{-}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{A}}^{-} = \tilde{\boldsymbol{A}}^{-} = 0.009338\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}. \tag{6.4}$$

By the discontinuous nature of the functions s_1 and s_2 in (6.2), it is not possible to solve equations (6.2) exactly, we search the crossing of zero (see Definition 4.1), by minimizing the sum of squares $s_1^2(\boldsymbol{\beta}) + s_2^2(\boldsymbol{\beta})$ over all possible values of $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. Note that the crossing of zero of the score function is equivalent to the minimizer of the sum of squared component scores so that the minimization procedure is justified.

We use a derivative free optimization algorithm proposed by [12]. The method is a pattern-search optimization method that does not require the objective function to be continuous. The algorithm starts from an initial estimate of the minimum and looks for a better nearby point using a set of $2d$ equal step sizes along the coordinate axes in each direction, first making a step in the direction of the previous move. For the object function we take the sum of the squared values of the component functions, which achieve a minimum at a crossing of zero. If in no direction an improvement is found, the step size is halved, and a new search for improvement is done, with the reduced step sizes. This is repeated until the step size has reached a prespecified minimum. A very clear exposition of the method is given in [17], section 4.3. In this paper also convergence proofs for the optimization algorithm are presented.

7. Lagrange approach

Instead of tackling the fact that our parameter space is essentially of dimension $d-1$ by the parametrization $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$ which locally maps $\mathbb{R}^{d-1}$ into the sphere $\mathcal{S}_{d-1}$, one can introduce the restriction $\|\boldsymbol{\alpha}\| = 1$ via a Lagrangian term. We then consider the problem of minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}^2 + \lambda\left\{\|\boldsymbol{\alpha}\|^2 - 1\right\}, \tag{7.1}$$

where $\hat{\psi}_{n\boldsymbol{\alpha}}$ is the LSE defined in Section 3 and λ is a Lagrange parameter which we add to the sum of squared errors to deal with the identifiability of the single-index model. We consider a Lagrange penalty for solving the optimization problem under the constraint that $\|\boldsymbol{\alpha}\| = 1$.

Analogously to the treatment given in Section 4, we consider the derivative of (7.1) w.r.t. $\boldsymbol{\alpha}$, where we ignore the non-differentiability of the LSE $\hat{\psi}_{n\boldsymbol{\alpha}}$. This leads to the set of equations,

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij}\left\{\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i) - Y_i\right\} + \lambda\alpha_j = 0, \qquad j = 1, \ldots, d. \tag{7.2}$$

Here λ has to satisfy

$$\lambda = \lambda \sum_{j=1}^{d} \alpha_j^2 = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\alpha}^T \boldsymbol{X}_i \left\{ Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\}. \tag{7.3}$$

Plugging in the above expression for λ in (7.2), we would consider the score equation

$$\begin{aligned} \mathbf{0} &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \left\{ Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\} - \boldsymbol{\alpha}^T \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \left\{ Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\} \right) \boldsymbol{\alpha} \\ &= \left(\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\alpha}^T \right) \int \boldsymbol{x} \left\{ y - \hat{\psi}_{n\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T \boldsymbol{x}\right) \right\} d\mathbb{P}_n(\boldsymbol{x}, y), \end{aligned} \tag{7.4}$$

where $\boldsymbol{I}$ is the $d \times d$ identity matrix.

A computer program was implemented to solve (7.4). It has the advantage that we do not have to deal with the parametrization $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$, but has the disadvantage that we cannot assume that $\hat{\boldsymbol{\alpha}}_n$ has exactly norm 1 because we again have to deal with crossings of zero instead of exact equality to zero. One way to circumvent this problem is to normalize the solution after each iteration by dividing by its norm. This approach seems to provide reasonable solutions, although it is not entirely satisfactory from a theoretical point of view. Also note that if the right-hand side of (7.3) equals zero, so $\lambda = 0$, the equation does not force the norm of α to be one; it only does so if $\lambda \neq 0$. Indeed, in our computer experiments, λ was never zero, so this problem did actually not occur, but λ will tend to zero with increasing sample sizes, so some numerical instability is to be expected.

For reasons of space we do not further describe all details of this approach, but instead restrict ourselves to showing a picture of the simple score estimate of ψ_0 for $n = 1000$ and $d = 10$ for the simulation where all the X_i variables and the random error variable ε are standard normal and independent, $\psi_0(\boldsymbol{x}) = x^3$ and $\boldsymbol{\alpha}_0 = (1/\sqrt{10}, \ldots, 1/\sqrt{10})^T$. It is clear that the estimate of ψ_0 will be rather accurate because of the information provided by the 10 covariates X_i (instead of, say, just two covariates X_1, X_2).

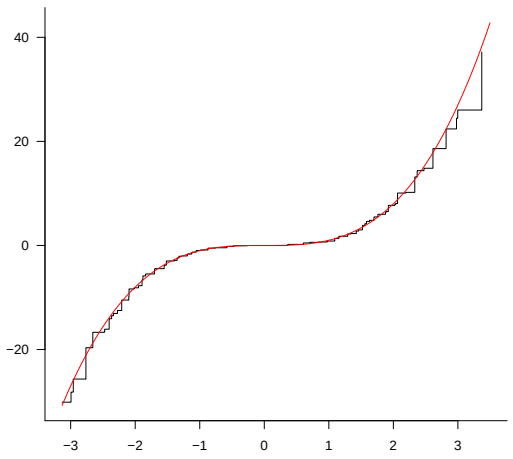

Fig 2: The real ψ_0 (red, solid) and the function $\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}$ (black, step function) for $\psi_0(\boldsymbol{x}) = x^3$, $\alpha_{0i} = 1/\sqrt{10}$ and $X_i \overset{i.i.d}{\sim} N(0,1), i = 1 \ldots 10$.

8. Simulations

In this section we illustrate the finite sample behavior of the single index score estimators proposed in Section 4 and Section 5. We consider the model

$$Y = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) + \varepsilon = \left(\boldsymbol{\alpha}_0^T \boldsymbol{X}\right)^3 + \varepsilon, \tag{8.1}$$

where $\alpha_{0i} = 1/\sqrt{d}, i = 1, \ldots, d$ and $\varepsilon \sim N(0,1)$, independent of $\boldsymbol{X}$. We consider two different dimensions $d = 2$ and $d = 3$ and two different distributions for the covariate vector $\boldsymbol{X}$, $X_i \overset{i.i.d}{\sim} U[1,2]$ and $X_i \overset{i.i.d}{\sim} N(0,1)$ for $i = 1, \ldots, d$. This yields four different simulation set-ups.

For the two dimensional models, we consider the parametrization $\mathbb{S}(\beta_0) = (\cos(\beta_0), \sin(\beta_0))^T$ and we use the parametrization given in (6.1) for the three dimensional models. In each simulation setting, we estimate $\boldsymbol{\alpha}_0$ by the simple score estimator (SSE) and the efficient score estimator (ESE) and compare the behavior of our proposed estimates with the least squares estimate (LSE) minimizing (1.1) over all possible $(\psi, \boldsymbol{\alpha})$ and the maximum rank correlation estimate (MRCE) proposed in [9]. For sample sizes $n = 100, 500, 1000, 2000, 5000$ and $n = 10000$ we generate $N = 5000$ datasets from Model (8.1) and show, in Tables 1-4, the mean and n times the covariance of the estimates. Tables 1-4 also show the asymptotic values to which the results should converge based on Theorem 4.1 and Theorem 5.1. The limiting distribution of the LSE is still unknown and therefore no asymptotic results are provided for the LSE. For the MRCE, [15] has an expression for the asymptotic covariance matrix in an (implicit) $(d-1)$-dimensional representation in his Theorem 4 on p. 133. If, in accordance with the methods of our paper, we turn this into an expression in terms of our d-dimensional representation, we obtain as the asymptotic covariance matrix of the MRCE $\boldsymbol{V}^- \boldsymbol{\Sigma} \boldsymbol{V}^-$ where

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T S(Y, \boldsymbol{\alpha}_0^T\boldsymbol{X})^2 g_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\right],$$

and $\boldsymbol{V}^-$ is the Moore-Penrose inverse of

$$\boldsymbol{V} = \mathbb{E}\left[\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T S_2(Y, \boldsymbol{\alpha}_0^T\boldsymbol{X}) g_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\right],$$

and g_0 is the density of $\boldsymbol{\alpha}_0^T\boldsymbol{X}$ and S and S_2 are defined by:

$$S(y,u) = E\left[1_{\{y>Y\}} - 1_{\{y<Y\}} \mid \boldsymbol{\alpha}_0^T\boldsymbol{X} = u\right], \qquad S_2(y,u) = \frac{\partial}{\partial u} S(y,u).$$

It is clear from our simulations that the factor 2 in front of $\boldsymbol{V}$ in (20) of [15] cannot be correct and indeed [4] have a note on p. 361 of their paper, attributed to Myoung-Jae Lee that this factor 2 should not be there.

For all simulation studies, the results shown in Tables 1-4 show convergence of n times the variance-covariance matrices towards its asymptotic values.

The performance of the ESE is slightly better than the performance of the SSE; the difference between the asymptotic limiting variances is smaller in the model with uniform$[1,2]$ covariates X_i than the difference in the model with standard normal covariates X_i. Tables 1-4 also illustrate that n times the variances of the estimates decrease if the dimension d of the model increases in the model with $X_i \sim U[1,2]$ but increases in the model with $X_i \sim N(0,1)$.

Although the model with standard normal covariates violates Assumptions A1, A2 and A4 given in Section 3, our proposed estimates perform reasonably well.

The performance of the MRCE is worse than the performances of our proposed score estimates in all simulation settings. In the model with standard normal covariates, the variances of the MRCE are remarkable larger than the variances of the score estimates and the LSE. Also, the asymptotic variances of the MRCE in these models are considerably larger than the variances of the score estimators. This might be caused by the fact that only the indicators $1_{\{y>Y_i\}}$ and $1_{\{y>Y_i\}}$ are used in the definition of the MRCE and not the actual values of the Y_i's.

The behavior of the LSE is rather remarkable. Tables 1 and 2 suggest an increase of n times the covariance matrix, whereas Tables 3 and 4 suggest a decrease.

The results presented in Tables 3 and 4 show that the performance of the LSE is better than the performance of the SSE for small sample sizes when $X_i \sim N(0,1)$. For the model with uniform covariates,

summarized in Tables 1 and 2 our proposed score estimates outperform the LSE. The variances for the LSE presented in Tables 1-4 suggest that the rate of convergence for the LSE is faster than the cube-root $n^{1/3}$-rate proved in [1]. The asymptotic distribution of the LSE falls beyond the scope of this paper and needs to be addressed in further research.

TABLE 1
Simulation model ($X_i \sim U[1,2], d = 2$): The mean value ($\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in}), i = 1,2$) and n times the variance-covariance ($\hat{\sigma}_{ij} = n{\cdot}\text{cov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn}), i,j = 1,2$) of the simple score estimate (SSE), the efficient score estimate (ESE), the least squares estimate (LSE) and the maximum rank correlation estimate (MRCE) for different sample sizes n with $N = 5000$ and $X_i \sim U[1,2]$. The line, preceded by ∞, gives the asymptotic values.

Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$
SSE	100	0.707249	0.706389	0.040773	0.040866	-0.040787
	500	0.707224	0.706890	0.035018	0.035057	-0.035033
	1000	0.707175	0.706992	0.033262	0.033273	-0.033265
	2000	0.707173	0.707016	0.034017	0.034012	-0.034014
	5000	0.707134	0.707070	0.034011	0.034012	-0.034011
	10000	0.707109	0.707100	0.033344	0.033350	-0.033347
	∞	0.707107	0.707107	0.032439	0.032439	-0.032439
ESE	100	0.707293	0.706359	0.039631	0.039758	-0.039663
	500	0.707230	0.706888	0.033888	0.033922	-0.033900
	1000	0.707185	0.706983	0.032302	0.032316	-0.032307
	2000	0.707175	0.707015	0.032992	0.032989	-0.032990
	5000	0.707130	0.707074	0.032925	0.032924	-0.032924
	10000	0.707111	0.707098	0.032278	0.032283	-0.032280
	∞	0.707107	0.707107	0.031516	0.031516	-0.031516
LSE	100	0.706848	0.706624	0.052397	0.052415	-0.052321
	500	0.707060	0.707002	0.053547	0.053570	-0.053542
	1000	0.707138	0.707000	0.053513	0.053573	-0.053535
	2000	0.707122	0.707053	0.055502	0.055519	-0.055506
	5000	0.707118	0.707079	0.059731	0.059756	-0.059741
	10000	0.707128	0.707077	0.061843	0.061868	-0.061854
	∞	0.707107	0.707107	?	?	?
MRCE	100	0.707301	0.706188	0.051226	0.051272	-0.051182
	500	0.707178	0.706910	0.044397	0.044428	-0.044402
	1000	0.707107	0.707047	0.041886	0.041904	-0.041891
	2000	0.707145	0.707039	0.041231	0.041248	-0.041237
	5000	0.707154	0.707048	0.040968	0.040976	-0.040971
	10000	0.707125	0.707083	0.040129	0.040141	-0.040135
	∞	0.707107	0.707107	0.035535	0.035535	-0.035535

9. Discussion

In this paper we propose a method for estimating the regression parameter in the monotone single index model. Our estimators are defined as zero-crossings of an unsmooth score function which contains the LSE of the monotone link function. The estimation approach extends the results of [7] for the current status linear regression model and as far as we know, it is the first time that $\sqrt{n}-$consistent estimates of the regression parameter are constructed based on the piecewise-constant LSE for the link function. Good performances, both asymptotically and numerically, of our estimation approach are illustrated by the asymptotic normality of our score estimators and by simulation studies that show comparable or even better behavior of our score estimates compared to the maximum rank correlation estimate proposed by [9] and the least squares estimate discussed in [1].

A limitation of our approach might be that we use a local parametrization of the sphere to obtain the

TABLE 2
Simulation model ($X_i \sim U[1,2], d = 3$): The mean value ($\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in}), i = 1,2,3$) and n times the variance-covariance ($\hat{\sigma}_{ij} = n \cdot \text{cov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn}), i,j = 1,2,3$) of the simple score estimate (SSE), the efficient score estimate (ESE), the least squares estimate (LSE) and the maximum rank correlation estimate (MRCE) for different sample sizes n with $N = 5000$ and $X_i \sim U[1,2]$. The line, preceded by ∞, gives the asymptotic values.

Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{33}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{13}$	$\hat{\sigma}_{23}$
SSE	100	0.577082	0.576760	0.577536	0.026000	0.026531	0.025168	-0.013676	-0.012352	-0.012797
	500	0.577066	0.577370	0.577505	0.020865	0.021461	0.020629	-0.010840	-0.010012	-0.010623
	1000	0.577128	0.577364	0.577505	0.020471	0.020915	0.020576	-0.010394	-0.010054	-0.010531
	2000	0.577232	0.577281	0.577512	0.020101	0.020454	0.020325	-0.010107	-0.009987	-0.010346
	5000	0.577286	0.577343	0.577412	0.019549	0.019796	0.019966	-0.009688	-0.009855	-0.010112
	10000	0.577324	0.577351	0.577370	0.019208	0.019732	0.019746	-0.009597	-0.009609	-0.010138
	∞	0.577350	0.577350	0.577350	0.019223	0.019223	0.019223	-0.009612	-0.009612	-0.009612
ESE	100	0.576080	0.577006	0.578298	0.025577	0.026523	0.024775	-0.013601	-0.011903	-0.012910
	500	0.576671	0.577423	0.577850	0.020379	0.020804	0.020022	-0.010557	-0.009793	-0.010250
	1000	0.576861	0.577388	0.577750	0.019936	0.020265	0.019954	-0.010099	-0.009803	-0.010174
	2000	0.577055	0.577293	0.577677	0.019503	0.019941	0.019749	-0.009830	-0.009657	-0.010108
	5000	0.577175	0.577351	0.577516	0.019054	0.019269	0.019449	-0.009431	-0.009612	-0.009843
	10000	0.577251	0.577359	0.577436	0.018714	0.019187	0.019205	-0.009344	-0.009364	-0.009845
	∞	0.577350	0.577350	0.577350	0.018677	0.018677	0.018677	-0.009338	-0.009338	-0.009338
LSE	100	0.576900	0.577212	0.576726	0.046654	0.047393	0.045991	-0.023955	-0.022584	-0.023394
	500	0.577271	0.577276	0.577256	0.047999	0.047934	0.047352	-0.024291	-0.023683	-0.023653
	1000	0.577343	0.577285	0.577294	0.049601	0.050000	0.049575	-0.024999	-0.024590	-0.024991
	2000	0.577440	0.577277	0.577267	0.050351	0.051728	0.051742	-0.025168	-0.025194	-0.026544
	5000	0.577379	0.577338	0.577305	0.054862	0.055341	0.054135	-0.028024	-0.026838	-0.027305
	10000	0.577335	0.577356	0.577345	0.058277	0.057892	0.058661	-0.028773	-0.029497	-0.029143
	∞	0.577350	0.577350	0.577350	?	?	?	?	?	?
MRCE	100	0.576976	0.576965	0.576916	0.046496	0.046523	0.044848	-0.024092	-0.022429	-0.022333
	500	0.577285	0.577414	0.577265	0.017158	0.016660	0.016967	-0.008421	-0.008723	-0.008247
	1000	0.577285	0.577414	0.577265	0.034317	0.033319	0.033935	-0.016843	-0.017445	-0.016493
	2000	0.577339	0.577325	0.577347	0.030230	0.030323	0.031574	-0.014493	-0.015744	-0.015825
	5000	0.577381	0.577371	0.577285	0.028795	0.028768	0.029245	-0.014163	-0.014630	-0.014611
	10000	0.577370	0.577355	0.577319	0.026680	0.027576	0.027700	-0.013279	-0.013402	-0.014297
	∞	0.577350	0.577350	0.577350	0.021436	0.021436	0.021436	-0.010718	-0.010718	-0.010718

estimates. We however showed that the asymptotic properties of the resulting regression parameter estimators is independent of the parametrization used and we moreover developed a computationally interesting Lagrangian approach that avoids the use of a local parametrization and allows for easy estimation in high dimensions. The theory for the Lagrangian procedure is complicated by the fact that we do no longer work on the sphere, which is rather unusual in the theory on single index models where identifiability assumptions are commonly used. Our simulations however suggest that its asymptotic properties are the same as those presented in this paper.

The presence/absence of smoothing procedures involved with the estimation techniques is one of the main difference between our two score estimates. Although the simple score estimator, avoiding any smoothing parameters is not an efficient estimate of the regression parameter, its performance in finite samples is not remarkably worse than the performance of the efficient score estimates. By avoiding the use of a smooth estimate for the derivative of the link function, the simple score approach has the advantage that no bandwidth selection procedures are needed, which makes it computationally more attractive.

TABLE 3
Simulation model ($X_i \sim N(0,1), d=2$): The mean value ($\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in}), i=1,2$) and n times the variance-covariance ($\hat{\sigma}_{ij} = n\cdot\text{cov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn}), i,j=1,2$) of the simple score estimate (SSE), the efficient score estimate (ESE), the least squares estimate (LSE) and the maximum rank correlation estimate (MRCE) for different sample sizes n with $N=5000$ and $X_i \sim N(0,1)$. The line, preceded by ∞, gives the asymptotic values.

Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$
SSE	100	0.705202	0.706163	0.201378	0.200975	-0.200327
	500	0.706395	0.707509	0.109547	0.109291	-0.109371
	1000	0.706824	0.707259	0.092557	0.092464	-0.092494
	2000	0.707100	0.707056	0.080981	0.080966	-0.080967
	5000	0.707026	0.707167	0.072021	0.071947	-0.071982
	10000	0.707091	0.707113	0.067685	0.067665	-0.067674
	∞	0.707107	0.707107	0.055556	0.055556	-0.055556
ESE	100	0.706800	0.706173	0.087513	0.087878	-0.087480
	500	0.706905	0.707200	0.038450	0.038468	-0.038453
	1000	0.706978	0.707190	0.031701	0.031672	-0.031685
	2000	0.707061	0.707133	0.027930	0.027924	-0.027926
	5000	0.707079	0.707128	0.023914	0.023907	-0.023911
	10000	0.707104	0.707106	0.022827	0.022828	-0.022827
	∞	0.707107	0.707107	0.018519	0.018519	0.018519
LSE	100	0.706748	0.706135	0.093819	0.094319	-0.093715
	500	0.706710	0.707309	0.068561	0.068260	-0.068383
	1000	0.706737	0.707389	0.061614	0.061325	-0.061459
	2000	0.707009	0.707161	0.061123	0.061109	-0.061111
	5000	0.707087	0.707110	0.060759	0.060722	-0.060738
	10000	0.707074	0.707131	0.061708	0.061692	-0.061699
	∞	0.707107	0.707107	?	?	?
MRCE	100	0.701888	0.704656	0.542166	0.539405	-0.530421
	500	0.705470	0.707715	0.364976	0.360968	-0.362295
	1000	0.706374	0.707364	0.336674	0.335793	-0.335971
	2000	0.707090	0.706892	0.328658	0.328860	-0.328632
	5000	0.706895	0.707229	0.317497	0.317000	-0.317202
	10000	0.707062	0.707107	0.313625	0.313331	-0.313456
	∞	0.707107	0.707107	0.268711	0.268711	-0.268711

10. Appendix

In this Section, we give the proofs of the results stated in Sections 3 and 4. The results given in Section 5 together with additional technical lemmas needed for proving our main results are given in the supplementary material [2]. We will write "Lemma X in Supplement Y" when we refer to these results in the remainder of the Appendix. Entropy results are used in our proofs. Before we start the proofs, we first introduce some notations and definitions used in the remainder of the Appendix.

We will denote the L_2-norm of a function f defined on $\mathcal{X}\times\mathbb{R}$ with respect to some probability measure $\mathbb{P}$ by $\|\cdot\|_{\mathbb{P}}$; i.e.,

$$\|f\|_{\mathbb{P}} = \mathbb{P}(f^2)^{1/2} = \left(\int_{\mathcal{X}} f^2(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x})\right)^{1/2}. \tag{10.1}$$

Also, we will denote by $\|\cdot\|_{B,\mathbb{P}}$ the Bernstein norm of a function f defined on $\mathcal{X}\times\mathbb{R}$ which is given by

$$\|f\|_{\mathbb{P},B} = \left(2\mathbb{P}\Big(e^{|f|} - |f| - 1\Big)\right)^{1/2}. \tag{10.2}$$

For both norms, $\mathbb{P}$ will taken to be P_0, the true joint probability measure of the $(\boldsymbol{X}, Y)$. Note that when f

TABLE 4
Simulation model ($X_i \sim N(0,1), d=3$): The mean value ($\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in}), i=1,2,3$) and n times the variance-covariance ($\hat{\sigma}_{ij} = n\cdot\text{cov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn}), i,j=1,2,3$) of the simple score estimate (SSE), the efficient score estimate (ESE), the least squares estimate (LSE) and the maximum rank correlation estimate (MRCE) for different sample sizes n with $N=5000$ and $X_i \sim N(0,1)$. The line, preceded by ∞, gives the asymptotic values.

Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{33}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{13}$	$\hat{\sigma}_{23}$
SSE	100	0.570983	0.575590	0.578029	0.263776	0.309305	0.282773	-0.144522	-0.114110	-0.165663
	500	0.575676	0.577118	0.578470	0.141363	0.161205	0.149813	-0.076137	-0.064089	-0.085601
	1000	0.576424	0.577186	0.578120	0.123481	0.124819	0.121335	-0.063075	-0.060038	-0.061653
	2000	0.576826	0.577144	0.577945	0.104429	0.104871	0.103671	-0.052625	-0.051672	-0.052170
	5000	0.577037	0.577330	0.577635	0.093606	0.097177	0.093909	-0.048362	-0.045185	-0.048793
	10000	0.577107	0.577377	0.577543	0.087822	0.092609	0.089642	-0.045356	-0.042420	-0.047258
	∞	0.577350	0.577350	0.577350	0.074074	0.074074	0.074074	-0.037037	-0.037037	-0.037037
ESE	100	0.571762	0.577007	0.579906	0.123299	0.141033	0.121794	-0.070078	-0.049452	-0.071866
	500	0.575761	0.577456	0.578542	0.056520	0.059089	0.051253	-0.032075	-0.023321	-0.027809
	1000	0.576414	0.577396	0.578128	0.043292	0.043200	0.041763	-0.022266	-0.020903	-0.020953
	2000	0.576780	0.577305	0.577918	0.036568	0.036203	0.036525	-0.018055	-0.018448	-0.018145
	5000	0.577012	0.577403	0.577620	0.030453	0.032063	0.032005	-0.015234	-0.015195	-0.016831
	10000	0.577125	0.577401	0.577517	0.029584	0.029680	0.030257	-0.014494	-0.015074	-0.015193
	∞	0.577350	0.577350	0.577350	0.024691	0.024691	0.024691	-0.012346	-0.012346	-0.012346
LSE	100	0.575121	0.574814	0.577588	0.173704	0.174857	0.173169	-0.085890	-0.086918	-0.086607
	500	0.576833	0.577377	0.577289	0.107186	0.104621	0.106974	-0.052394	-0.054652	-0.052262
	1000	0.577022	0.577354	0.577415	0.101079	0.098209	0.100486	-0.049366	-0.051622	-0.048862
	2000	0.577337	0.577418	0.577177	0.092130	0.091427	0.089505	-0.047012	-0.045113	-0.044391
	5000	0.577311	0.577469	0.577224	0.090412	0.088701	0.089865	-0.044678	-0.045745	-0.044061
	10000	0.577258	0.577498	0.577272	0.088953	0.085219	0.089801	-0.042188	-0.046740	-0.043056
	∞	0.577350	0.577350	0.577350	?	?	?	?	?	?
MRCE	100	0.568656	0.570735	0.571815	0.794208	0.796010	0.802239	-0.389900	-0.384086	-0.387755
	500	0.576584	0.576612	0.576154	0.513244	0.520888	0.524345	-0.253712	-0.257288	-0.265869
	1000	0.577444	0.576655	0.576699	0.487464	0.482635	0.475335	-0.246720	-0.240780	-0.234323
	2000	0.577241	0.576968	0.577275	0.436314	0.434728	0.436539	-0.217338	-0.218922	-0.217178
	5000	0.577312	0.577253	0.577265	0.416947	0.430323	0.426823	-0.210246	-0.206836	-0.219833
	10000	0.577239	0.577371	0.577334	0.398551	0.418207	0.410919	-0.202886	-0.195591	-0.215307
	∞	0.577350	0.577350	0.577350	0.358281	0.358281	0.358281	-0.179141	-0.179141	-0.179141

is only a function of $\boldsymbol{x} \in \mathcal{X}$ then

$$\|f\|_{P_0} = \left(\int_{\mathcal{X}} f^2(\boldsymbol{x}) dG(\boldsymbol{x})\right)^{1/2} = \left(\int_{\mathcal{X}} f^2(\boldsymbol{x}) g(\boldsymbol{x}) d\boldsymbol{x}\right)^{1/2},$$

by Assumption A5.

For a class of functions $\mathcal{F}$ on $\mathcal{R}$ equipped with a norm $\|\cdot\|$, we let $N_B(\zeta, \mathcal{F}, \|\cdot\|)$ denote the minimal number N for which there exists pairs of functions $\{[g_j^L, g_j^U], j=1,\ldots,N\}$ such that $\|g_j^U - g_j^L\| \leq \zeta$ for all $j=1,\ldots,N$ and such that for each $g \in \mathcal{F}$ there is a $j \in \{1,\ldots,N\}$ such that $g_j^L \leq g \leq g_j^U$. The $\zeta-$entropy with bracketing of $\mathcal{F}$ is defined as $H_B(\zeta, \mathcal{F}, \|\cdot\|) = \log(N_B(\zeta, \mathcal{F}, \|\cdot\|))$.

Results on entropy calculations used in proving our main results are given in Section 2. Our proofs use inequalities for empirical processes described in Lemma 3.4.2 and Lemma 3.4.3 of [19].

Lemma 3.4.2 ([19]) Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{\mathbb{P}} \leq \delta$ and $\|f\|_\infty \leq M$ for every f in $\mathcal{F}$. Then

$$\mathbb{E}_{\mathbb{P}}\left[\|\mathbb{G}_n\|_{\mathcal{F}}\right] \lesssim J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P}}) \left(1 + \frac{J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P}})}{\sqrt{n}\delta^2} M\right),$$

where

$$J_n(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + H_B(\epsilon, \mathcal{F}, \|\cdot\|)}d\epsilon$$

Lemma 3.4.3 ([19]) Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{\mathbb{P},B} \leq \delta$ for every f in $\mathcal{F}$. Then

$$\mathbb{E}_{\mathbb{P}}\Big[\|\mathbb{G}_n\|_{\mathcal{F}}\Big] \lesssim J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P},B}) \left(1 + \frac{J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P},B})}{\sqrt{n}\delta^2}\right),$$

In the sequel, and whenever the ϵ-bracketing entropy of some class $\mathcal{F}$ with respect to some norm $\|\cdot\|$ is bounded above by $C\epsilon^{-1}$ for some constant $C > 0$ (which may depend on n), we will write for all $e > \epsilon$

$$J_n(d) = \int_0^e (1 + C/\epsilon)^{1/2} d\epsilon \tag{10.3}$$

Moreover, we will use the inequality

$$J_n(d) \leq d + 2C^{1/2}e^{1/2} \tag{10.4}$$

which is an immediate consequence of the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \geq 0$.

10.1. Appendix A: The least squares estimator (LSE) $\hat{\psi}_{n\boldsymbol{\alpha}}$

In this section we first prove Proposition 3.1. We next show in Lemma 10.1 that the LSE $\hat{\psi}_{n\boldsymbol{\alpha}}$ is of order $O_p(\log n)$ uniformly in $\mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. This result is used in the proof of Proposition 3.2, given at the end of this section.

Proof of Proposition 3.1. Note that with $\boldsymbol{X} \sim g$ we can write

$$L_{\boldsymbol{\alpha}}(\psi) = \mathbb{E}\Big[\Big(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) - \psi(\boldsymbol{\alpha}^T\boldsymbol{X})\Big)^2\Big].$$

Thus,

$$\mathbb{E}\Big[\Big(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) - \mathbb{E}(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) \mid \boldsymbol{\alpha}^T\boldsymbol{X})\Big)^2\Big] = \min_{\psi \in \mathcal{M}'} L_{\boldsymbol{\alpha}}(\psi),$$

with $\mathcal{M}'$ the set of all bounded Borel-measurable function defined on $\mathcal{I}_{\boldsymbol{\alpha}}$. Therefore, if the minimizing function $u \mapsto \mathbb{E}(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) \mid \boldsymbol{\alpha}^T\boldsymbol{X} = u)$ is monotone increasing on $\mathcal{I}_{\boldsymbol{\alpha}}$, then this implies that it necessarily minimizes $L_{\boldsymbol{\alpha}}$ over $\mathcal{M}$. Furthermore, such a minimizer is unique by strict convexity of $L_{\boldsymbol{\alpha}}$. □

Lemma 10.1.

$$\max_{\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)} \sup_{\boldsymbol{x} \in \mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| = O_p(\log n).$$

Proof. The proof of this lemma is similar to that of Lemma 4.4 of [1]. For a fixed $\boldsymbol{\alpha}$ it follows from the min-max formula of an isotonic regression that we have for all $\boldsymbol{x} \in \mathcal{X}$

$$\min_{1 \leq k \leq n} \frac{\sum_{i=1}^k Y_i^{\boldsymbol{\alpha}}}{k} \leq \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) \leq \max_{1 \leq k \leq n} \frac{\sum_{i=k}^n Y_i^{\boldsymbol{\alpha}}}{n-k+1}.$$

Hence,

$$\min_{1 \leq i \leq n} Y_i \leq \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) \leq \max_{1 \leq i \leq n} Y_i$$

and this in turn implies that

$$\max_{\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)} \sup_{\boldsymbol{x} \in \mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| \leq \max_{1 \leq i \leq n} |Y_i|.$$

Using similar arguments as in [1], we use Assumption A7 to show that $\max_{1 \leq i \leq n} |Y_i| = O_p(\log n)$, which completes the proof. □

Proof of Proposition 3.2. By the definition of the LSE of the unknown monotone link function, $\hat{\psi}_{n\boldsymbol{\alpha}}$ maximizes the map $\psi \mapsto \mathbb{M}_n$ over $\mathcal{M}$ where,

$$\mathbb{M}_n(\psi, \boldsymbol{\alpha}) = \int_{\mathcal{X}\times\mathbb{R}} \Big(2y\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x})\Big) d\mathbb{P}_n(\boldsymbol{x}, y). \tag{10.5}$$

Moreover, $\psi_{\boldsymbol{\alpha}}$ maximizes the map $\psi \mapsto \mathbb{M}$ over $\mathcal{M}$, where

$$\mathbb{M}(\psi, \boldsymbol{\alpha}) = \int_{\mathcal{X}\times\mathbb{R}} \Big(2y\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x})\Big) dP_0(\boldsymbol{x}, y). \tag{10.6}$$

Define the function $f_{\psi,\boldsymbol{\alpha}}$ by,

$$f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x}, y) = 2y\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x})$$

Note that by definition of the LSE as the maximizer of (10.5), we have

$$\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\hat{\psi}_{n\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y)\Big) d\mathbb{P}_n(\boldsymbol{x}, y) \geq 0.$$

Moreover for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ and $\psi \in \mathcal{M}$, we have,

$$\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x}, y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y)\Big) dP_0(\boldsymbol{x}, y) = -d^2_{\boldsymbol{\alpha}}(\psi, \psi_{\boldsymbol{\alpha}}).$$

where, for any $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ and for any two elements ψ_1 and ψ_2 in $\mathcal{M}$ we define the squared distance

$$d^2_{\boldsymbol{\alpha}}(\psi_1, \psi_2) = \int_{\mathcal{X}} \Big(\psi_2(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_1(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)^2 g(\boldsymbol{x}) d\boldsymbol{x}.$$

