

# Estimation and convergence rates in the distributional single index model

**Journal Article****Author(s):**

Balabdaoui, Fadoua; Henzi, Alexander; Looser, Lukas

**Publication date:**

2024

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000667298>

**Rights / license:**

[Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International](#)

**Originally published in:**

Statistica Neerlandica, <https://doi.org/10.1111/stan.12336>

# Estimation and convergence rates in the distributional single index model

Fadoua Balabdaoui<sup>1</sup> | Alexander Henzi | Lukas Looser

Seminar for Statistics, ETHZ, Zurich, Switzerland

## Correspondence

Fadoua Balabdaoui, Seminar for Statistics, ETHZ, Rämistrasse 101, 8092 Zurich, Switzerland.  
Email: [fadouab@ethz.ch](mailto:fadouab@ethz.ch)

The distributional single index model is a semiparametric regression model in which the conditional distribution functions  $P(Y \leq y|X = x) = F_0(\theta_0(x), y)$  of a real-valued outcome variable  $Y$  depend on  $d$ -dimensional covariates  $X$  through a univariate, parametric index function  $\theta_0(x)$ , and increase stochastically as  $\theta_0(x)$  increases. We propose least squares approaches for the joint estimation of  $\theta_0$  and  $F_0$  in the important case where  $\theta_0(x) = \alpha_0^\top x$  and obtain convergence rates of  $n^{-1/3}$ , thereby improving an existing result that gives a rate of  $n^{-1/6}$ . A simulation study indicates that the convergence rate for the estimation of  $\alpha_0$  might be faster. Furthermore, we illustrate our methods in an application on house price data that demonstrates the advantages of shape restrictions in single index models.

## KEYWORDS

monotone regression, isotonic distributional regression, single index model

## 1 | INTRODUCTION

Consider the classical regression framework in which one aims to predict a response variable  $Y \in \mathbb{R}$  with covariates  $X \in \mathcal{X} \subseteq \mathbb{R}^d$ . The popular generalized linear models (GLMs) assume that

$$\mathbb{E}[Y|X = x] = g_\phi(\alpha_0^\top x),$$

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Statistica Neerlandica* published by John Wiley & Sons Ltd on behalf of Netherlands Society for Statistics and Operations Research.

where  $Y$  follows an exponential family distribution,  $\alpha_0$  is unknown, and  $g_\phi$  is a monotone transformation known up to a dispersion parameter  $\phi$  that does not depend on the covariates. Balabdaoui, Durot, and Jankowski (2019) study a semiparametric variant of this model, the monotone single index model, where the function  $g_\phi$  is replaced by an unknown monotone function  $\psi_0$  that is estimated nonparametrically, jointly with  $\alpha_0$ . The focus of this article is an extension of the monotone single index model introduced by Henzi, Kleger, and Ziegel (2023), called the distributional single index model, which aims at estimating conditional cumulative distribution functions (CDFs) of  $Y$  given  $X$  rather than only its conditional expectation. The model assumes that

$$\mathbb{P}(Y \leq y | X = x) = F_0(\theta_0(x), y), \quad (1)$$

where  $y \mapsto F_0(z, y)$  is an unknown conditional distribution function for all fixed  $z \in \mathbb{R}$ ,  $\theta_0 : \mathbb{R}^d \rightarrow \mathbb{R}$  a mapping of the  $d$ -dimensional covariates to  $\mathbb{R}$ , and monotonicity of  $\psi_0$  is replaced by the assumption of stochastic monotonicity. Stochastic monotonicity means that  $F_0(z, y)$  is non-increasing in  $z$  for all fixed  $y \in \mathbb{R}$ , so graphically, the conditional CDFs  $F_0(z, y)$  shift to the right as  $z$  increases, or in simple words,  $Y$  tends to attain larger values when  $\theta_0(X)$  is large. In this article, we are interested in the special case where  $\theta_0(x) = \alpha_0^\top x$  is a linear function. The most popular families in GLMs—Gaussian, Binomial, Poisson, Gamma, Inverse Gaussian—satisfy the stochastic monotonicity assumption of the distributional single index model, save for a change of sign of  $\alpha_0$  for decreasing link functions. Thus, the model can be regarded as a semiparametric, distributional extension of GLMs. If  $Y$  has finite expectation, then

$$\mathbb{E}[Y | X = x] = \int_0^\infty (1 - F_0(\alpha_0^\top x, y)) dy - \int_{-\infty}^0 F_0(\alpha_0^\top x, y) dy,$$

is increasing in  $\alpha_0^\top x$ , so the assumption of stochastic monotonicity is stronger than monotonicity of the conditional expectation in this case. When  $Y$  is binary, the distributional single index model becomes a special case of the monotone single index model. Both the monotone single index model and the distributional single index model build on the idea of single index model introduced by Härdle, Hall, and Ichimura (1993), and we refer the interested readers to the literature reviews in Balabdaoui, Durot, and Jankowski (2019) and Henzi et al. (2023) for a comprehensive discussion of related work.

Rates for the estimation of the conditional CDFs in the distributional single index model have already been obtained by Henzi et al. (2023). They showed that for an independent and identically distributed (i.i.d.) sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  from model (1), if  $\hat{\theta}_n$  is a uniformly consistent estimator for  $\theta_0$  converging at a rate of  $o_p((\log(n)/n)^{1/2})$  and if  $\hat{F}_n$  is computed on the data  $(\hat{\theta}_n(X_1), Y_1), \dots, (\hat{\theta}_n(X_n), Y_n)$  with isotonic distributional regression (Henzi, Ziegel, & Gneiting, 2021; Mösching & Dümbgen, 2020), then

$$\sup_{y \in \mathbb{R}, x \in \mathcal{X}_{\varepsilon_n}} |\hat{F}_n(\hat{\theta}_n(x), y) - F_0(\theta_0(x), y)| = o_p((\log(n)/n)^{1/6}), \quad (2)$$

under certain regularity conditions. Here  $\mathcal{X}_\varepsilon = \{x \in \mathcal{X} : \theta_0(x) \pm \varepsilon \in I\}$  for an interval  $I$  on which  $\theta(X)$  has density bounded away from zero and infinity, and  $\varepsilon_n > 0$  is a certain sequence converging to zero. When  $\hat{\theta}_n$  and  $\hat{F}_n$  are computed on independent samples, a faster rate of  $o_p((\log(n)/n)^{1/3})$  is

achieved, if  $\hat{\theta}_n$  converges to  $\theta_0$  at least at this rate. Henzi et al. (2023) provide no theoretical results on the estimation of the index function, and the rate of  $o_p((\log(n)/n)^{1/6})$  is likely to be suboptimal, because if  $\theta_0 = \hat{\theta}_n$  it should be  $o_p((\log(n)/n)^{1/3})$  by theorem 3.3 of Mösching and Dümbgen (2020), or if  $Y$  is binary the results of Balabdaoui, Durot, and Jankowski (2019) yield  $O_p(n^{-1/3})$  for the estimation of  $F_0$  and  $\theta_0$  when the latter is linear.

In this article, we focus on the linear case  $\theta_0(x) = \alpha_0^\top x$ , and propose to estimate  $(F_0, \alpha_0)$  by minimizing weighted least squares criteria of the form

$$L_n(Q; F, \alpha) = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} (1\{Y_i \leq t\} - F(\alpha^\top X_i, t))^2 dQ(t), \quad (3)$$

where  $Q$  is a Borel measure. We obtain a rate of  $O_p(n^{-1/3})$  when  $Q$  has a finite support or it is compactly supported Lebesgue continuous with a bounded density. Furthermore, we investigate an approach with  $Q$  equal to the empirical distribution of  $Y_1, \dots, Y_n$ , which has favorable invariance properties under transformations of the response variable, but the consistency and convergence rates of which remain an open challenge.

The article is structured as follows. In Section 2 we describe the estimation method in detail. Convergence rates are derived in Section 3. In Section 4, we present the invariance property result which holds when  $Q$  is taken to be the empirical distribution function of the responses. In Section 5 we discuss computational aspects and present a simulation study and an application on house price data. We conclude with a discussion in Section 6, and the proofs are deferred to Appendix. Throughout the article, we denote the joint distribution of  $(X, Y)$  by  $\mathbb{P}$ , the marginals by  $\mathbb{P}^X$  and  $\mathbb{P}^Y$ , and the conditional distributions by  $\mathbb{P}^{Y|X=x}$  and  $\mathbb{P}^{X|Y=y}$ , respectively. The empirical distributions of  $n$  independent observations are denoted by  $\mathbb{P}_n, \mathbb{P}_n^X, \mathbb{P}_n^Y$ . We denote by  $\text{supp}(P)$  the support of a probability measure  $P$ , and by  $A^\circ$  the interior of a set  $A$ . The expectation operator  $\mathbb{E}[\cdot]$  is understood to be with respect to  $\mathbb{P}$ , unless explicitly defined differently.

## 2 | ESTIMATION

Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be a sample of covariates and response variable from model (1), where from now on we always assume that  $\theta_0(x) = \alpha_0^\top x$ . Define  $C_\alpha = \{\alpha^\top x : x \in \mathcal{X}\}$ , and let  $\mathcal{F}_\alpha : C_\alpha \times \mathbb{R} \rightarrow [0, 1]$  be the class of bivariate functions  $F$  for which  $y \mapsto F(z, y)$  is a CDF for all fixed  $z \in \mathbb{R}$ , and  $z \mapsto F(z, y)$  is non-increasing for all fixed  $y \in \mathbb{R}$ . The function  $F_0$  and the parameter  $\alpha_0$  in (1) are not identifiable, since  $\check{F}_0(z, y) = F_0(z/c, y)$  and  $\check{\alpha}_0 = c \cdot \alpha_0$  for  $c > 0$  yield the same conditional distributions. Hence, we assume that  $\alpha_0 \in S_{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ , and define the class of candidate functions for estimation by

$$\mathcal{F} = \{(F, \alpha) : \alpha \in S_{d-1}, F \in \mathcal{F}_\alpha\}.$$

To estimate  $(F_0, \alpha_0)$ , we propose to minimize the least squares criteria of the form given in (3). The following proposition describes the solutions of this minimization problem.

**Proposition 1.** *Assume that  $Q$  is locally finite.*

- (i) *For a fixed  $\alpha \in S_{d-1}$ , let  $z_1 < \dots < z_m$  be the distinct values of  $\alpha^\top X_1, \dots, \alpha^\top X_n$ , with multiplicities  $n_1, \dots, n_m$ . The minimizer of  $L_n(Q; F, \alpha)$  in  $\mathcal{F}$  is uniquely*

defined in the first argument on  $\{z_1, \dots, z_m\}$  and in the second argument on  $\text{supp}(Q)$ , and it is given by

$$\hat{F}_{n,\alpha}(z_i, y) = \min_{k=1, \dots, i} \max_{l=i, \dots, m} \left( \frac{1}{n_k + \dots + n_l} \sum_{j=k}^l \sum_{s: \alpha^\top X_s = z_j} 1\{Y_s \leq y\} \right), \quad i = 1, \dots, m. \quad (4)$$

(ii) Let  $S^X = \{\alpha \in S_{d-1} : \alpha^\top X_i \neq \alpha^\top X_j, i, j = 1, \dots, n, i \neq j\}$ . The minimum of  $L_n(Q; F, \alpha)$  is achieved for a pair  $(\hat{F}_{n, \hat{\alpha}_n}, \hat{\alpha}_n)$  with  $\hat{\alpha}_n \in S^X$  and  $\hat{F}_{n, \hat{\alpha}_n}$  given by (4). The minimizer is not unique.

The estimator  $\hat{F}_{n,\alpha}$  in (4) is called the isotonic distributional regression in Henzi et al. (2021), and the fact that it is a minimizer is due to theorem 2.1 of that article; the condition that  $Q$  is locally finite is only necessary to ensure uniqueness in part (i). It follows directly from (4) that  $y \mapsto \hat{F}_{n,\alpha}(\alpha^\top X_i, y)$  is indeed a CDF for  $i = 1, \dots, m$ . For a fixed  $\alpha$ , the estimator  $\hat{F}_{n,\alpha}$  depends on  $Q$  only through its support, as can be seen from (4). It suffices to compute it at the distinct values  $y_1 < \dots < y_k$  of  $Y_1, \dots, Y_n$ , since for  $y \geq y_1$ ,

$$1\{Y_i \leq y\} = 1\{Y_i \leq y_{l(y)}\}, \quad i = 1, \dots, n, \quad \text{with} \quad l(y) = \max\{j \in \{1, \dots, k\} : y_j \leq y\},$$

and  $1\{Y_i \leq y\} = 0$  if  $y < y_1$ . Part (ii) of the proposition follows by the same steps as Prop. 2.2 in Balabdaoui, Durot, and Jankowski (2019). Note that the minimizers  $\hat{\alpha}_n$  and, hence,  $\hat{F}_{n, \hat{\alpha}_n}$  do depend on  $Q$ , which appears in the criterion (3). To lighten the notation, we write  $\hat{F}_{n, \hat{\alpha}_n} = \hat{F}_n$  in the following, and only use the subscript when it is necessary to indicate the dependence on  $\hat{\alpha}_n$ . To define  $\hat{F}_n$  beyond the set  $\{\hat{\alpha}_n^\top X_1, \dots, \hat{\alpha}_n^\top X_n\} \times \text{supp}(Q)$ , we let

$$\hat{F}_n(z, y) = \begin{cases} 0, & y < y_1, \\ \hat{F}_n(z, y_j), & y \in [y_j, y_{j+1}), j = 1, \dots, k - 1, \\ 1, & y \geq y_k, \end{cases} \quad (5)$$

for  $z \in \{\hat{\alpha}_n^\top X_1, \dots, \hat{\alpha}_n^\top X_n\}$  and  $y \in \mathbb{R}$ , and

$$\hat{F}_n(z, y) = \begin{cases} \hat{F}_n(z_1, y), & z < z_1, \\ \frac{z_{j+1} - z}{z_{j+1} - z_j} \hat{F}_n(z_j, y) + \frac{z - z_j}{z_{j+1} - z_j} \hat{F}_n(z_{j+1}, y), & z \in [z_j, z_{j+1}), j = 1, \dots, m - 1, \\ \hat{F}_n(z_m, y), & z \geq z_m. \end{cases}$$

We apply these interpolation methods in our empirical studies in Section 5. For the theory, any other interpolation methods satisfying the monotonicity constraints in both arguments is admissible.

In the forecasting literature, the loss function (3) with  $Q$  equal to the Lebesgue measure  $\lambda$  is known under the name continuous ranked probability score (CRPS), which is a widely used proper scoring rule for the estimation of distribution functions and for forecast evaluation (Gneiting & Raftery, 2007). The criterion with general Borel measures  $Q$  are the so-called threshold weighted forms of the CRPS (Gneiting & Ranjan, 2011). At a first sight, the CRPS seems to be a

natural choice for the loss function since it weighs all thresholds equally, but it has the drawback that  $\mathbb{E}[L_n(\lambda; F_0, \alpha_0)]$  is finite only if the conditional distributions corresponding to  $F_0(\alpha_0^\top x, \cdot)$  have finite first moment; see (21) in Gneiting and Raftery (2007). This is an unnecessary assumption if the goal is the estimation of the conditional CDFs, rather than conditional expectations, and it complicates proofs of consistency. We therefore focus on finite measures  $Q$ .

### 3 | CONVERGENCE RATES

#### 3.1 | Assumptions

We proceed to establish consistency results for the bundled estimator  $\hat{F}_n(\hat{\alpha}_n^\top x, y)$  and for the separated estimators  $\hat{\alpha}_n$  and  $\hat{F}_n(z, y)$ . The proofs and assumptions are closely related to those by Balabdaoui, Durot, and Jankowski (2019) for the monotone single index model.

**Assumption 1.** The set  $\mathcal{X}$  is bounded and convex.

**Assumption 2.** The measure  $Q$  and the distribution of  $(X, Y)$  satisfy one of the following assumptions.

- (i) The distribution of  $X$  admits a Lebesgue density  $p_X$  which is bounded from below by  $\underline{p}_X > 0$  and from above by  $\bar{p}_X < \infty$ , and  $Q$  has finite support, putting mass only on points  $t_1 < \dots < t_p$ .
- (ii) For all  $y \in \text{supp}(\mathbb{P}^Y)$ , the distribution of  $X$  conditional on  $Y = y$  admits a Lebesgue density bounded from below by  $\underline{p}_X > 0$  and from above by  $\bar{p}_X < \infty$ , with constants not depending on  $y$ . The measure  $Q$  has support on  $[a, b]$  and admits a Lebesgue density  $q$  bounded from above by  $c < \infty$ .

**Assumption 3.** For all  $t \in \text{supp}(Q)$  the function  $z \mapsto F_0(z, t)$  is continuously differentiable on  $C_{\alpha_0}$  with derivative  $F_0^{(1)}(z, t)$ , and  $0 < |F_0^{(1)}(z, t)| \leq K_t$  for all  $z \in C_{\alpha_0}^\circ$  and some  $K_t < \infty$ .

**Assumption 4.** For all  $\alpha \in S_{d-1}$ , the random variable  $\alpha^\top X$  admits a Lebesgue density bounded from below by  $\underline{q} > 0$  and from above by  $\bar{q} > 0$ .

**Assumption 5.** The density  $p_X$  of  $X$  is continuous on  $\mathcal{X}$ .

Assumptions 1, 4 and 5 correspond to (A1), (A4) and (A6) in Balabdaoui, Durot, and Jankowski (2019), respectively, and Assumption 3 is a direct extension of their condition (A5) to our case.

In the next sections, we present one of the main convergence results of this work, derived under the assumptions above. The case  $Q = \mathbb{P}_n^Y$  would have been a natural choice. One referee raised the point of whether one can derive rates of convergence in this case when the distribution of  $Y$  is compactly supported. Unfortunately, compactness of the support does not solve the issue that empirical process associated with the estimation problem at hand contains a term that cannot be handled with the classical results such as Lemma 3.4.2 or Lemma 3.4.3 of van der Vaart and Wellner (1996). The reason behind the additional difficulties is that this term in question is of the form

$$\int \psi_n(t)(\mathbb{P}_n^Y - \mathbb{P}^Y)(t),$$

where  $\psi_n$  is random function involving the empirical measure  $\mathbb{P}_n$ ; additional details are provided in Section 4. More sophisticated tools need to be used in this context. The problem is beyond the scope of this article but worth investigating in future research.

### 3.2 | Convergence rate for the bundled estimator

The results for convergence rates for both types of  $Q$  in Assumption 2 are presented in a unified framework. For the bundled estimator, we obtain the following result.

**Theorem 1.** *Under Assumptions 1 and 2, it holds that*

$$\left( \int_{\mathbb{R}} \int_{\mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) \right)^{1/2} = O_p(n^{-1/3}).$$

The proof of Theorem 1 applies theorem 3.4.1 and Lemma 3.4.2 of van der Vaart and Wellner (1996), and it is given in Section A.1. In the following, we introduce empirical process notation, provide auxiliary results that are of independent interest, and discuss the techniques and problems involved in the proof.

In accordance with Assumption 1, assume  $\|x\| \leq R$  for all  $x \in \mathcal{X}$  and some  $R > 0$ , so that  $|\alpha^\top x| \leq R$  for  $\alpha \in S^{d-1}$ . In the proofs, the following function classes appear,

$$\begin{aligned} \mathcal{H} &= \{h : [-2R, 2R] \rightarrow [0, 1], \text{ nonincreasing}\}, \\ \mathcal{G} &= \{g : \mathcal{X} \rightarrow [0, 1], g(x) = h(\alpha^\top x), (\alpha, h) \in S_{d-1} \times \mathcal{H}\}, \end{aligned}$$

where the support in the class  $\mathcal{H}$  has to be extended to  $[-2R, 2R]$  for technical reasons. Nonincreasing functions  $\tilde{h} : [-R, R] \rightarrow [0, 1]$  are considered as elements of  $\mathcal{H}$  by constant extrapolation at the boundaries. Denote the  $L_2$ -norm of functions from  $\mathcal{X}$  to  $\mathbb{R}$ , with respect to a Borel measure  $\mu$ , by

$$\|f\|_\mu = \left( \int_{\mathcal{X}} f(x)^2 d\mu(x) \right)^{1/2}.$$

For integration with respect to the Lebesgue measure over a set  $A$ , we write  $\|f\|_A$ . The bracketing entropy of a function class  $\mathcal{T}$  with respect to some norm  $\|\cdot\|$  is denoted by  $N_B(\epsilon, \mathcal{T}, \|\cdot\|)$ , and the bracketing integral is defined as

$$\tilde{J}(\delta, \mathcal{T}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_B(\epsilon, \mathcal{T}, \|\cdot\|)} d\epsilon.$$

The following proposition, which relies on theorem. 2.7.5 of van der Vaart and Wellner (1996) and a result of Feige and Schechtman (2002), is crucial for all our results.