This can be seen as follows:

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x}, y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y)\Big) dP_0(\boldsymbol{x}, y) \\
&= \int_{\mathcal{X}\times\mathbb{R}} \Big(2\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x}) + \psi^2_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big) dP_0(\boldsymbol{x}, y) \\
&= -\int_{\mathcal{X}} \Big(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)^2 g(\boldsymbol{x}) d\boldsymbol{x} = -d^2_{\boldsymbol{\alpha}}(\psi, \psi_{\boldsymbol{\alpha}}),
\end{aligned}$$

where we use that $\mathbb{E}\{Y|\boldsymbol{\alpha}^T\boldsymbol{X} = u\} = \psi_{\boldsymbol{\alpha}}(u)$ This implies that, for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ and $\psi \in \mathcal{M}$, we have,

$$\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x}, y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y)\Big) d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \geq d^2_{\boldsymbol{\alpha}}(\psi, \psi_{\boldsymbol{\alpha}}).$$

We write,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n\boldsymbol{\alpha}}, \psi_{\boldsymbol{\alpha}}) \geq \epsilon\right\} \\
&\leq \mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0), d_{\boldsymbol{\alpha}}(\hat{\psi}_{n\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}})\geq\epsilon} \left\{\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\hat{\psi}_{n\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y)\Big) d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) - d^2_{\boldsymbol{\alpha}}(\hat{\psi}_{n\boldsymbol{\alpha}}, \psi_{\boldsymbol{\alpha}})\right\} \geq 0, \right. \\
&\qquad\qquad \left. \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n\boldsymbol{\alpha}}, \psi_{\boldsymbol{\alpha}}) \geq \epsilon\right\} \\
&\leq \mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0), \psi\in\mathcal{M}_{RK}, d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\geq\epsilon} \left\{\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x}, y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x}, y)\Big) d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) - d^2_{\boldsymbol{\alpha}}(\psi_{\boldsymbol{\alpha}}, \psi)\right\} \geq 0, \right. \\
&\qquad \left. \max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| \leq K\right\} + \mathbb{P}\left\{\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| > K\right\}.
\end{aligned}$$

Fix $\nu > 0$. Since

$$\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| = O_p(\log n),$$

by Lemma 10.1, we can find $K_1 > 0$ large enough such that

$$\mathbb{P}\left\{\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| > K_1 \log n\right\} < \nu/2.$$

Define

$$\mathcal{M}_{RK} = \Big\{\psi \text{ monotone non-decreasing on } [-R,R] \text{ and bounded by } K\Big\}, \tag{10.7}$$

and consider the related class

$$\mathcal{F}_{RK} = \Big\{f(\boldsymbol{x},y) = 2y\Big(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big) - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})^2 + \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})^2, \\ (\boldsymbol{\alpha},\psi) \in \mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)\times\mathcal{M}_{RK} \text{ and } (\boldsymbol{x},y)\in\mathcal{X}\times\mathbb{R}\Big\}, \tag{10.8}$$

and for some $v > 0$

$$\mathcal{F}_{RKv} := \Big\{f \in \mathcal{F}_{RK} : d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}}) \le v \text{ for all } \boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)\Big\}. \tag{10.9}$$

Note now that the class $\mathcal{F}_{RKv}$ is included in the class $\mathcal{H}_{RC\delta}$ defined in Lemma 2.4 of Supplement B of [2] with $C = 2K^2$ and $\delta = 2Kv$. This holds true provided that $K_0 \le K$, and $K \ge 1$ which we can assume for n large enough since K will be chosen to be of order $\log n$. To see the claimed inclusion, it is enough to show that if m is an nondecreasing function $[-R,R]$ then m^2 can be written as the difference of two monotone functions. This is true because $m^2 = m^2\mathbb{I}_{m\ge 0} - (-m^2)\mathbb{I}_{m<0}$, and m^2 and $-m^2$ are nondecreasing on the subsets $\{m \ge 0\}$ and $\{m < 0\}$ respectively. When restricting attention to the event that $\hat{\psi}_{n\boldsymbol{\alpha}}$ is bounded by K for n large enough, we can consider only monotone functions $\psi \in \mathcal{M}_{RK}$. Using the expression of $\psi_{\boldsymbol{\alpha}}$ the latter is bounded by $K_0 \le K$. On the other hand, for any function $f \in \mathcal{F}_{RKv}$, there exist nondecreasing monotone functions f_1 and f_2 such that $\psi^2 - \psi_{\boldsymbol{\alpha}}^2 = f_2 - f_1$, such that $\|f_1\|_\infty, \|f_2\|_\infty \le K^2 + K_0^2 \le 2K^2$. Using that $K \ge 1$ implies that $\|2\psi\|_\infty, \|2\psi_{\boldsymbol{\alpha}}\|_\infty \le 2K \le 2K^2$. To finish, note that for any $\boldsymbol{\alpha}$ we have that $\int_{\mathcal{X}} \left(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 dG(\boldsymbol{x}) \le v^2$ we also have that

$$\begin{aligned}\int_{\mathcal{X}} \left(\psi^2(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}^2(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 dG(\boldsymbol{x}) &\le (2K)^2 \int_{\mathcal{X}} \left(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 dG(\boldsymbol{x}) \\ &\le 4K^2v^2.\end{aligned}$$

The calculation above implies that we can take $\delta = 2Kv$. Using the result of Lemma 2.4 in Supplement B of [2], it follows that the related class $\widetilde{\mathcal{F}}_{RKv} = \tilde{D}^{-1}\mathcal{F}_{RKv}$ with $\tilde{D} = 16M_0C = 32M_0K^2$ and a given $v > 0$ satisfies

$$\begin{aligned}H_B\big(\epsilon, \widetilde{\mathcal{F}}_{RKv}, \|\cdot\|_{B,P_0}\big) &\le H_B\big(\epsilon, \widetilde{\mathcal{H}}_{RKv}, \|\cdot\|_{B,P_0}\big) \\ &\le \frac{A}{\epsilon}\end{aligned}$$

for some constant $A > 0$ (depending only on d and the other parameters of the problem), and that for all $\tilde{f} \in \widetilde{F}_{RKv}$ we have $\|\tilde{f}\|_{B,P_0} \le \tilde{D}^{-1}\delta = (32M_0K^2)^{-1}2Kv = (16M_0)^{-1}K^{-1}v \equiv A_0K^{-1}v$. It follows from Lemma 3.4.3 of [19] that

$$\begin{aligned}\mathbb{E}\Big[\|\mathbb{G}_n\|_{\widetilde{F}_{RKv}}\Big] &\lesssim J_n(A_0K^{-1}v), \left(1 + K^2\frac{J_n(A_0K^{-1}v)}{\sqrt{n}A_0^2v^2}\right), \quad \text{where } J_n \text{ is defined in (10.3)} \\ &\le A_0K^{-1}v + 2A_0^{1/2}K^{-1/2}v^{1/2}A^{1/2}, \quad \text{using the inequality in (10.4)} \\ &\le B_0K^{-1/2}(v + v^{1/2})\end{aligned}$$

for some constant $B_0 > 0$, where we used the fact that $K^{-1/2} \geq K^{-1}$. Therefore,

$$\begin{aligned}\mathbb{E}\Big[\|\mathbb{G}_n\|_{\widetilde{F}_{RKv}}\Big] &\lesssim B_0 K^{-1/2}(v+v^{1/2})\left(1+K^2\frac{B_0K^{-1/2}(v+v^{1/2})}{\sqrt{n}A_0^2v^2}\right)\\ &\leq C_0K^{-1/2}(v+v^{1/2})\left(1+C_0K^{3/2}\frac{1+v^{1/2}}{\sqrt{n}v^{3/2}}\right).\end{aligned}$$

Using the definition of the class $\widetilde{F}_{RKv}$ the preceding display implies that

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{F}_{RKv}}\Big] \lesssim C_0K^{3/2}(v+v^{1/2})\left(1+C_0K^{3/2}\frac{1+v^{1/2}}{\sqrt{n}v^{3/2}}\right). \tag{10.10}$$

Now with $K = K_1 \log n$, we have that

$$\begin{aligned}&\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}}) \geq \epsilon\right\}\\ &\leq \mathbb{P}\left\{\sup_{\substack{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0),\psi\in\mathcal{M}_{RK},\\ d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\geq\epsilon}} \left\{\int_{\mathcal{X}\times\mathbb{R}}\Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y)-f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d(\mathbb{P}_n-P_0)(\boldsymbol{x},y) - d^2_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\right\} \geq 0,\right.\\ &\qquad\qquad \left.\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| \leq K\right\} + \nu/2\end{aligned}$$

$$\begin{aligned}&\leq \sum_{s=0}^{\infty}\mathbb{P}\left\{\sup_{\substack{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0),\psi\in\mathcal{M}_{RK},\\ 2^s\epsilon\leq d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\leq 2^{s+1}\epsilon}} \sqrt{n}\int_{\mathcal{X}\times\mathbb{R}}\Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y)-f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d(\mathbb{P}_n-P_0)(\boldsymbol{x},y) \geq \sqrt{n}2^{2s}\epsilon^2,\right\} + \nu/2\\ &\leq \sum_{s=0}^{\infty}\mathbb{P}\left\{\sup_{h\in\mathcal{F}_{RK2^{s+1}\epsilon}}\sqrt{n}\int_{\mathcal{X}\times\mathbb{R}} h(\boldsymbol{x},y)d(\mathbb{P}_n-P_0)(\boldsymbol{x},y) \geq \sqrt{n}2^{2s}\epsilon^2\right\} + \nu/2\end{aligned} \tag{10.11}$$

where, we now show that there exists a constant $C > 0$ such that with $\epsilon = M(\log n)n^{-1/3}$,

$$\mathbb{E}\left\{\sup_{h\in\mathcal{F}_{RK2^{s+1}\epsilon}}\sqrt{n}\left|\int_{\mathcal{X}\times\mathbb{R}} h(\boldsymbol{x},y)d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\right|\right\} \leq CM^{1/2}(\log n)^2 n^{-1/6}2^{(s+1)/2} \tag{10.12}$$

An application of Markov's inequality, together with (10.11), then yields, with $\epsilon = M(\log n)n^{-1/3}$

$$\begin{aligned}\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}}) \geq \epsilon\right\} &\leq \sum_{s=0}^{\infty}\frac{CM^{1/2}(\log n)^2n^{-1/6}2^{(s+1)/2}}{\sqrt{n}2^{2s}\epsilon^2} + \nu/2\\ &= \sum_{s=0}^{\infty}\frac{C(\log n)^2n^{-1/6}2^{(s+1)/2}}{\sqrt{n}2^{2s}M^{3/2}(\log n)^2n^{-2/3}} + \nu/2 = \frac{2^{1/2}C}{M^{3/2}}\sum_{s=0}^{\infty}\frac{1}{2^{3s/2}} + \nu/2 \leq \nu,\end{aligned}$$

for M sufficiently large. The result of Proposition 3.2 hence follows by showing (10.12). Using the obtained bound in (10.10) with $v = 2^{s+1}\epsilon$ and using that $2^{s+1} \geq 1, s \geq 0$ and $\epsilon \leq 1$ for n large enough we get some

some constant $D_0 > 0$

$$\begin{aligned}
&\mathbb{E}\left\{\sup_{h\in\mathcal{F}_{RK2^{s+1}\epsilon}}\sqrt{n}\left|\int_{\mathcal{X}\times\mathbb{R}}h(\boldsymbol{x},y)d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\right|\right\}\\
&\lesssim (\log n)^{3/2}M^{1/2}2^{(s+1)/2}(\log n)^{1/2}n^{-1/6}\left(1+D_0(\log n)^{3/2}\frac{1+M^{1/2}2^{(s+1)/2}(\log n)^{1/2}n^{-1/6}}{\sqrt{n}M^{3/2}2^{3(s+1)/2}(\log n)^{3/2}n^{-1/2}}\right)\\
&= (\log n)^2 n^{-1/6}M^{1/2}2^{(s+1)/2}\left(1+D_0'\frac{1+M^{1/2}2^{(s+1)/2}(\log n)^{1/2}n^{-1/6}}{2^{3(s+1)/2}}\right), \quad \text{with } D_0' = D_0M^{-3/2}\\
&\le 2(\log n)^2 n^{-1/6}M^{1/2}2^{(s+1)/2}, \quad \text{for } s\ge 0 \text{ and } n \text{ large enough.}
\end{aligned}$$

This proves the desired result:

$$\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d^2_{\boldsymbol{\alpha}}(\hat{\psi}_{n\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}}) = \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\int\left\{\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})-\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 dG(\boldsymbol{x}) = O_p\left((\log n)^2 n^{-2/3}\right).$$

□

10.2. Appendix B: Asymptotic behavior of the simple score estimator

In this section we prove Theorem 4.1 given in Section 4. The proof is decomposed into three parts: In Section 10.2.1 we first prove the existence of a crossing of zero of ϕ_n defined in (4.5). The proof of consistency and asymptotic normality of $\hat{\boldsymbol{\alpha}}_n$ are given in Section 10.2.2 and Section 10.2.3.

10.2.1. Proof of existence of a crossing of zero

Let ϕ be the population version of ϕ_n defined by

$$\phi(\boldsymbol{\beta}) := \int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{x}\left\{y-\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}dP_0(\boldsymbol{x},y), \tag{10.13}$$

where $\psi_{\boldsymbol{\alpha}}$ is defined by

$$\psi_{\boldsymbol{\alpha}}(u) := \mathbb{E}\left[\psi_0\left(\boldsymbol{\alpha}^T\boldsymbol{X}\right)|\boldsymbol{\alpha}^T\boldsymbol{X}=u\right] \equiv \mathbb{E}\left[\psi_0\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}=u\right].$$

We have the following result:

Proposition 10.1.

$$\phi_n(\boldsymbol{\beta}) = \phi(\boldsymbol{\beta}) + o_p(1),$$

uniformly in $\boldsymbol{\beta}\in\mathcal{C} := \left\{\boldsymbol{\beta}\in\mathbb{R}^{d-1} : \mathbb{S}(\boldsymbol{\beta})\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)\right\}$.

Proof. For any $\boldsymbol{\beta}\in\mathcal{C}$, we write,

$$\begin{aligned}
\phi_n(\boldsymbol{\beta}) &= \int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{x}\left\{y-\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}d\mathbb{P}_n(\boldsymbol{x},y)\\
&\quad+\int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{x}\left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)-\hat{\psi}_{n\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}d\mathbb{P}_n(\boldsymbol{x},y)\\
&= \phi(\boldsymbol{\beta})+\int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{x}\left\{y-\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\\
&\quad+\int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{x}\left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)-\hat{\psi}_{n\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\\
&\quad+\int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{x}\left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)-\hat{\psi}_{n\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}dP_0(\boldsymbol{x},y)\\
&= \phi(\boldsymbol{\beta})+I+II+III.
\end{aligned}\tag{10.14}$$

To find the rate of convergence of the term I in (10.14) we will use Lemma 2.5 in Supplement B of [2]. Note first that for $1 \leq i \leq d-1$ the i-th component of the vector $\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x}$ can be written as $s(\boldsymbol{\beta})_{i1}x_1 + \ldots + s(\boldsymbol{\beta})_{id}x_d$, where by Assumption A8 the functions s_{ij} are assumed to be uniformly bounded with partial derivatives that are also uniformly bounded on the bounded convex set $\mathcal{C}$ to which $\boldsymbol{\beta}$ belongs. If B_1 is the same constant found in Lemma 2.5 in Supplement B of [2] then the ϵ-bracketing entropy is bounded above by $B_1 K_0/\epsilon$. Applying Lemma 3.4.3 of [19], Markov's inequality and Lemma 2.5 in Supplement B to each of the empirical processes corresponding to the term $s(\boldsymbol{\beta})_{ij}x_j$ for $1 \leq j \leq d$ yields (with $D = 8MRK_0$, and M is a constant bounding the sum of $s(\boldsymbol{\beta})_{ij}$ and their partial derivatives) for $A > 0$

$$
\begin{aligned}
P\Big(|I| \geq An^{-1/2}\Big) &\leq \frac{D}{A} J_n(B_2)\left(1 + \frac{J_n(B_2)}{\sqrt{n}B_2^2}\right), \text{ where } J_n \text{ is defined in (10.3) and } B_2 \text{ is the same} \\
&\qquad\qquad \text{constant of Lemma 2.5 in Supplement B} \\
&\leq \frac{D}{A} B_3 \left(1 + \frac{B_3}{\sqrt{n}B_2^2}\right), \quad \text{using the inequality in (10.4) with } B_3 = B_2 + 2B_1^{1/2}K_0^{1/2}B_2^{1/2} \\
&\asymp \frac{1}{A}.
\end{aligned}
$$

This implies that $I = O_p(n^{-1/2})$. For the last term III in (10.14) we get by an application of the Cauchy-Schwarz inequality and by Proposition 3.2 that this term is $O_p(n^{-1/3}\log n)$, i.e.

$$
\begin{aligned}
III &\leq \left(\int \left\|\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x}\right\|_2^2 dG(\boldsymbol{x})\right)^{1/2} \left(\int \left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) - \hat{\psi}_{n\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}^2 dG(\boldsymbol{x})\right)^{1/2} \\
&= O_p(n^{-1/3}\log n),
\end{aligned}
$$

where we also use that $\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x}$ is bounded in L_2 norm, a straightforward implication of Assumption A1 (boundedness of $\mathcal{X}$ and Assumption A8 (uniform boundedness of the components of the matrix $\boldsymbol{J}_{\mathbb{S}}$). The result now follows by showing that the term II is $o_p(1)$. Consider the class of functions

$$
\begin{aligned}
\mathcal{G}_{jRKv} \quad = \quad \Big\{ & g(\boldsymbol{x}, y) = s(\boldsymbol{\beta})x_j \left\{\psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T\boldsymbol{x}\right) - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right\}, \\
& \qquad \text{such that } (\boldsymbol{\alpha}, \boldsymbol{\beta}, \psi) \in \mathcal{S}_{d-1} \times \mathcal{C} \times \mathcal{M}_{RK} \text{ and } (\boldsymbol{x}, y) \in \mathcal{X} \times \in \mathbb{R}, \\
& \qquad \text{and} \sup_{\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)} d_{\boldsymbol{\alpha}}(\psi_{\boldsymbol{\alpha}}, \psi) \leq v \Big\}
\end{aligned}
$$

with s a function satisfying (4.6). Then, $\mathcal{G}_{jRKv} \subset \mathcal{Q}_{jRK} - \mathcal{Q}_{jRK}$, where $\mathcal{Q}_{jRC}$ is the same class defined in (2.5). Here, we choose K large enough such that $K \geq K_0$. If follows from (2.7) in the proof of Lemma 2.5 in Supplement B, that (at the cost of increasing the constant L in (2.7))

$$
H_B\Big(\epsilon, \tilde{\mathcal{G}}_{jRKv}, \|\cdot\|_{P_0}\Big) \leq \frac{LK}{\epsilon},
$$

where $\tilde{\mathcal{G}}_{jRKv} = (16M_0K)^{-1}\mathcal{G}_{jRKv}$. Also, we have for all $g \in \mathcal{G}_{jRKv}$

$$
\|g\|_{P_0} \leq MRv.
$$

Fix $\nu > 0$ and let s_{ij} be the $i \times j$ entry of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ for $1 \leq i \leq d-1$ and $1 \leq j \leq d$. Also, let

$$
II_{ij} = \int_{\mathcal{X}\times\mathbb{R}} s_{ij}(\boldsymbol{\beta})x_j \left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) - \hat{\psi}_{n\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y).
$$

Using Lemma 10.1 and Proposition 3.2 there exists some constant $K_1 > 0$ large enough (and independent

of n) such that with $K = K_1 \log n$ and $v = K_1 \log n\ n^{-1/3}$ we have that for $A > 0$

$$
\begin{aligned}
& P\left(|II_{ij}| \geq An^{-1/2}\right) \\
& = P\left(|II_{ij}| \geq An^{-1/2}, \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| \leq K, \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\psi_{\boldsymbol{\alpha}},\psi) \leq v\right) + \nu/2 \\
& \lesssim \frac{K}{A} J_n(MRv)\left(1 + \frac{J_n(MRv)}{\sqrt{n}M^2R^2v^2}\right) + \nu/2, \quad \text{where } J_n \text{ is defined in (10.3)} \\
& \leq \frac{K}{A}\left(MRv + 2(MRL)^{1/2}K^{1/2}v^{1/2}\right)\left(1 + \frac{MRv + 2(MRL)^{1/2}K^{1/2}v^{1/2}}{\sqrt{n}M^2R^2v^2}\right) + \nu/2, \quad \text{using the inequality in (10.4)} \\
& \leq \frac{\tilde{M}}{A}(\log n)^2\ n^{-1/6}\left(1 + \frac{1}{\log n M^2R^2}\right) + \nu/2, \quad \text{for some constant } \tilde{M} > 0 \\
& \leq \nu
\end{aligned}
$$

for n large enough. We conclude that $II_{ij} = o_p(n^{-1/2})$ which in turn implies that $II = o_p(n^{-1/2})$.

□

Proof of Theorem 4.1 (Existence). Using Proposition 10.1 we get, analogously to the development in [7], the relation

$$\phi_n(\boldsymbol{\alpha}) = \phi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + R_n(\boldsymbol{\beta}), \tag{10.15}$$

where $R_n(\boldsymbol{\alpha}) = o_p(1) + o(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, uniformly in $\boldsymbol{\beta} \in \mathcal{C}$ and where ϕ' is the derivative of ϕ defined in (10.13). Using Lemma 3.1 in Supplement C of [2], we get that the derivative of ϕ at $\boldsymbol{\beta}_0$ is given by the matrix

$$\phi'(\boldsymbol{\beta}_0) = (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\, \mathbb{E}\left[\psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\, \mathrm{Cov}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\right] \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0) = (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\, \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0) = \boldsymbol{B},$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined in (4.7) and (4.10) respectively. We now define, for $h > 0$, the functions

$$\tilde{R}_{n,h}(\boldsymbol{\beta}) = \frac{1}{h^{d-1}} \int K_h(u_1 - \beta_{01}) \dots K_h(u_{d-1} - \beta_{d-1})\, R_n(u_1, \dots, u_{d-1})\, du_1 \dots\, du_{d-1},$$

where

$$K_h(\boldsymbol{x}) = h^{-1}K(x/h), \qquad x \in \mathbb{R},$$

letting K be one of the usual smooth kernels with support $[-1, 1]$.