**Proposition 2.** *Let  $\mu$  be a Lebesgue continuous distribution with support in a bounded set contained in a ball of radius  $R > 0$  with density bounded from above by  $D > 0$ . Then,*

$$\log(N_B(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathcal{X}})) \leq \frac{2^{(d+1)/2} d^{1/4} R^{(d-1)/2} (1 + \sqrt{R})(d\sqrt{A} + K\sqrt{R})\sqrt{D}}{\varepsilon}$$

for universal constants  $A, K > 0$ .

Due to Proposition 2, the entropy of the class of functions  $x \mapsto F(\alpha^\top x, y)$  for  $(F, \alpha) \in \mathcal{F}$  and  $y$  fixed is of the same order as the entropy of the monotone function class with values in  $[0, 1]$ . If  $Q$  has finite support, this is sufficient to obtain the cubic convergence rate. However, as one would expect, the constants in the bounds increase with the size of the support, and it is not possible to extend the same proof strategy to Lebesgue continuous  $Q$ . For this case, a bound for the entropy of the class

$$\mathcal{M} := \left\{ h : \mathbb{R}^d \times \mathbb{R} \rightarrow [0, 1], h(x, y) = \int_{[y, \infty)} F(\alpha^\top x, t)^2 dQ(t), (F, \alpha) \in \mathcal{F} \right\}, \quad (6)$$

is required. We find such a bound by constructing a suitable discretization of the support of  $Q$ .

*Remark 1.* One might think that a simpler way to bound the entropy of the class  $\mathcal{M}$  would be via the results of Gao and Wellner (2007) on the entropy of multivariate monotone function. Indeed, the function  $(z, y) \mapsto F(z, y)$  is bivariate monotone, and due to Proposition 2, the fact that we have  $\alpha^\top x$  in the first argument only increases the entropy by a constant factor. However, according to theorem 1.1 of Gao and Wellner (2007), the entropy of the class of bivariate monotone functions is of order  $1/\varepsilon^2$ , which leads to a diverging entropy integral. Even with the relaxation discussed on p. 326 of van der Vaart and Wellner (1996), which allows to integrate only from  $\min(u\delta^2, \delta)/3$  for small  $u > 0$  in the entropy integral, it is not possible to achieve the cubic rate with this entropy bound.

### 3.3 | Convergence rate for the separated estimators

The rate for the separated estimators  $\hat{F}_n(z, y)$  and  $\hat{\alpha}_n$  relies on Theorem 1 and is proved in a similar way as in theorem 5.2 and corollary 5.3 of Balabdaoui, Durot, and Jankowski (2019). Note that under our model assumptions, the parameters  $F$  and  $\alpha$  are indeed identifiable. More precisely, if  $F(\alpha^\top X, t) = F_0(\alpha_0^\top X, t)$  almost surely for a fixed  $t$ , then  $F(z, t) = F_0(z, t)$  for  $(z, t) \in C_{\alpha_0} \times \text{supp}(Q)$ , and  $\alpha = \alpha_0$ . This is shown in an analogous way as in proposition 5.1 of Balabdaoui, Durot, and Jankowski (2019), and it is proven in Section A.5 for completeness.

**Theorem 2.** *Let Assumptions 1, 2, and 4 hold true. Assume that for each  $t$  the function  $F_0(\cdot, t)$  is left-continuous, nonconstant and does not have discontinuity points on the boundary of  $C_{\alpha_0}$ . Furthermore, assume that from each subsequence  $(n_k)_{k \in \mathbb{N}}$  we can extract another subsequence  $(n_{k_l})_{l \in \mathbb{N}}$  which satisfies*

$$\lim_{l \rightarrow \infty} \int_{\mathbb{R}} \int_{\mathcal{X}} (F_0(\alpha_0^\top x, t) - \hat{F}_{n_{k_l}}(\hat{\alpha}_{k_l}^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) = 0, \quad (7)$$

almost surely. Then,



- (i)  $\hat{\alpha}_n$  converges to  $\alpha_0$  in probability in the euclidean norm,
- (ii) for all continuity points  $(z, t)$  of  $F_0$  in  $C_{\alpha_0}^{\circ} \times \text{supp}(Q)$ , we have that  $\hat{F}_n(z, t)$  converges to  $F_0(z, t)$  in probability.

**Remark 2.** The condition (7) in Theorem 2 holds under our assumptions due to Theorem 1.

**Theorem 3.** Define  $\underline{c} = \inf C_{\alpha_0}$  and  $\bar{c} = \sup C_{\alpha_0}$ . Under Assumptions 1–5, we have that

- (i)  $\|\alpha_0 - \hat{\alpha}_n\| = O_P(n^{-1/3})$ ;
- (ii) if  $\sup_{t \in \text{supp}(Q)} K_t < \infty$ , then

$$\left( \int_{\mathbb{R}} \int_{\underline{c} + v_n}^{\bar{c} - v_n} (F_0(z, t) - \hat{F}_n(z, t))^2 dz dQ(t) \right)^{1/2} = O_P(n^{-1/3}), \quad (8)$$

for all sequences  $v_n$  such that  $\underline{c} + v_n \leq \bar{c} - v_n$  and  $n^{1/3}v_n \rightarrow \infty$  for  $n \rightarrow \infty$ .

Part (ii) of Theorem 3 can be regarded as analogous to the result (2) derived by Henzi et al. (2023, theorem 5.1), with the weighted  $L_2$ -norm replacing the supremum norm. Henzi et al. (2023) do not assume a linear index function, but they impose the assumption that the index function is estimated the rate of  $(\log(n)/n)^{1/2}$ , rather than deriving a convergence rate, which we do in part (i) of the above theorem.

## 4 | EMPIRICAL DISTRIBUTION AS WEIGHTING MEASURE

The methods proposed so far require the specification of a weighting measure  $Q$ . An interesting variant of the criterion (3), which does not require an explicit weighting choice, arises when  $Q$  equals the empirical distribution  $\mathbb{P}_n^Y$ ; that is,

$$L_n(\mathbb{P}_n^Y; F, \alpha) = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} (1\{Y_i \leq t\} - F(\alpha^\top X_i, t))^2 d\mathbb{P}_n^Y(t) = \frac{1}{n^2} \sum_{i,j=1}^n (1\{Y_i \leq Y_j\} - F(\alpha^\top X_i, Y_j))^2.$$

According to the following lemma, for this choice of  $Q$  the estimator  $\hat{\alpha}_n$  and the pointwise error of the CDFs at the observed values of the response variable do not depend on the scale of the observations  $Y$ .

**Lemma 1.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be strictly increasing on the support of  $Y$ , and  $f^{-1}(t) = \inf\{s \in \mathbb{R} : f(s) \geq t\}$ . Then, the following hold with probability one.

- (i) A tuple  $(\hat{F}_{n, \hat{\alpha}_n}, \hat{\alpha}_n)$  minimizes  $L_n(\mathbb{P}_n^Y; \cdot)$  if and only if  $(\tilde{F}_{n, \hat{\alpha}_n}, \hat{\alpha}_n)$  with  $\tilde{F}_{n, \hat{\alpha}_n}(t, z) = \hat{F}_{n, \hat{\alpha}_n}(z, f^{-1}(t))$  is a minimizer of  $L_n(\mathbb{P}_n^{f(Y)}; \cdot)$ , and it holds that  $L_n(\mathbb{P}_n^Y; \hat{F}_{n, \hat{\alpha}_n}, \hat{\alpha}_n) = L_n(\mathbb{P}_n^{f(Y)}; \tilde{F}_{n, \hat{\alpha}_n}, \hat{\alpha}_n)$ .
- (ii) With  $\tilde{F}_0(z, t) = F_0(z, f^{-1}(t))$  and  $t_i = f(Y_i)$ , we have

$$\tilde{F}_{n, \hat{\alpha}_n}(\hat{\alpha}_n^\top X_i, t_i) - \tilde{F}_0(\alpha^\top X_i, t_i) = \hat{F}_{n, \hat{\alpha}_n}(\hat{\alpha}_n^\top X_i, Y_i) - F_0(\alpha^\top X_i, Y_i), \quad i = 1, \dots, n.$$

If  $\tilde{F}_{n,\hat{\alpha}_n}$  and  $\hat{F}_{n,\hat{\alpha}_n}$  are interpolated as in (5), then the above equality holds for all  $y \in \mathbb{R}$  and  $t = f(y)$ .

The above result is generally not true for  $Q \neq \mathbb{P}_n^Y$  in (3). The invariance property aligns well with the fact that the transformed outcome  $f(Y)$  again follows a distributional single index model with the same parameter  $\alpha_0$  and the corresponding CDFs  $t \mapsto F_0(\alpha_0^\top x, f^{-1}(t))$ . However, it turns out that deriving convergence rates for this criterion is substantially more difficult than for fixed measures  $Q$ , because the integral in the function class  $\mathcal{M}$  in (6) is now over the random measure  $\mathbb{P}_n$  instead of the fixed measure  $Q$ . We suspect that the rate for this estimator should still be of order  $O_p(n^{-1/3})$ , and our simulations confirm this intuition in certain examples. However, a completely different strategy of proof seems necessary to prove this rate.

## 5 | EMPIRICAL RESULTS

### 5.1 | Simulations

We investigate the convergence of our estimators in simulations. For  $d = 2, 3$ , we simulate  $X_j \sim \text{Unif}(0, 1)$ ,  $j = 1, \dots, d$ , independently, and generate the response variable in two ways,

$$Y^{(1)} = (\alpha_0^\top X)^3 \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1), \quad Y^{(2)} = (\alpha_0^\top X)^3 \eta, \quad \eta \sim \text{Exp}(1). \quad (9)$$

For the weighting measure  $Q$ , we consider the empirical distribution  $\mathbb{P}_n^Y$ , the uniform distribution on  $[-10, 10]$  and the Gaussian distribution with variance 4 truncated to the interval  $[-4, 10]$  for the simulations with Gaussian noise, and the uniform distribution on  $[0, 50]$  and the truncated Gamma distribution with shape 3 and scale 1 for the simulations with exponentially distributed noise, respectively. The rationale is that the uniform distribution over a large set provides a rather rough choice for the weighting, whereas the truncated distributions more closely follow the actual outcome distributions, up to truncation to a compact interval.