Furthermore, we define:

$$\tilde{\phi}_{n,h}(\boldsymbol{\beta}) = \phi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \tilde{R}_{nh}(\boldsymbol{\beta}).$$

Clearly:

$$\lim_{h\downarrow 0} \tilde{\phi}_{n,h}(\boldsymbol{\beta}) = \phi_n(\boldsymbol{\beta}) \qquad \text{and} \qquad \lim_{h\downarrow 0} \tilde{R}_{nh}(\boldsymbol{\beta}) = R_n(\boldsymbol{\beta}),$$

for each continuity point $\boldsymbol{\beta}$ of ϕ_n.

We now reparametrize, defining

$$\boldsymbol{\gamma} = \phi'(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \qquad \boldsymbol{\gamma}_0 = \phi'(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0.$$

This gives:

$$\phi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \tilde{R}_{nh}(\boldsymbol{\beta}) = \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 + \tilde{R}_{nh}\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}\right),$$

By (10.15), the mapping

$$\boldsymbol{\gamma} \mapsto \boldsymbol{\gamma}_0 - R_n\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}\right),$$

maps, for each $\eta > 0$, the ball $B_\eta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\} \leq \eta\}$ into $B_{\eta/2}(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\} \leq \eta/2\}$ for all large n, with probability tending to one, where $\|\cdot\|$ denotes the Euclidean norm, implying that the *continuous* map

$$\boldsymbol{\gamma} \mapsto \boldsymbol{\gamma}_0 - \tilde{R}_{nh}\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}\right),$$

maps $B_\eta(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_2 \leq \eta\}$ into itself for all large n and small h. So for large n and small h there is, by Brouwer's fixed point theorem, a point $\boldsymbol{\gamma}_{nh}$ such that

$$\boldsymbol{\gamma}_{nh} = \boldsymbol{\gamma}_0 - \tilde{R}_{nh}\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}_{nh}\right).$$

Defining $\boldsymbol{\beta}_{nh} = \boldsymbol{B}^{-1}\boldsymbol{\gamma}_{nh}$, we get:

$$\tilde{\phi}_{n,h}(\boldsymbol{\beta}_{nh}) = \phi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_{nh} - \boldsymbol{\beta}_0) + \tilde{R}_{nh}(\boldsymbol{\beta}_{nh}) = \mathbf{0}. \tag{10.16}$$

By compactness, $(\boldsymbol{\beta}_{n,1/k})_{k=1}^{\infty}$ must have a subsequence $(\boldsymbol{\beta}_{n,1/k_i})$ with a limit $\tilde{\boldsymbol{\beta}}_n$, as $i \to \infty$.

Suppose that the jth component ϕ_{nj} of ϕ_n does not have a crossing of zero at $\tilde{\boldsymbol{\beta}}_n$. Since ϕ_{nj} only has finitely many jump discontinuities, since there can only be discontinuities at a changing of ordering of the values $\boldsymbol{\alpha}^T\boldsymbol{X}_i$, there must be a closed ball $B_\delta(\tilde{\boldsymbol{\beta}}_n) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_n\| \leq \delta\}$ of $\tilde{\boldsymbol{\beta}}_n$ such that $\{\bar{\phi}_{nj}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in B_\delta(\tilde{\boldsymbol{\beta}}_n)\}$ has a constant sign in the closed ball B_δ, say $\bar{\phi}_{nj}(\boldsymbol{\beta}) > 0$ for $\boldsymbol{\beta} \in \bar{B}_\delta(\tilde{\boldsymbol{\beta}}_n)$. Again using that ϕ_{nj} only has finitely many jump discontinuities, this means that

$$\bar{\phi}_{n,j}(\boldsymbol{\beta}) \geq c > 0, \qquad \text{for all } \boldsymbol{\beta} \in \bar{B}_\delta(\tilde{\boldsymbol{\beta}}_n).$$

This means that the jth component $\tilde{\phi}_{n,h,j}$ of $\tilde{\phi}_{n,h}$ satisfies

$$\begin{aligned}
&\tilde{\phi}_{n,h,j}(\boldsymbol{\beta}) = \left[\phi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]_j + \tilde{R}_{nh,j}(\boldsymbol{\beta}) \\
&= \frac{1}{h^{d-1}} \int \left\{\left[\phi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]_j + R_{nj}(u_1, \ldots, u_{d-1})\right\} K_h(u_1 - \beta_{01}) \ldots K_h(u_{d-1} - \beta_{d-1})\, du_1 \ldots\, du_{d-1} \\
&\geq \frac{1}{h^{d-1}} \int \left\{\left[\phi'(\boldsymbol{\beta}_0)(\boldsymbol{u} - \boldsymbol{\beta}_0)\right]_j + R_{nj}(u_1, \ldots, u_{d-1})\right\} K_h(u_1 - \beta_1) \ldots K_h(u_{d-1} - \beta_{d-1})\, du_1 \ldots\, du_{d-1} - c/2 \\
&\geq c\, \frac{1}{h^{d-1}} \int K_h(u_1 - \beta_1) \ldots K_h(u_d - \beta_d)\, du_1 \ldots\, du_d - c/2 \\
&= c/2,
\end{aligned}$$

for $\boldsymbol{\beta} \in B_{\delta/2}(\tilde{\boldsymbol{\beta}}_n)$ and sufficiently small h, contradicting (10.16), since $\boldsymbol{\beta}_{nh}$, for $h = 1/k_i$, belongs to $B_{\delta/2}(\tilde{\boldsymbol{\beta}}_n)$ for large k_i.

□

10.2.2. Proof of consistency of $\hat{\boldsymbol{\alpha}}_n$

Proof. Since $\hat{\boldsymbol{\beta}}_n$ is contained in the compact set $\mathcal{C}$, the sequence $(\hat{\boldsymbol{\beta}}_n)$ has a subsequence $(\hat{\boldsymbol{\beta}}_{n_k} = \hat{\boldsymbol{\beta}}_{n_k}(\omega))$, converging to an element $\boldsymbol{\beta}_*$. Let $\boldsymbol{\alpha}_{n_k} = \mathbb{S}(\hat{\boldsymbol{\beta}}_{n_k})$. If $\hat{\boldsymbol{\beta}}_{n_k} = \hat{\boldsymbol{\beta}}_{n_k}(\omega) \longrightarrow \boldsymbol{\beta}_*$, we get by continuity of the map $\mathbb{S}$ that $\boldsymbol{\alpha}_{n_k} \to \boldsymbol{\alpha}_* = \mathbb{S}(\boldsymbol{\beta}_*)$. By Proposition 3.2, we also have

$$\hat{\psi}_{n_k,\hat{\boldsymbol{\alpha}}_{n_k}}\left(\mathbb{S}(\boldsymbol{\beta}_{n_k})^T\boldsymbol{x}\right) \longrightarrow \psi_{\boldsymbol{\alpha}_*}(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}),$$

where $\psi_{\boldsymbol{\alpha}}$ is defined in (3.1). By Proposition 10.1 and the fact that in the limit, the crossing of zero becomes a root of the continuous limiting function, we get,

$$\lim_{k\to\infty} \phi_{n_k}(\boldsymbol{\beta}_{n_k}) = \phi(\boldsymbol{\beta}_*) = \mathbf{0} \tag{10.17}$$

where,

$$\begin{aligned}\phi(\boldsymbol{\beta}_*) &= \int \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\left\{y-\psi_{\boldsymbol{\alpha}_*}\left(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\right)\right\}\,dP_0(\boldsymbol{x},y)\\ &= \int \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\left\{\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)-\psi_{\boldsymbol{\alpha}_*}\left(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\right)\right\}\,dG(\boldsymbol{x})\\ &= \int \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\left[\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)-\mathbb{E}\left\{\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)|\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X}=\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\right\}\right]\,dG(\boldsymbol{x})\\ &= \mathbb{E}\left[\mathrm{Cov}\left[\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})^T\boldsymbol{X},\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\right)|\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X}\right]\right] \qquad (10.18)\end{aligned}$$

We next conclude that,

$$\begin{aligned}0 &= (\boldsymbol{\beta}_0-\boldsymbol{\beta}_*)^T\phi(\boldsymbol{\beta}_*)\\ &= \mathbb{E}\left[\mathrm{Cov}\left[(\boldsymbol{\beta}_0-\boldsymbol{\beta}_*)^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{X},\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X}+(\mathbb{S}(\boldsymbol{\beta}_0)-\mathbb{S}(\boldsymbol{\beta}_*))^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X}\right]\right]\end{aligned}$$

which can only happen if $\boldsymbol{\beta}_0=\boldsymbol{\beta}_*$ where we use the positivity of the random variable $\mathrm{Cov}((\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^T\boldsymbol{X}, \psi_0(\boldsymbol{\alpha}^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X})$ shown in Lemma 3.2 in Supplement C of [2] and Assumption A6 which guarantees that the random variable $\mathrm{Cov}[(\boldsymbol{\beta}_0-\boldsymbol{\beta}_*)^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})^T\boldsymbol{X},\psi_0(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}+(\mathbb{S}(\boldsymbol{\beta}_0)-\mathbb{S}(\boldsymbol{\beta}))^T\boldsymbol{X})|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}]$ is not equal to 0 almost surely for all $\boldsymbol{\beta}\neq\boldsymbol{\beta}_0$. Note that,

$$\begin{aligned}&\mathrm{Cov}\left[(\boldsymbol{\beta}_0-\boldsymbol{\beta})^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})^T\boldsymbol{X},\psi_0\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}+(\mathbb{S}(\boldsymbol{\beta}_0)-\mathbb{S}(\boldsymbol{\beta}))^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}=u\right]\\ &= \mathrm{Cov}\left[(\mathbb{S}(\boldsymbol{\beta}_0)-\mathbb{S}(\boldsymbol{\beta})+o(\boldsymbol{\beta}-\boldsymbol{\beta}_0))^T\boldsymbol{X},\psi_0\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}+(\mathbb{S}(\boldsymbol{\beta}_0)-\mathbb{S}(\boldsymbol{\beta}))^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}=u\right]\\ &= \mathrm{Cov}\left[(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^T\boldsymbol{X},\psi_0(\boldsymbol{\alpha}^T\boldsymbol{X}+(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X}=u\right]+o(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\end{aligned}$$

where the first term in the expression above is positive for all $\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)$ by Lemma 3.2 in Supplement C.

□

10.2.3. Proof of asymptotic normality of $\hat{\boldsymbol{\alpha}}_n$

We define ϕ_n at $\hat{\boldsymbol{\beta}}_n$ by putting

$$\phi_n(\hat{\boldsymbol{\beta}}_n)=\mathbf{0}. \qquad (10.19)$$

Note that, with this definition, we use the representation of the components as a convex combination of the left and right limit at $\hat{\boldsymbol{\beta}}_n$:

$$\phi_{n,j}(\hat{\boldsymbol{\beta}}_n)=\gamma_j\phi_{n,j}(\hat{\boldsymbol{\beta}}_n-)+(1-\gamma_j)\phi_{n,j}(\hat{\boldsymbol{\beta}}_n+)=0, \qquad (10.20)$$

where $\phi_{n,j}$ denotes the jth component of ϕ_n and where we can choose $\gamma_j\in[0,1]$ in such a way that (10.20) holds since we have a crossing of zero componentwise. Note that this does not change the location of the crossing of zero. Since the following asymptotic representations are also valid for one-sided limits as used in (10.20) we can use Definition (10.19) and assume $\phi_n(\hat{\boldsymbol{\beta}}_n)=\mathbf{0}$. We show

$$\begin{aligned}\phi_n(\hat{\boldsymbol{\beta}}_n) &= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\int\left\{\boldsymbol{x}-\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y-\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}\,d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\\ &+\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\int\left\{\boldsymbol{x}-\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y-\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\boldsymbol{\beta}_n)^T\boldsymbol{x}\right)\right\}\,dP_0(\boldsymbol{x},y)\\ &+o_p\left(n^{-1/2}+\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|\right), \qquad (10.21)\end{aligned}$$

where from now on we will use the notation $\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x})$ to denote $\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}=\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x})$ for all $\boldsymbol{\beta}\in\mathcal{C}$ and $\boldsymbol{x}\in\mathcal{X}$.

Since $\hat{\boldsymbol{\beta}}_n \to_p \boldsymbol{\beta}_0$ and since the function $\boldsymbol{\beta} \to \psi_{\mathbb{S}(\boldsymbol{\beta})}(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x})$ has derivative $\psi_0'(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\big(\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\big)$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ for all $\boldsymbol{x} \in \mathcal{X}$ (See Lemma 3.1 in Supplement C), we get by Definition (10.19) and a Taylor expansion at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ that,

$$\begin{aligned}&\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\} d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, y)\\ &= \boldsymbol{B}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) + o_p\left(n^{-1/2} + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right),\end{aligned} \tag{10.22}$$

where $\boldsymbol{B}$ is defined in (4.10). We conclude that,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_d N_d\left(\mathbf{0}, \boldsymbol{\Pi}\right),$$

where $\boldsymbol{\Pi}$ is defined in (4.13). The asymptotic normality of the estimator $\hat{\boldsymbol{\alpha}}_n$ then follows by noting that,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) = J_{\mathbb{S}}(\boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right) \to_d N_d\left(\mathbf{0}, J_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{\Pi}\left(J_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\right).$$

To prove (10.21) we first define the piecewise constant function $\bar{E}_{n,\boldsymbol{\beta}}$

$$\bar{E}_{n,\boldsymbol{\beta}}(u) = \begin{cases} \mathbb{E}\left[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = \tau_{i,\boldsymbol{\beta}}\right] & \text{if } \psi_{\boldsymbol{\alpha}}(u) > \hat{\psi}_{n\boldsymbol{\alpha}}(\tau_i) \text{ for all } u \in (\tau_i, \tau_{i+1}),\\ \mathbb{E}\left[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = s\right] & \text{if } \psi_{\boldsymbol{\alpha}}(s) = \hat{\psi}_{n\boldsymbol{\alpha}}(s) \text{ for some } s \in (\tau_i, \tau_{i+1}),\\ \mathbb{E}\left[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = \tau_{i+1,\boldsymbol{\beta}}\right] & \text{if } \psi_{\boldsymbol{\alpha}}(u) < \hat{\psi}_{n\boldsymbol{\alpha}}(\tau_i) \text{ for all } u \in (\tau_i, \tau_{i+1}),\end{cases}$$

where the $\tau_{i,\boldsymbol{\beta}}$ denote the sequence of jump points of the monotone LSE $\hat{\psi}_{n\boldsymbol{\alpha}} = \hat{\psi}_{n\mathbb{S}(\boldsymbol{\beta})}$. We then have

$$\int \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\left\{y - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) = \mathbf{0}. \tag{10.23}$$

This follows from the fact that $\hat{\psi}_{n\boldsymbol{\alpha}}$, i.e. the minimizer of the quadratic criterion $\int_{\mathcal{X}\times\mathbb{R}}\left(y - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 d\mathbb{P}_n(\boldsymbol{x}, y)$ over monotone functions $\psi \in \mathcal{M}$, is the left derivative of the greatest convex minorant of the cumulative sum diagram $\{(0,0), \left(i, \sum_{j=1}^i Y_i^\alpha\right), i = 1, \ldots, n\}$. (See also [8], p.332). By Lemma 3.6 in Supplement C of [2] we also know that $\psi'_{\boldsymbol{\alpha}}$ stays away from zero for all $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$ in a neighborhood of $\boldsymbol{\alpha}_0 = \mathbb{S}(\boldsymbol{\beta}_0)$. Using the same techniques as in [8], we can find a constant $C > 0$ such that for all $i = 1, \ldots, d$ and $u \in \mathcal{I}_{\boldsymbol{\alpha}}$,

$$\left|\mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right) - \bar{E}_{ni,\boldsymbol{\beta}}(u)\right| \le C\left|\psi_{\boldsymbol{\alpha}}(u) - \hat{\psi}_{n\boldsymbol{\alpha}}(u)\right| \tag{10.24}$$

where $\bar{E}_{ni,\boldsymbol{\beta}}$ denotes the ith component of $E_{n,\boldsymbol{\beta}}$. In the sequel, we will use $\boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n) = O_p(1)$, an immediate consequence of consistency of $\hat{\boldsymbol{\alpha}}_n$ and Assumption A8. Now, as a consequence of (10.23), we can write

$$\begin{aligned}\phi_n(\hat{\boldsymbol{\beta}}_n) &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y)\\ &\quad + \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y)\\ &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T(I + II).\end{aligned} \tag{10.25}$$

The term II can be written as

$$\begin{aligned}II &= \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\\ &\quad + \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y)\\ &\quad + \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y)\\ &= II_a + II_b + II_c.\end{aligned} \tag{10.26}$$

We first note that by Lemma 3.4 in Supplement C, the functions $u \mapsto \mathbb{E}(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u)$ are uniformly bounded by R for all $\boldsymbol{\beta} \in \mathcal{C}$ and $i \in \{1, \ldots, d\}$. Also, they admit a bounded variation, with a total variation that is uniformly bounded for all $\boldsymbol{\beta} \in \mathcal{C}$ and $i \in \{1, \ldots, d\}$. By definition of $\bar{E}_{n,\boldsymbol{\beta}}$ its i-th component, $\bar{E}_{ni,\boldsymbol{\beta}}$ is also uniformly bounded by R and has a finite total variation which cannot exceed the total variation of $u \mapsto \mathbb{E}(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u)$. Using Lemma 3.5 of Supplement C , we can find two monotone functions f_1 and f_2 such that $u \mapsto \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right) - \bar{E}_{n,\boldsymbol{\beta}}(u) = f_2(u) - f_1(u)$ with $f_1, f_2 \in \mathcal{M}_{RC_1}$ for some constant $C_1 > 0$. Also, we know that $\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n} \in \mathcal{M}_{RK}$ with $K = K_1 \log n$ with increasing probability as $n \to \infty$ provided that $K_1 > 0$ is chosen large enough. Noting that for any bounded increasing functions f_1, f_2, f_3 we have that $(f_2 - f_1)f_3$ is again bounded and has a bounded variation, it follows that the class of functions, $\mathcal{F}_a$ say, involved in term II_a is included in $\mathcal{H}_{RK'v}$ defined in Lemma 2.4 in Supplement B . Here, the constant $K' = K_2 \log n$ for some large enough constant $K_2 > 0$, and $v = C_2(\log n)^2 n^{-1/3}$ for some constant $C_2 > 0$ using (10.24) and Proposition 3.2. Using Lemma 2.4 in Supplement B, we can show that (when the event $\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n} \in \mathcal{M}_{RK}$ occurs)

$$H_B\left(\epsilon, \widetilde{\mathcal{F}}_a, \|\cdot\|_{B,P_0}\right) \leq \frac{B_1}{\epsilon}, \quad \text{for some constant } B_1 > 0,$$

with $\widetilde{\mathcal{F}}_a = \tilde{D}^{-1}\mathcal{F}_a$ with $\tilde{D} \asymp K' = K_2 \log n$. Also, for any element $\tilde{f} = \tilde{D}^{-1}f \in \widetilde{\mathcal{F}}$ we have that

$$\|\tilde{f}\|_{B,P_0} \leq B_2\tilde{D}^{-1}v = C_2(\log n)n^{-1/3} = \delta_n, \quad \text{for some constant } C_2 > 0.$$

Let $II_{a,i}$ be the term corresponding to i-th component of $\boldsymbol{X}$. Using Markov's inequality we have for a fixed $A > 0$, $\nu > 0$ and n large enough that

$$\begin{aligned}
P\left(|II_{a,i}| \geq An^{-1/2}\right) &= P\left(|II_{a,i}| \geq An^{-1/2}, \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| \leq K\right) + \nu/2 \\
&\lesssim \frac{\tilde{D}}{A} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \nu/2, \quad \text{where } J_n \text{ is defined in (10.3)} \\
&\lesssim \frac{\log n}{A} B_2 \delta_n^{1/2}\left(1 + \frac{B_2}{\sqrt{n}\delta_n^{3/2}}\right) + \nu/2, \quad \text{for some constant } B_2 > 0, \\
&\qquad \text{using the inequality in (10.4) and taking } n \text{ large enough} \\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_3}{(\log n)^{3/2}}\right) + \nu/2, \quad \text{with } B_3 = B_2 C_2^{-3/2} \\
&\leq \nu
\end{aligned}$$

for$n$ large enough. We conclude that $II_{a,i} = o_p(n^{-1/2})$ which in turn implies that

$$II_a = o_p(n^{-1/2}).$$

We turn now to II_b. Using Lemma 3.1 in Supplement C and a Taylor expansion of $\boldsymbol{\beta} \mapsto \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)$ we get,

$$\begin{aligned}
\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) = \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\left[\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\left(\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right)\psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right] \\
+ o(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \qquad (10.27)
\end{aligned}$$

so that

$$\begin{aligned}
II_b &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \\
&= o_p\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right)
\end{aligned}$$

using consistency of $\hat{\boldsymbol{\beta}}_n$. We next consider the term II_c. Using uniform boundedness of $\boldsymbol{J}_{\mathbb{S}}$ on $\mathcal{C}$ and the inequality in (10.24) it follows that

$$\begin{aligned}
\|II_c\| &\lesssim \int \left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}\right)\right\}^2 dG(\boldsymbol{x}) \\
&= O_p((\log n)^2 n^{-2/3}) = o_p(n^{-1/2})
\end{aligned}$$

uniformly in $\beta \in \mathcal{C}$. We conclude that (10.25) can be written as

$$\begin{aligned}
\phi_n(\hat{\boldsymbol{\beta}}_n) &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\
&\quad + o_p\left(n^{-1/2} + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right) \\
&= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\
&\quad + o_p\left(n^{-1/2} + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right) \\
&= I_a + I_b + o_p\left(n^{-1/2} + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right). \qquad (10.28)
\end{aligned}$$

We show below that $I_b = o_p\left(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right)$ such that the limiting distribution of the score estimator follows from the analysis of the term I_a which can be rewritten as

$$\begin{aligned}
I_a &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \qquad (10.29)
\end{aligned}$$

where we recall that $\psi_{\boldsymbol{\alpha}}(u) = \mathbb{E}\left(\psi_0(\boldsymbol{\alpha}^T\boldsymbol{X}|\boldsymbol{\alpha}^T X = u\right)$. For the second term on the right-hand side of (10.29) we have by (10.27)

$$\begin{aligned}
&\boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \\
&= -\Bigg\{\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}\left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}^T dP_0(\boldsymbol{x}, y) \\
&\qquad\qquad \times \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\Bigg\}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&\quad + o_p(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0). \qquad (10.30)
\end{aligned}$$

For the first term on the right-hand side of (10.29) we have that

$$\begin{aligned}
&\boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
&= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) + o_p(n^{-1/2}) + o_p(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0). \qquad (10.31)
\end{aligned}$$

Indeed, since this amounts to showing that

$$\begin{aligned}
A &= \left(J_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n) - J_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
&= o_p(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0), \qquad (10.32) \\
B &= \int \left(\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right)(y - \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}))d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) = o_p(n^{-1/2}) \qquad (10.33)
\end{aligned}$$

and

$$C = \int \left(\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right)\left(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right)d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) = o_p(n^{-1/2}). \qquad (10.34)$$

We start by proving (10.32). Using again that $u \mapsto \mathbb{E}\left(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}=u\right)$ is a bounded function with a uniformly bounded total variation, and that x_i is a fixed (and deterministic) function, we can show that the class of functions involved in A, $\mathcal{F}_A$ say, satisfies $\mathcal{F}_A \subset x_i\mathcal{H}_{RC_1v} + \mathcal{H}_{RC_1v}$ with v and C_1 are some constants that are independent of n (since $\psi_{\boldsymbol{\alpha}}$, $\boldsymbol{X}$ and $u \mapsto E[\boldsymbol{X}|\boldsymbol{\alpha}^T X = u]$ are all bounded by constants independent of n). Now it follows by Lemma 2.4 in Supplement B that $H_B\big(\epsilon, \widetilde{\mathcal{H}}_{RC_1v}, \|\cdot\|_{B,P_0}\big) \lesssim 1/\epsilon$ with $\widetilde{\mathcal{H}}_{RC_1v} = (16M_0C_1)^{-1}\mathcal{H}_{RC_1v}$ and $\|\tilde{h}\|_{B,P_0} \lesssim C_2$ for some constant $C_2 > 0$ that is independent of n for all $\tilde{h} \in \widetilde{\mathcal{H}}_{RC_1v}$. Hence, using arguments similar to those of the proof of $II_a = o_p(n^{-1/2})$ we can show that

$$\int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) = O_p(n^{-1/2}).$$

Using a Taylor expansion of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$ gives the desired rate in (10.32).