The index  $\alpha_0$  is parameterized in spherical coordinates with  $\theta_0 \in [0, 2\pi]$  and values  $\theta_0 = \pi/4, \pi/3, \pi/2$  for  $d = 2$ , and  $\theta_0 \in [0, \pi] \times [0, 2\pi]$  and values  $\theta_0 = (\pi/4, \pi/2), (\pi/3, \pi/3), (\pi/2, \pi/4)$  for  $d = 3$ . To perform estimation, we parameterize  $\alpha$  in spherical coordinates and do a grid search followed by local numerical optimization. For  $d = 2$ , we choose 40 equidistant points  $\theta_1 = 0 < \theta_2 < \dots < \theta_{40} = 2\pi$ , evaluate the criterion (3) at  $\alpha_j = (\cos(\theta_j), \sin(\theta_j))$ , and perform numerical optimization of (3) with respect to  $\theta$  in  $\alpha = (\cos(\theta), \sin(\theta))$  around the  $\theta_j$  for which the minimal value of the criterion is attained. The procedure for  $d = 3$  is analogous, and for the grid we take all combinations of 20 equidistant points  $\theta_{1,j} \in [0, \pi]$  and 40 points  $\theta_{2,k} \in [0, 2\pi]$ ,  $j = 1, \dots, 20$ ,  $k = 1, \dots, 40$ . Numerical optimization is performed with `optimize` in R (R Core Team, 2022) for  $d = 2$ , and `nmkb` from the package `dfoptim` (Varadhan, Borchers, & Bechard, 2020) for  $d = 3$ . Estimation of the conditional CDFs uses the `isodistrreg` package (Henzi et al., 2021). A general implementation of our estimator and replication material for Section 5 are available on [https://github.com/AlexanderHenzi/distr\\_single\\_index](https://github.com/AlexanderHenzi/distr_single_index).

To estimate the rates of convergence, we simulate 100 realizations of the examples described above for each of the sample sizes  $n = 2^m$ ,  $m = 8, 9, \dots, 13$ , and compute the index error  $\|\hat{\alpha}_n - \alpha_0\|$ , the bundled error  $L(\hat{F}_n, \hat{\alpha}_n)$ , and of the error of the CDFs  $L_{\text{CDF}}(\hat{F}_n)$ . The integrals in  $L(\hat{F}_n, \hat{\alpha}_n)$  and  $L_{\text{CDF}}(\hat{F}_n)$  are estimated with the mean of the integrand evaluated at 5,000 draws

for  $y \sim \mathbb{P}^Y$  and  $x \sim \mathbb{P}^X$ , or  $z \sim \text{Unif}(c, \bar{c})$ , respectively. We then estimate the convergence rate with the slope coefficient from regressing  $\log(\text{err}_{n,i})$ , for all  $n$  and samples  $i = 1, \dots, 100$ , on  $-\log(n)$ , for each setting and error measure. The estimates and SEs are shown in Table 1. Naturally, there are many factors influencing the convergence rates estimates, such as noise in the estimation, different constants in different examples, and, most importantly, the fact that the rates of the errors are only estimated on a grid of finite sample sizes. Therefore, even if one might expect the same asymptotic rates in the examples that we consider, there are some deviations due to different constants and finite sample effects. However, Table 1 suggests that the rate of  $\hat{\alpha}_n$  is faster than  $n^{-1/3}$ , as in the experiments of Balabdaoui, Durot, and Jankowski (2019), and the rates for the bundled estimator and for the CDF are around  $n^{-1/3}$ . There are no systematic differences between the results for  $Q = \mathbb{P}_n^Y$  and for the other approaches, with average rates over all settings of 0.51, 0.28, and 0.39 for the index, CDF, and bundled estimator for the empirical weighting measure, and 0.51, 0.30, and 0.38 for the other weighting methods. This suggests that the same rates should hold for  $Q = \mathbb{P}_n^Y$ .

*Remark 3.* For dimension  $d = 1$ , the computation of  $\hat{\alpha}_n$  is a one-dimensional optimization problem, and  $\hat{\alpha}_n$  can be approximated to a high accuracy provided that the grid for the initial grid search is fine enough. For  $d > 2$  the grid search becomes expensive, and there are no guarantees that a pair  $(\tilde{\alpha}_n, \tilde{F}_n)$  chosen by our implementation is a global minimizer of our target function, which is nonsmooth and nonconvex. Estimation in the monotone single index model for the mean suffers from the same optimization difficulties, and although there has been extensive research on implementation and alternative methods for estimating  $\hat{\alpha}_n$  (Balabdaoui & Groeneboom, 2021; Balabdaoui, Groeneboom, & Hendrickx, 2019; Groeneboom, 2018; Groeneboom & Hendrickx, 2019), the computation of  $\hat{\alpha}_n$  remains a challenge, especially in higher dimensions.

## 5.2 | Illustration on house price data

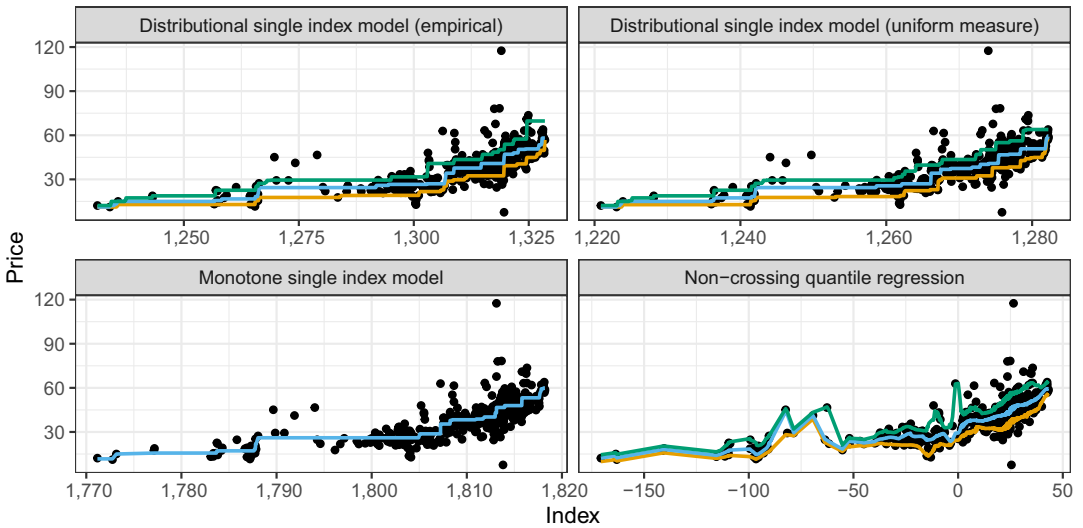
We illustrate the distributional single index model in a data example by Jiang and Yu (2023, section 4.4). The dataset, which is available on <https://doi.org/10.24432/C5J30W>, contains 414 real estate transaction records from Taipei City and New Taipei City. The dependent variable is the price per unit area, and the covariates are the number of convenience stores in the living circle on foot, the building age, the transaction year and month, and the distance to the nearest metro station. The transaction time is transformed to a numerical variable with values in between 2,012.67 and 2,013.58, and it is a proxy variable which captures effects such as trends in the house prices, or different policy regimes over time that might influence the prices.

Figure 1 depicts the index values  $\hat{\alpha}_n^\top X_i$  and prices  $Y_i$ ,  $i = 1, \dots, 414$ , for the distributional single index model, the monotone single index model, and for the noncrossing quantile regression estimator by Jiang and Yu (2023); the results for the latter are equal to those shown in (Jiang & Yu, 2023, figure 3(c)) and reproduced with the code from the supplement of their article. We implemented the distributional index model with the empirical measure and with the uniform measure over a large set including all observed prices. For the distributional methods, the lines in the figure show estimated conditional quantiles at levels  $\tau = 0.1, 0.5, 0.9$ , which are obtained by inversion of the CDFs for our estimator. Jiang and Yu (2023) center all covariates around their mean before estimation. With shape restricted estimation methods, such centering is not

**TABLE 1** Estimated convergence rates and SEs (in parentheses) in the simulation examples (9), with  $\theta_0$  giving the spherical coordinates for  $\alpha_0$ .

$Q$	Spherical coordinates $\theta_0$							
	Simulation	Error type	$\pi/4$	$\pi/3$	$\pi/2$	$(\pi/4, \pi/2)$	$(\pi/3, \pi/3)$	$(\pi/2, \pi/4)$
Empirical	Exponential	Index	0.49 (0.04)	0.50 (0.04)	0.63 (0.05)	0.48 (0.03)	0.47 (0.03)	0.49 (0.03)
		CDF	0.25 (0.01)	0.23 (0.01)	0.36 (0.01)	0.27 (0.01)	0.19 (0.02)	0.27 (0.01)
		Bundled	0.36 (0.01)	0.36 (0.01)	0.37 (0.01)	0.37 (0.01)	0.37 (0.01)	0.37 (0.01)
Truncated	Gaussian	Index	0.49 (0.04)	0.50 (0.05)	0.60 (0.05)	0.48 (0.03)	0.55 (0.03)	0.48 (0.03)
		CDF	0.30 (0.01)	0.30 (0.01)	0.37 (0.02)	0.33 (0.01)	0.33 (0.01)	0.34 (0.01)
		Bundled	0.38 (0.01)	0.38 (0.01)	0.40 (0.01)	0.39 (0.01)	0.39 (0.01)	0.39 (0.01)
Uniform	Exponential	Index	0.50 (0.04)	0.53 (0.04)	0.54 (0.05)	0.51 (0.03)	0.48 (0.03)	0.46 (0.03)
		CDF	0.24 (0.02)	0.24 (0.02)	0.38 (0.03)	0.29 (0.02)	0.21 (0.02)	0.27 (0.02)
		Bundled	0.37 (0.01)	0.37 (0.01)	0.40 (0.02)	0.39 (0.01)	0.39 (0.01)	0.40 (0.01)
Truncated	Gaussian	Index	0.46 (0.04)	0.48 (0.05)	0.55 (0.05)	0.48 (0.03)	0.55 (0.03)	0.50 (0.03)
		CDF	0.28 (0.01)	0.29 (0.02)	0.37 (0.02)	0.32 (0.02)	0.30 (0.01)	0.33 (0.02)
		Bundled	0.38 (0.01)	0.38 (0.01)	0.41 (0.01)	0.39 (0.01)	0.40 (0.01)	0.40 (0.01)
Uniform	Exponential	Index	0.56 (0.04)	0.55 (0.04)	0.61 (0.05)	0.45 (0.03)	0.49 (0.03)	0.48 (0.03)
		CDF	0.21 (0.02)	0.21 (0.02)	0.38 (0.02)	0.27 (0.03)	0.19 (0.03)	0.24 (0.03)
		Bundled	0.37 (0.01)	0.36 (0.01)	0.39 (0.01)	0.37 (0.01)	0.38 (0.01)	0.39 (0.01)
Truncated	Gaussian	Index	0.49 (0.04)	0.46 (0.04)	0.56 (0.05)	0.48 (0.03)	0.53 (0.03)	0.49 (0.03)
		CDF	0.26 (0.02)	0.27 (0.02)	0.38 (0.02)	0.29 (0.02)	0.22 (0.02)	0.29 (0.02)
		Bundled	0.38 (0.01)	0.38 (0.01)	0.41 (0.01)	0.39 (0.01)	0.39 (0.01)	0.39 (0.01)