Now we turn to term B in (10.33). Fix $\nu > 0$ and $i \in \{1, \ldots, d\}$. Using consistency of $\hat{\boldsymbol{\beta}}_n$ and Lemma 3.3 in Supplement C, then for all $\eta > 0$ there exists n large enough such that

$$\left|\mathbb{E}\left(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right| \leq \eta.$$

with probability at least $1 - \nu/2$. Thus, for $L > 0$ we have that for n large enough

$$\begin{aligned}
&P(|B_i| > Ln^{-1/2}) \\
&= P\left(\left|\int\left(\mathbb{E}\left(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right)(y - \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}))d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\right| > Ln^{-1/2}\right) \\
&\leq \nu/2 + \frac{1}{L}E[\|\mathbb{G}_n\|_{\mathcal{F}'}], \quad \text{where } \mathcal{F}' \text{ is defined in (2.10)} \\
&\leq \nu/2 + \frac{C_1}{L}\eta \quad \text{for some constant } C_1 > 0,
\end{aligned}$$

where B_i denotes the ith component of B defined in (10.33) and where we have used the result of Lemma 2.7. Choosing η such that $\eta \leq \nu L C_1^{-1}/2$ gives the claimed rate of convergence in (10.33).

To establish the convergence rate of C, we first note that, for $i \in \{1, \ldots, d\}$, we have that $\boldsymbol{x} \mapsto \mathbb{E}(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X} = \mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x})\psi_{\hat{\boldsymbol{\alpha}}_n}(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x})$ belongs to the class $\mathcal{G}_{RC_1} - \mathcal{G}_{RC_1}$ for some constant $C_1 > 0$ where $\mathcal{G}_{RK}$ was defined in (2.1). This follows from using again the fact that the function $u \mapsto E\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right)$ is uniformly bounded and has a uniform total variation for all $\beta \in \mathcal{C}$, that $\psi_{\boldsymbol{\alpha}}$ is a bounded monotone function, and the fact that $(f_1 - f_2)f_3$ is a bounded function with bounded total variation for any increasing and bounded functions f_1, f_2 and f_3, where we again use Lemma 3.5 in Supplement C to write the function $u \mapsto E\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right)$ as the difference $f_1 - f_2$. Note now that both $\boldsymbol{x} \mapsto x_i$ and $\boldsymbol{x} \mapsto \psi_0(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})$ are fixed and bounded functions, and that the order bracketing entropy of a class does not get altered after multiplication its members by such functions (similarly for addition). It follows from Lemma 2.2 and Lemma 2.3 in Supplement B of [2], that the ϵ-bracketing entropy of the class of functions involved in term C with respect to $\|\cdot\|_{P_0}$ is bounded above by B/ϵ for some constant B.

Furthermore, using consistency of $\hat{\boldsymbol{\alpha}}_n$ and Lemma 3.3 of Supplement C, we can find for any fixed $\nu > 0$ an $\eta > 0$ such that $\sup_{\boldsymbol{x}}|\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \psi_{\hat{\boldsymbol{\alpha}}_n}(\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})| \leq \eta$ with probability at least $1 - \nu/2$ for n large enough. Hence, at the cost of increasing the constant B, both the $\|\cdot\|_\infty$ and $\|\cdot\|_{P_0}$ norms of the functions of the class involved in term C are bounded above by $B\eta$. Using Markov's inequality and Lemma 3.4.2 of [19] it follows that for all $L > 0$

$$\begin{aligned}
&P(|C_i| > Ln^{-1/2}) \\
&= P\left(\left|\int\left(x_i - \mathbb{E}\left(\boldsymbol{X}_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right)\left(\psi_0(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}) - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right)d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\right| \geq Ln^{-1/2}\right) \\
&\leq \nu/2 + \frac{1}{L}J_n(B\eta)\left(1 + B\eta\frac{J_n(B\eta)}{\sqrt{n}B^2\eta^2}\right) \leq \nu/2 + \frac{1}{L}\left(B_1\eta^{1/2} + \frac{B_1}{B}\frac{1}{\sqrt{n}}\right) \leq \nu
\end{aligned}$$

taking η small enough and n large enough. We conclude that $C = o_p(n^{-1/2})$. Now we come back to term I_b given by

$$I_b = \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T\int\left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y).$$

Note first that

$$I_b = \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)$$
$$= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T I_b', \tag{10.35}$$

since

$$\begin{aligned}
&\int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y)\\
&= \mathbb{E}\left[\left(\boldsymbol{X} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)\left(\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{X} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right]\left(\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)\right]\\
&= \mathbf{0}.
\end{aligned}$$

Let $\mathcal{F}_b$ denote the class of functions involved in term I_b' defined in (10.35), where in the definition of this class we consider the event where $\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}$ is bounded. Given the arguments used recurrently above we can directly state that the ϵ-bracketing entropy of this class is no larger than $A_1 \log n/\epsilon$ for some constant $A_1 > 0$ with increasing probability. Also, the $\|\cdot\|_\infty$ and $\|\cdot\|_{P_0}$ norms of the members of the class $\mathcal{F}_b$ are respectively bounded above with increasing probability by $A_1 \log n$ and $A_1 \log n\, n^{-1/3} = \eta_n$ at the cost of taking a larger A_1. For a fixed $\nu > 0$ and $L > 0$ we have for $i \in \{1, \ldots, d\}$,using Lemma 3.4.2 of [19],

$$\begin{aligned}
&P\left(\left|\int \left\{\boldsymbol{x}_i - \mathbb{E}\left(\boldsymbol{X}_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\right| > Ln^{-1/2}\right)\\
&\leq \nu/2 + \frac{A_2}{L}(\log n)^{1/2}\eta_n^{1/2}\left(1 + \frac{A_2(\log n)^{1/2}\eta_n^{1/2}}{\sqrt{n}\eta_n^2}(\log n)\right), \quad \text{for some constant } A_2 > 0\\
&\leq \nu/2 + \frac{A_2}{L}(\log n)^{1/2}\eta_n^{1/2}\left(1 + \frac{A_2(\log n)^{3/2}}{\sqrt{n}\eta_n^{3/2}}\right), \quad \text{for some constant } A_2 > 0\\
&\lesssim \nu/2 + \frac{A_2}{L}(\log n)n^{-1/6}\left(1 + \frac{A_2}{A_1^{3/2}}\right) \leq \nu,
\end{aligned}$$

for n large enough. This implies that $I_b = o_p(n^{-1/2})$. We conclude by (10.30), (10.31), (10.32), (10.33), (10.34) and Definition (10.19) that,

$$\begin{aligned}
\boldsymbol{B}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = \int &\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\} d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, y)\\
&+ o_p\left(n^{-1/2} + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right),
\end{aligned} \tag{10.36}$$

where

$$\boldsymbol{B} = \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \mathbb{E}\Big[\psi_0'(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\,\mathrm{Cov}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\Big]\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right).$$

We get,

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = \sqrt{n}\boldsymbol{B}^{-1}\int &\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\} d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, y)\\
&+ o_p\left(1 + \sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right)\\
\to_d N(\mathbf{0}, \boldsymbol{\Pi}),&
\end{aligned} \tag{10.37}$$

where

$$\boldsymbol{\Pi} = \boldsymbol{B}^{-1}\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\boldsymbol{\Sigma}\,\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\,\boldsymbol{B}^{-1} \in \mathbb{R}^{(d-1)\times(d-1)}.$$

The asymptotic limiting distribution of the single index score estimator $\hat{\boldsymbol{\alpha}}_n$ now follows by an application of the Delta method and we conclude that

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) = J_{\mathbb{S}}(\boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right) \to_d N_d\left(\mathbf{0}, J_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{\Pi}\left(J_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\right).$$

Finally, the result of Theorem 4.1 follows by Lemma 4.1. This completes the proof.

Acknowledgements

The research of the third author was supported by the Research Foundation Flanders (FWO) [grant number 11W7315N]. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. For the simulations we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government - department EWI.

References

[1] Balabdaoui, F., Durot, C. and Jankowski, H. (2016). Least squares estimation in the monotone single index model. *arXiv preprint arXiv:1610.06026*.
[2] Balabdaoui, F., Groeneboom, P. and Hendrickx, K. (2017). Score estimation in the monotone single index model: supplement. working paper.
[3] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney Wiley Series in Probability and Mathematical Statistics. MR0326887 (48 ##5229)
[4] Cavanagh, C. and Sherman, R. P. (1998). Rank estimators for monotonic index models. *J. Econometrics* **84** 351–381. MR1630210
[5] Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis* **86** 213–226.
[6] Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics* 505–530.
[7] Groeneboom, P. and Hendrickx, K. (2017). Current status linear regression. To appear in Annals of Statistics, available at https://arxiv.org/abs/1601.00202.
[8] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints*. Cambridge Univ. Press, Cambridge.
[9] Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* **35** 303–316.
[10] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171 (94d:62134)
[11] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association* **84** 986–995.
[12] Hooke, R. and Jeeves, T. A. (1961). "Direct Search"Solution of Numerical and Statistical Problems. *Journal of the ACM (JACM)* **8** 212–229.
[13] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58** 71–120.
[14] Kuchibhotla, A. K. and Patra, R. K. (2016). Efficient Estimation in Single Index Models through Smoothing splines. available at https://arxiv.org/abs/1612.00068.
[15] Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61** 123–137. MR1201705
[16] Tanaka, H. (2008). Semiparametric least squares estimation of monotone single index models and its application to the iterative least squares estimation of binary choice models Technical Report.
[17] Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM J. Optim.* **7** 1–25. MR1430554
[18] van der Vaart, A. W. (1998). *Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge University Press, Cambridge. MR1652247 (2000c:62003)
[19] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes. Springer Series in Statistics*. Springer-Verlag, New York. With applications to statistics. MR1385671 (97g:60035)
[20] Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis* **97** 1162–1184.