Notes: The rates are the slope coefficient from regressing  $\log(\text{err}_{n,i})$  on  $-\log(n)$ , with  $n = 2^8, 2^9, \dots, 2^{13}$  and samples  $i = 1, \dots, 100$ , for each error measure, simulation setting, and weighting measure for our estimator. The rate in the index estimation is mostly close to  $1/2$ , while the rates for the other measures are around  $1/3$ .



**FIGURE 1** Pairs  $(\hat{\alpha}_n^\top X_i, Y_i)$ ,  $i = 1, \dots, 414$ , for the distributional single index model, the monotone single index model, and for noncrossing quantile regression. The lines for the distributional methods are estimated conditional quantile curves at the levels  $\tau = 0.1, 0.5, 0.9$ .

necessary since it does not change the order of the projections  $\hat{\alpha}_n^\top X_i$ . As the scatterplots suggest, the order of the index values  $\hat{\alpha}_n^\top X_i$ ,  $i = 1, \dots, 414$ , obtained with the three methods are very similar, and the pairwise Spearman correlations between them are indeed all above 0.98. In the given data application, all methods have advantages and disadvantages. The computation of the estimator by Jiang and Yu (2023) is fast, but it involves several tuning parameters, namely, an initial quantile level for estimation, set to  $\tau = 0.5$ , bandwidths for kernel smoothing, and a prespecified grid of quantiles on which the estimator is computed and evaluated, chosen to be  $\tau = 0.1, 0.2, \dots, 0.9$ . Estimation for our method and for the monotone single index model is slower, since we take a fine grid for the grid search over  $\alpha$  and perform local optimization in several regions to ensure a good approximation of the minimum. However, the parameters of the shape restricted methods are more easily interpretable due to the monotone dependence on  $\hat{\alpha}_n^\top X$ . One can draw the—reasonable—conclusions that the price is increasing in the number of closely situated convenience stores and over time, and decreasing in the distance to the nearest metro station and in the age of the building; see Table 2. The interpretation is more difficult for the estimator by Jiang and Yu (2023). Although the signs of  $\hat{\alpha}_n$  in their estimator agree with those of the shape restricted methods, the conditional quantile curves are nonmonotone and interpolate the prices for some of the observations.

## 6 | DISCUSSION

In this article, we proposed estimators for the distributional single index model, and proved a convergence rate of  $O_p(n^{-1/3})$  both for bundled and separated estimators. This greatly improves upon the  $(\log(n)/n)^{1/6}$ -rate known so far. There are several avenues for future research. Consistency for our transformation-invariant estimator proposed in Section 4 is an open challenge, which goes beyond the techniques applied for the convergence rates in this article. A possible future research

TABLE 2 Estimates  $\hat{\alpha}_n$  for the different methods.

Method	Number stores	Building age	Transaction date	Distance metro
Distributional index model (empirical)	0.706	-0.263	0.658	-0.013
Distributional index model (uniform)	0.750	-0.186	0.634	-0.008
Monotone single index model	0.415	-0.122	0.902	-0.006
Noncrossing quantile regression	0.152	-0.060	0.987	-0.004

Notes: For the noncrossing quantile regression, we show the entries for NCCQR<sub>0</sub> from table 6 of Jiang and Yu (2023), standardized to norm 1 for comparability.

direction is to study convergence under more general weighting measures  $Q$  with possibly an unbounded support. This would allow analyzing whether there is an optimal choice of  $Q$  in terms of the estimation error for  $\alpha_0$ . As for the monotone single index model, our simulations also suggest that  $\alpha_0$  is estimated at a faster rate. Deriving this rate, as well as a comparison to the estimators for  $\alpha_0$  in the monotone single index model, would be an interesting direction for future work.

## ACKNOWLEDGMENTS

We are grateful to an anonymous referee for helpful comments. Open access funding provided by Eidgenossische Technische Hochschule Zurich.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

Fadoua Balabdaoui  <https://orcid.org/0000-0001-8042-4193>

## REFERENCES

- Balabdaoui, F., Durot, C., & Jankowski, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B), 3276–3310.
- Balabdaoui, F., & Groeneboom, P. (2021). *Profile least squares estimators in the monotone single index model*. In *Advances in contemporary statistics and econometrics — Festschrift in honor of Christine Thomas-Agnan* (pp. 3–22). Cham, Switzerland: Springer.
- Balabdaoui, F., Groeneboom, P., & Hendrickx, K. (2019). Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics*, 46(2), 517–544.
- Feige, U., & Schechtman, G. (2002). On the optimality of the random hyperplane rounding technique for max cut. *Random Structures & Algorithms*, 20(3), 403–440.
- Gao, F., & Wellner, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9), 1751–1764.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.
- Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3), 411–422.
- Groeneboom, P. (2018). *Algorithms for computing estimates in the single index model*. Retrieved from. <https://github.com/pietg/single.index>
- Groeneboom, P., & Hendrickx, K. (2019). Estimation in monotone single-index models. *Statistica Neerlandica*, 73(1), 78–99.

- Härdle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, 21(1), 157–178.
- Henzi, A., Kleger, G.-R., & Ziegel, J. F. (2023). Distributional (single) index models. *Journal of the American Statistical Association*, 118(541), 489–503.
- Henzi, A., Ziegel, J. F., & Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 83(5), 963–993.
- Jiang, R., & Yu, K. (2023). No-crossing single-index quantile regression curve estimation. *Journal of Business & Economic Statistics*, 41(2), 309–320.
- Lavrič, B. (1993). Continuity of monotone functions. *Universitatis Masarykianae Brunensis*, 29(1-2), 1–4.
- Mösching, A., & Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14(1), 24–49.
- Murphy, S. A., van der Vaart, A. W., & Wellner, J. A. (1999). Current status regression. *Mathematical Methods of Statistics*, 8(3), 407–425.
- R Core Team. (2022). *R: A Language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- van der Vaart, A. W. (1998). *Asymptotic statistics Cambridge series in statistical and probabilistic mathematics* (Vol. 3). Cambridge, England: Cambridge University Press.
- van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes Springer series in statistics*. New York, NY: Springer-Verlag With applications to statistics.
- Varadhan, R., Borchers, H. W., & Bechard, V. (2020). *dfoptim: Derivative-free optimization*. R package version 2020.10-1.

**How to cite this article:** Balabdaoui, F., Henzi, A., & Looser, L. (2024). Estimation and convergence rates in the distributional single index model. *Statistica Neerlandica*, 1–28. <https://doi.org/10.1111/stan.12336>

## APPENDIX A. PROOFS

### A.1 Proof of Theorem 1

The proof of Theorem 1 is slightly different for the two cases in Assumption 2, which involve different entropy calculations. We first give a proof for the theorem with an unspecified constant in an entropy bound, and then derive the constant for the two cases in separate lemmas.

*Proof of Theorem 1.* The proof applies theorem 3.4.1 and Lemma 3.4.2 of van der Vaart and Wellner (1996).

$$\begin{aligned}\mathbb{M}_n(F, \alpha) &= \int_{\mathbb{R}} \int_{\mathbb{R}^d \times \mathbb{R}} (1\{y \leq t\} - F(\alpha^\top x, t))^2 d\mathbb{P}_n(x, y) dQ(t), \\ \mathbb{M}(F, \alpha) &= \int_{\mathbb{R}} \int_{\mathbb{R}^d \times \mathbb{R}} (1\{y \leq t\} - F(\alpha^\top x, t))^2 d\mathbb{P}(x, y) dQ(t).\end{aligned}$$

Expanding the squares and using the fact that  $\mathbb{E}[1\{Y \leq t\} | X = x] = F_0(\alpha_0^\top x, t)$  yields

$$\mathbb{M}(F, \alpha) - \mathbb{M}(F_0, \alpha_0) = \int_{\mathbb{R}} \int_{\mathbb{R}} (F(\alpha^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}(x) dQ(t) =: d((F, \alpha), (F_0, \alpha_0))^2.$$



Furthermore, we have

$$\mathbb{M}_n(F, \alpha) - \mathbb{M}(F, \alpha) = \int_{\mathbb{R}} \int_{\mathbb{R}^d \times \mathbb{R}} (1\{y \leq t\} - F(\alpha^\top x, t))^2 dQ(t) d(\mathbb{P}_n(x, y) - \mathbb{P}(x, y)),$$

or, when rescaling with  $\sqrt{n}$  and using empirical process notation,

$$\sqrt{n}(\mathbb{M}_n(F, \alpha) - \mathbb{M}(F, \alpha)) = \mathbb{G}_n \int_{\mathbb{R}} (1\{y \leq t\} - F(\alpha^\top x, t))^2 dQ(t).$$

We now analyze the functions of the form

$$\ell(x, y) = \int_{\mathbb{R}} (1\{y \leq t\} - F(\alpha^\top x, t))^2 dQ(t),$$

with  $(F, \alpha) \in \mathcal{F}$ , and denote the class of such functions by  $\mathcal{L}$ . Also, let  $\mathcal{L}_\delta$  contain all functions of type

$$\tilde{\ell}(x, y) = \ell(x, y) - \int_{\mathbb{R}} (1\{y \leq t\} - F_0(\alpha_0^\top x, t))^2 dQ(t),$$

with  $\ell \in \mathcal{L}$  and for which

$$\delta^2 \geq \|\tilde{\ell}\|_{\mathbb{P}}^2 = d((F, \alpha), (F_0, \alpha_0))^2 = \mathbb{M}(F, \alpha) - \mathbb{M}(F_0, \alpha_0).$$

The elements in  $\mathcal{L}_\delta$  are obtained by shifting elements of  $\mathcal{L}$  by a fixed function, so we have  $N_B(\varepsilon, \mathcal{L}_\delta, \|\cdot\|_{\mathbb{P}}) \leq N_B(\varepsilon, \mathcal{L}, \|\cdot\|_{\mathbb{P}})$ . To apply Lemma 3.4.2 of van der Vaart and Wellner (1996), we have to find an upper bound for the bracketing entropy of the class  $\mathcal{L}$ . Since  $Q$  is a finite measure, we have

$$\ell(x, y) = \underbrace{1 - Q([t, \infty))}_{=: f(t)} + \underbrace{\int_{\mathbb{R}} F(\alpha^\top x, t)^2 dQ(t)}_{=: g(x)} + \underbrace{\int_{[y, \infty)} F(\alpha^\top x, t) dQ(t)}_{=: h(x, y)}.$$

The function  $f$  above does not contribute to the entropy, and  $g$  does not depend on  $y$  and belongs to the class  $\mathcal{G}$ , for which we know from Assumptions 1 and 2 and Proposition 2 that  $\log(N_B(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathbb{P}^x})) \leq \tilde{C}/\varepsilon$  for a constant  $\tilde{C} > 0$ . In separate lemmas below, we show that the entropy of the functions of the form  $h$  above, with  $(F, \alpha) \in \mathcal{F}$ , is bounded from above by  $\tilde{D}/\varepsilon$  for some constant  $\tilde{D}$ . Let now  $[l, u]$  be an  $\varepsilon$ -bracket containing  $g$  and  $[L, U]$  an  $\varepsilon$ -bracket containing  $h$ . We interpret  $l, u$  as functions of  $(x, y)$  which are constant in  $y$ . Then the functions  $U + u + 1 - Q([t, \infty))$ ,  $L + l + 1 - Q([t, \infty))$  form a  $(2\varepsilon)$ -bracket containing  $\ell$ , because

$$\begin{aligned} \|U + u - L - l\|_{\mathbb{P}}^2 &= \int_{\mathbb{R}^d \times \mathbb{R}} \{(U - L)^2 + (u - l)^2 + 2(U - L)(u - l)\} d\mathbb{P}(x, y) \\ &\leq 2\varepsilon^2 + 2 \left( \int_{\mathbb{R}^d \times \mathbb{R}} (U - L)^2 d\mathbb{P}(x, y) \right)^{1/2} \left( \int_{\mathbb{R}^d \times \mathbb{R}} (u - l)^2 d\mathbb{P}(x, y) \right)^{1/2} \\ &\leq 4\varepsilon^2. \end{aligned}$$



Consequently, the number of  $\varepsilon$ -brackets required to cover  $\mathcal{L}$  is bounded from above by  $2(\tilde{C} + \tilde{D})/\varepsilon =: \kappa/\varepsilon$ , which yields the following bound on the entropy integral,

$$\tilde{J}(\delta, \mathcal{L}, \|\cdot\|_{\mathbb{P}}) = \int_0^\delta \sqrt{1 + \log N_B(\varepsilon, \mathcal{L}, \|\cdot\|)} d\varepsilon \leq \int_0^\delta 1 + \left(\frac{\kappa}{\varepsilon}\right)^{1/2} d\varepsilon = \delta + 2\kappa^{1/2}\delta^{1/2}.$$

Lemma 3.4.2 of van der Vaart and Wellner (1996) with  $M = 2$  implies

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathbb{G}_n \int_{\mathbb{R}} (1\{y \leq t\} - F(\alpha^\top x, t))^2 dQ(t) - \int_{\mathbb{R}} (1\{y \leq t\} - F_0(\alpha_0^\top x, t))^2 dQ(t) \right\|_{\mathcal{L}_\delta} \right] \\ & \leq (\delta + 2\kappa^{1/2}\delta^{1/2}) \left( 1 + 2\frac{\delta + 2\kappa^{1/2}\delta^{1/2}}{\delta^2 n^{1/2}} \right). \end{aligned}$$

Consequently, with

$$\begin{aligned} \tilde{\phi}_n(\delta) & := (\delta + 2\kappa^{1/2}\delta^{1/2}) \left( 1 + 2\frac{\delta + 2\kappa^{1/2}\delta^{1/2}}{\delta^2 n^{1/2}} \right) \\ \phi_n(\delta) & := \tilde{\phi}_n(\delta)/\phi_n(1), \end{aligned}$$

we have

$$\mathbb{E} \left[ \sup_{(F, \alpha): d((F, \alpha), (F_0, \alpha_0)) \leq \delta} |(\mathbb{M}_n - \mathbb{M})(F, \alpha) - (\mathbb{M}_n - \mathbb{M})(F_0, \alpha_0)| \right] \leq \phi_n(\delta),$$

and, for  $r_n = n^{2/3}$ ,  $r_n^2 \phi_n(1/r_n) \leq n^{1/2}$ . Since  $(\hat{F}_n, \hat{\alpha}_n)$  maximizes  $\mathbb{M}_n$  by definition, theorem 3.4.2 of van der Vaart and Wellner (1996) implies that  $n^{1/3} d((\hat{F}_n, \hat{\alpha}_n), (F_0, \alpha_0)) = O_p(1)$ . ■

For the entropy of the function class

$$\mathcal{M} = \left\{ h : \mathbb{R}^d \times \mathbb{R} \rightarrow [0, 1], h(x, y) = \int_{[y, \infty]} F(\alpha^\top x, t) dQ(t), (F, \alpha) \in \mathcal{F} \right\},$$

we begin with the simpler case that  $Q$  has finite support.

**Lemma 2.** Under Assumptions 1 and 2 (i), we have

$$\log(N_B(\varepsilon, \mathcal{M}, \|\cdot\|_{\mathbb{P}})) \leq \frac{\tilde{C} + p}{\varepsilon}, \quad \varepsilon \in (0, 1),$$

where  $\tilde{C} = \tilde{C}(d, R, \bar{p}_X)$  is the constant from Proposition 2, and  $p$  is the cardinality of the finite support of  $Q$ .

*Proof of Lemma 2.* Recall that  $Q$  puts all its mass on the points  $t_1 < \dots < t_p$ . Let  $l_i, u_i$ ,  $i = 1, \dots, N$  be  $\varepsilon$ -brackets covering  $\mathcal{G}$ , and let  $l_{i(j)}, u_{i(j)}$  be an  $\varepsilon$ -bracket containing  $x \mapsto F(\alpha^\top x, t_j)$ ,  $j = 1, \dots, m$ . Then,

$$L(x, y) := \sum_{j: t_j \geq y} Q(\{t_j\}) l_{i(j)}(x), \quad U(x, y) := \sum_{j: t_j \geq y} Q(\{t_j\}) u_{i(j)}(x),$$

are an  $\varepsilon$ -bracket containing  $h$ , because

$$\begin{aligned} \|U - L\|_{\mathbb{P}}^2 &= \int_{\mathbb{R}^d \times \mathbb{R}} \left( \sum_{j: t_j \geq y} Q(\{t_j\})(u_{i(j)}(x) - l_{i(j)}(x)) \right)^2 d\mathbb{P}(x, y) \\ &\leq \int_{\mathbb{R}^d \times \mathbb{R}} \sum_{j: t_j \geq y} Q(\{t_j\})(u_{i(j)}(x) - l_{i(j)}(x))^2 d\mathbb{P}(x, y) \\ &\leq \int_{\mathbb{R}^d \times \mathbb{R}} \sum_{j=1}^p Q(\{t_j\})(u_{i(j)}(x) - l_{i(j)}(x))^2 d\mathbb{P}(x, y) \\ &= \int_{\mathbb{R}^d} \sum_{j=1}^p Q(\{t_j\})(u_{i(j)}(x) - l_{i(j)}(x))^2 d\mathbb{P}(x) \\ &\leq \varepsilon^2. \end{aligned}$$

Moreover, there are  $pN$  functions of the form of  $L, U$ , corresponding to  $N$  choices for  $l_{i(j)}, u_{i(j)}$  and  $p$  choices of  $t_j$ . So for  $\varepsilon \in (0, 1)$ , we have

$$\log(N_B(\varepsilon, \mathcal{M}, \|\cdot\|_{\mathbb{P}})) \leq \tilde{C}/\varepsilon + p \leq \frac{\tilde{C} + p}{\varepsilon}.$$

■

For  $Q$  with Lebesgue continuous distribution, the entropy bound is as follows.

**Lemma 3.** *Under Assumptions 1 and 2 (ii), we have*

$$\log(N_B(\varepsilon, \mathcal{M}, \|\cdot\|_{\mathbb{P}})) \leq \frac{3\tilde{C} \max(1, c) + b - a + 1}{\varepsilon}, \quad \varepsilon \in (0, 1),$$

where  $\tilde{C} = \tilde{C}(d, R, \bar{p}_X)$  is the constant from Proposition 2.

*Proof of Lemma 3.* We assume that  $Q$  is Lebesgue continuous on  $[a, b]$  with density bounded from above by  $c < \infty$ . Discretize the interval  $[a, b]$  with a net of suitable size, namely, let  $N' = \lceil (b - a)/\varepsilon \rceil$  and define

$$t_j := a + (j - 1)(b - a)/N', \quad h_j(x) := \int_{[a, t_j]} F(\alpha^\top x, t) dQ(t), \quad j = 1, \dots, N' + 1.$$

The functions  $h_j$  are contained in the class  $\mathcal{G}$ . Let  $l_i, u_i, i = 1, \dots, N$  be  $\varepsilon$ -brackets for  $\mathcal{G}$ , such that  $N_B(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathbb{P}^X | Y=y}) \leq \tilde{C}/\varepsilon$  for all  $y \in \text{supp}(\mathbb{P}^Y)$ . For  $j = 1, \dots, m$  let  $i(j)$  be an index such that  $l_{i(j)} \leq h_j \leq u_{i(j)}$ , and for  $t \in (-\infty, b]$ , define

$$r(t) := \max\{j \in \{1, \dots, N' + 1\} : t_j \leq t\}, \quad s(t) := \begin{cases} \min(r(t) + 1, N' + 1), & t \geq a, \\ r(t), & t < a \end{cases},$$

and the functions

$$L(x, y) := l_{i(r(y))}(x), \quad U(x, y) := u_{i(s(y))}(x), \quad y \in (-\infty, b],$$

with  $L(x, y) := U(x, y) := 0$  for  $y > b$ . Note that there are at most  $N(N' + 1)$  such functions for all choices of  $r(y) \in \{1, \dots, N' + 1\}$  and  $i(j) \in \{1, \dots, N\}, j = 1, \dots, N' + 1$ . By construction, we have

$$L(x, y) \leq \int_{[y, \infty]} F(\alpha^\top x, t) dQ(t) \leq U(x, y), \quad y \in \mathbb{R}.$$

We show that  $L, U$  form an  $\varepsilon$ -bracket. First, notice that

$$\begin{aligned} \|U - L\|_{\mathbb{P}}^2 &= \int_{\mathbb{R}^d \times \mathbb{R}} (U(x, y) - L(x, y))^2 d\mathbb{P}(x, y) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}} (u_{i(s(y))}(x) - l_{i(r(y))}(x))^2 d\mathbb{P}(x, y) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^d} (u_{i(s(y))}(x) - l_{i(r(y))}(x))^2 d\mathbb{P}^{X|Y=y}(x) d\mathbb{P}^Y(y). \end{aligned}$$

We separate the outer integral into three parts. The lower part, over  $(-\infty, a)$ , satisfies

$$\begin{aligned} \int_{-\infty}^a \int_{\mathbb{R}^d} (u_{i(s(y))}(x) - l_{i(r(y))}(x))^2 d\mathbb{P}^{X|Y=y}(x) d\mathbb{P}^Y(y) &= \int_{-\infty}^a \int_{-\infty}^a (u_{i(1)}(x) - l_{i(1)}(x))^2 d\mathbb{P}^{X|Y=y}(y) d\mathbb{P}^Y(x) \\ &\leq \int_{-\infty}^a (u_{i(1)}(x) - l_{i(1)}(x))^2 d\mathbb{P}^{X|Y=y}(y), \end{aligned}$$

since  $l_{i(1)}, u_{i(1)}$  are  $\varepsilon$ -brackets. The upper part over  $(b, \infty)$  equals 0 because  $L(x, y) = U(x, y) = 0$  for  $y > b$ . For the middle part over  $[a, b]$ , let  $y$  in  $[t_j, t_{j+1})$ . Then,

$$\int_{\mathbb{R}^d} (u_{i(s(y))}(x) - l_{i(r(y))}(x))^2 d\mathbb{P}^{X|Y=y}(x) = \int_{\mathbb{R}^d} (u_{i(j+1)}(x) - l_{i(j)}(x))^2 d\mathbb{P}^{X|Y=y}(x),$$

and we expand the integrand as follows

$$\begin{aligned} &\int_{\mathbb{R}^d} (u_{i(j+1)}(x) - l_{i(j)}(x))^2 d\mathbb{P}^{X|Y=y}(x) \\ &= \int_{\mathbb{R}^d} (u_{i(j+1)}(x) - h_{j+1}(x) + h_{j+1}(x) - h_j(x) + h_j(x) - l_{i(j)}(x))^2 d\mathbb{P}^{X|Y=y}(x). \end{aligned} \tag{A1}$$

Since  $l(i(k)), i = 1, \dots, N$  are  $\varepsilon$ -brackets, we have

$$\int_{\mathbb{R}^d} (u_{i(j+1)}(x) - h_{j+1}(x))^2 + (h_j(x) - l_{i(j)}(x))^2 d\mathbb{P}^{X|Y=y}(x) \leq 2\varepsilon^2,$$

and also, because  $t_{j+1} - t_j \leq (b - a)/\varepsilon$

$$\begin{aligned} \int_{\mathbb{R}^d} (h_{j+1}(x) - h_j(x))^2 d\mathbb{P}^{X|Y=y}(x) &= \int_{\mathbb{R}^d} \left( \int_{(t_j, t_{j+1}]} F(\alpha^\top x, t) dQ(t) \right)^2 d\mathbb{P}^{X|Y=y}(x) \\ &\leq \int_{\mathbb{R}^d} \left( \int_{(t_j, t_{j+1}]} 1 dQ(t) \right)^2 d\mathbb{P}^{X|Y=y}(x) \end{aligned}$$

$$\begin{aligned} &\leq \int_{\mathbb{R}^d} (c\varepsilon)^2 d\mathbb{P}^{X|Y=y}(x) \\ &\leq (c\varepsilon)^2. \end{aligned}$$

The cross-terms can be bounded by applying the Cauchy–Schwarz inequality,

$$\begin{aligned} &\int_{\mathbb{R}^d} (u_{i(j+1)}(x) - h_{j+1}(x))(h_{j+1}(x) - h_j(x)) d\mathbb{P}^{X|Y=y}(x) \\ &\leq \left( \int_{\mathbb{R}^d} (u_{i(j+1)}(x) - h_{j+1}(x))^2 d\mathbb{P}^{X|Y=y}(x) \right)^{1/2} \left( \int_{\mathbb{R}^d} (h_{j+1}(x) - h_j(x))^2 d\mathbb{P}^{X|Y=y}(x) \right)^{1/2} \\ &\leq \max(1, c)\varepsilon^2, \end{aligned}$$

applying the bounds from above; the other cross terms are bounded in an analogous way. Hence,

$$\begin{aligned} \int_{[a,b]} \int_{\mathbb{R}^d} (u_{s(y)}(x) - l_{r(y)}(x))^2 d\mathbb{P}^{X|Y=y} d\mathbb{P}^Y &= \sum_{j=1}^{N'} \int_{[t_j, t_{j+1})} \int_{\mathbb{R}^d} (u_{s(y)}(x) - l_{r(y)}(x))^2 d\mathbb{P}^{X|Y=y} d\mathbb{P}^Y \\ &\leq \sum_{j=1}^{N'} \int_{[t_j, t_{j+1})} 9 \max(1, c^2)\varepsilon^2 d\mathbb{P}^Y \\ &\leq 9 \max(1, c^2)\varepsilon^2, \end{aligned}$$

where the factor 9 is due to the fact that one obtains three square terms and six cross-terms from expanding the square in (A1). So we have

$$\int_{\mathbb{R}} \int_{\mathbb{R}^d} (u_{s(y)}(x) - l_{r(y)}(x))^2 \leq 9 \max(1, c^2)\varepsilon^2.$$

Consequently, we obtain

$$\log(N_B(\varepsilon, \mathcal{M}, \|\cdot\|_{\mathbb{P}})) \leq \frac{3\tilde{C} \max(1, c)}{\varepsilon} + \frac{b - a + 1}{\varepsilon} \quad \blacksquare$$

## A.2 Proof of Proposition 2

*Proof.* Fix  $\varepsilon \in (0, 1)$ . By Lemma 21 of Feige and Schechtman (2002), we know that  $S_{d-1}$  can be partitioned into  $N$  subsets of equal size with diameter at most  $\varepsilon$  such that  $N \leq (A/\varepsilon^2)^d$ , for a universal constant  $A$ . Let  $\alpha_1, \dots, \alpha_N$  be points in these  $N$  subsets. Furthermore, from Theorem 2.7.5 of van der Vaart and Wellner (1996), we can find  $N' \leq \exp(K/\varepsilon)$  brackets  $[h_i^L, h_i^U], i = 1, \dots, N'$  with respect to the norm  $\|\cdot\|_{[-2R, 2R]}$ .

Let  $g \in \mathcal{G}$ . Then,  $g(x) = h(\alpha^\top x)$  for some  $\alpha \in S_{d-1}$  and  $h \in \mathcal{H}$ . Let  $i \in \{1, \dots, N\}$  and  $j \in \{1, \dots, N'\}$  such that  $\|\alpha - \alpha_i\| \leq \varepsilon^2$  and  $h_j^L \leq h \leq h_j^U$ . Now, it follows from the Cauchy–Schwarz inequality that

$$\alpha^\top x = (\alpha - \alpha_i)^\top x + \alpha_i^\top x \in [\alpha_i^\top x - \varepsilon^2 R, \alpha_i^\top x + \varepsilon^2 R] \subset [-2R, 2R].$$

By monotonicity of  $h$  this implies that

$$h(\alpha_i^\top x + \varepsilon^2 R) \leq h(\alpha^\top x) \leq h(\alpha_i^\top x - \varepsilon^2 R),$$

and hence

$$h_j^L(\alpha_i^\top x + \varepsilon^2 R) \leq h(\alpha^\top x) \leq h_j^U(\alpha_i^\top x - \varepsilon^2 R). \quad (\text{A2})$$

Now, using the Minkowski inequality, we have that

$$\begin{aligned} & \left( \int_{\mathcal{X}} \left\{ h_j^U(\alpha_i^\top x - \varepsilon^2 R) - h_j^L(\alpha_i^\top x + \varepsilon^2 R) \right\}^2 dx \right)^{1/2} \\ & \leq \left( \int_{\mathcal{X}} \left\{ h_j^U(\alpha_i^\top x - \varepsilon^2 R) - h(\alpha_i^\top x - \varepsilon^2 R) \right\}^2 dx \right)^{1/2} \\ & \quad + \left( \int_{\mathcal{X}} \left\{ h(\alpha_i^\top x - \varepsilon^2 R) - h(\alpha_i^\top x + \varepsilon^2 R) \right\}^2 dx \right)^{1/2} \\ & \quad + \left( \int_{\mathcal{X}} \left\{ h_j^L(\alpha_i^\top x + \varepsilon^2 R) - h(\alpha_i^\top x + \varepsilon^2 R) \right\}^2 dx \right)^{1/2} \\ & =: I_1 + I_2 + I_3. \end{aligned}$$

Note that for any  $\alpha = (\alpha^{(1)}, \dots, \alpha^{(d)}) \in \mathcal{S}_{d-1}$ , there exists  $j \in \{1, \dots, d\}$  such that  $|\alpha^{(j)}| \geq 1/\sqrt{d}$ . Without loss of generality we assume that  $|\alpha^{(1)}| \geq 1/\sqrt{d}$ . Consider the change of variable  $\varphi(x) = t$  where

$$t_1 = \alpha_1^\top x - \varepsilon^2 R \text{ and } t_j = x_j, \text{ for } j = 2, \dots, d.$$

Then,

$$\begin{aligned} I_1 & \leq \left( \int_{\varphi(\mathcal{X})} \left\{ h_j^U(t_1) - h(t_1) \right\}^2 dt \frac{1}{\alpha_j^{(1)}} \right)^{1/2} \\ & \leq \left( \sqrt{d} \int_{-2R}^R \int_{-R}^R \dots \int_{-R}^R \left\{ h_j^U(t_1) - h(t_1) \right\}^2 dt \right)^{1/2} \\ & \leq d^{1/4} (2R)^{(d-1)/2} \left( \int_{-2R}^R \left\{ h_j^U(t_1) - h(t_1) \right\}^2 dt_1 \right)^{1/2} \\ & \leq d^{1/4} (2R)^{(d-1)/2} \left( \int_{-2R}^{2R} \left\{ h_j^U(t_1) - h(t_1) \right\}^2 dt_1 \right)^{1/2} \\ & \leq d^{1/4} (2R)^{(d-1)/2} \varepsilon, \end{aligned}$$

where above used that  $t_1 = \alpha_j^\top x - \varepsilon^2 R \in [-2R, R]$  for all  $x \in \mathcal{X}$ . Using a similar reasoning, we can bound  $I_3$  by the same constant. Now, we turn to  $I_2$ . With the same change of variable, we have that

$$\begin{aligned}
& \left( \int_{\mathcal{X}} \{h(\alpha_i^\top x - \varepsilon^2 R) - h(\alpha_i^\top x + \varepsilon^2 R)\}^2 dx \right)^{1/2} \\
& \leq d^{1/4} (2R)^{(d-1)/2} \left( \int_{-2R}^R \{h(z) - h(z + 2\varepsilon^2 R)\}^2 dz \right)^{1/2} \\
& \leq d^{1/4} (2R)^{(d-1)/2} \left( \int_{-2R}^R \{h(z) - h(z + 2\varepsilon^2 R)\} dz \right)^{1/2},
\end{aligned}$$

using monotonicity of  $h$  and that  $h(z) - h(z + 2\varepsilon^2 R) \in [0, 1]$  for all  $z \in [-2R, R]$ . Now,

$$\begin{aligned}
\int_{-2R}^R \{h(z) - h(z + 2\varepsilon^2 R)\} dz &= \int_{-2R}^R h(z) dz - \int_{-2R}^R h(z + 2\varepsilon^2 R) dz \\
&= \int_{-2R}^R h(z) dz - \int_{-2R+2\varepsilon^2 R}^{R+2\varepsilon^2 R} h(z) dz \\
&= \int_{-2R}^{-2R+2\varepsilon^2 R} h(z) dz - \int_R^{R+2\varepsilon^2 R} h(z) dz \\
&\leq 2\varepsilon^2 R.
\end{aligned}$$

Thus,

$$\begin{aligned}
\left( \int_{\mathcal{X}} \left\{ h_j^U(\alpha_i^\top x - \varepsilon^2 R) - h_j^L(\alpha_i^\top x + \varepsilon^2 R) \right\}^2 dx \right)^{1/2} &\leq d^{1/4} (2R)^{(d-1)/2} \varepsilon + d^{1/4} (2R)^{(d-1)/2} \sqrt{2} \sqrt{R} \varepsilon \\
&\leq d^{1/4} (2R)^{(d-1)/2} (1 + \sqrt{R}) \varepsilon.
\end{aligned}$$

If we put  $B = 2d^{1/4} (2R)^{(d-1)/2} (1 + \sqrt{R})$ , then the previous calculations and the inequality (A2) imply that

$$N_B(B\varepsilon, \mathcal{G}, \|\cdot\|_{\mathcal{X}}) \leq NN',$$

and hence

$$\begin{aligned}
\log(N_B(B\varepsilon, \mathcal{G}, \|\cdot\|_{\mathcal{X}})) &\leq \log N + \log N' \\
&\leq d \log \frac{A}{\varepsilon^2} + \frac{2K\sqrt{R}}{\varepsilon} \\
&= 2d \log \frac{\sqrt{A}}{\varepsilon} + \frac{2K\sqrt{R}}{\varepsilon} \\
&\leq \frac{2(d\sqrt{A} + K\sqrt{R})}{\varepsilon},
\end{aligned}$$

which in turn implies that

$$\log(N_B(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathcal{X}})) \leq \frac{2^{(d+1)/2} d^{1/4} R^{(d-1)/2} (1 + \sqrt{R}) (d\sqrt{A} + 2K\sqrt{R})}{\varepsilon}.$$

Finally, since the Lebesgue density of  $\mu$  is bounded from above by  $C$ , the previous bound implies

$$\log(N_B(\varepsilon, \mathcal{G}, \|\cdot\|_\mu)) \leq \frac{2^{(d+1)/2} d^{1/4} R^{(d-1)/2} (1 + \sqrt{R})(d\sqrt{A} + 2K\sqrt{R})C}{\varepsilon}.$$

■

### A.3 Proof of Theorem 2

*Proof.* For simplicity of notation, index the subsequence by  $n$ , and choose an  $\omega$  in the underlying probability space such that (7) holds true. Recall that  $\hat{F}_n$  is non increasing in the first entry and nondecreasing in the second entry for every  $n$ . Lemma 2.5. in van der Vaart (1998) can be adapted to this case. Therefore  $\hat{F}_n$  converges pointwise along a subsequence to a bivariate function  $G$  at each point of continuity of  $G$  that lies in  $\text{supp}(Q)$ . The limit  $G$  has the property that  $G(\cdot, t)$  is left continuous and nonincreasing for each  $t \in \text{supp}(Q)$  and  $G(z, \cdot)$  nondecreasing for every  $z$ . Furthermore,  $\hat{\alpha}_n \in \mathcal{S}_{d-1}$  is a sequence in a compact space and hence converges along a further subsequence to  $\beta_0$  in the Euclidean distance.

Our goal is to show that  $G = F_0$  and  $\alpha_0 = \beta_0$ . Recall that if the  $L_2$  distance between two functions is zero then they coincide almost surely. We have

$$\begin{aligned} & \int_{\mathcal{X} \times \mathbb{R}} (G(\beta_0^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) \\ &= \int_{\mathcal{X} \times \mathbb{R}} (G(\beta_0^\top x, t) - G(\hat{\alpha}_n^\top x, t) + \hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t) + G(\hat{\alpha}_n^\top x, t) - \hat{F}_n(\hat{\alpha}_n^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) \\ &\leq 3I_{n,1} + 3I_{n,2} + 3I_{n,3}, \end{aligned}$$

by applying the Cauchy–Schwarz inequality, where

$$\begin{aligned} I_{n,1} &= \int_{\mathcal{X} \times \mathbb{R}} (G(\beta_0^\top x, t) - G(\hat{\alpha}_n^\top x, t))^2 d\mathbb{P}^X(x) dQ(t), \\ I_{n,2} &= \int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t), \\ I_{n,3} &= \int_{\mathcal{X} \times \mathbb{R}} (G(\hat{\alpha}_n^\top x, t) - \hat{F}_n(\hat{\alpha}_n^\top x, t))^2 d\mathbb{P}^X(x) dQ(t). \end{aligned}$$

We show that for  $n \rightarrow \infty$  the terms  $I_{n,1}$ ,  $I_{n,2}$ ,  $I_{n,3}$  converge to zero almost surely, so  $G = F_0$  almost surely.

Recall that  $\hat{\alpha}_n$  converges to  $\beta_0$ . Therefore, at all continuity points of  $G_0$  we have that  $G_0(\hat{\alpha}_n^\top x, t)$  converges to  $G_0(\beta_0^\top x, t)$ . Note that  $G_0$  is bounded and monotone in both variables. Lavrič (1993) shows that the set of all discontinuity points of the bivariate, monotone function  $G$  may not be countable but has Lebesgue measure 0. When using that both  $Q$  and  $\mathbb{P}^X$  are equivalent to the Lebesgue measure, under our assumptions, we have that  $I_{n,1} \rightarrow 0$  by Lebesgue's dominated convergence Theorem. The second integral  $I_{n,2}$  converges to 0 directly by (7). Finally, we rewrite the third integral to

$$I_{n,3} = \int_{\mathcal{X} \times \mathbb{R}} (G(z, t) - \hat{F}_n(z, t))^2 dQ_n(z) dQ(t),$$

where  $\mathbb{Q}_n$  denotes the distribution of  $\hat{\alpha}_n^\top X$  and  $X$  is a random variable that is independent of the data, but has distribution  $\mathbb{P}^X$ . As at each point of continuity of  $G$ , the function  $\hat{F}_n$  converges to  $G$  and the set of discontinuity points of  $G$  has Lebesgue measure 0, Assumption 4 and Lebesgue's dominated convergence theorem imply that  $I_{n,3} \rightarrow 0$ .

If necessary, modify  $G$  to not have discontinuity points at the boundary. By Proposition 3 it follows that  $\beta_0 = \alpha_0$  and  $G = F_0$  everywhere on  $C_{\alpha_0} \times \text{supp}(Q)$ . As we have found almost sure convergence along a subsequence, we follow that the statements hold true for convergence in probability. ■

#### A.4 Proof of Theorem 3

*Proof.* We apply Lemma 2.5. from Murphy, van der Vaart, and Wellner (1999). Rewrite the integrated error as follows,

$$\begin{aligned} \int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) &= \int_{\mathcal{X} \times \mathbb{R}} (G_1(x, t) + G_2(x, t))^2 d\mathbb{P}^X(x) dQ(t) \\ &= \mathbb{E}[(G_1(X, T) + G_2(X, T))^2], \end{aligned}$$

where the expectation is a shorthand notation of integrating with respect to a random variable  $(X, T)$  whose distribution is the product measure of  $\mathbb{P}^X$  and  $Q$ . The functions  $G_1$  and  $G_2$  are  $G_1(x, t) = \hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\hat{\alpha}_n^\top x, t) = \tilde{G}_1(\hat{\alpha}_n^\top x, t)$  and  $G_2(x, t) = F_0(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t)$ . The Cauchy-Schwarz inequality and the tower property of conditional expectations yield

$$\begin{aligned} \mathbb{E}[G_1(X, T)G_2(X, T)]^2 &= \mathbb{E}[\tilde{G}_1(\hat{\alpha}_n^\top X, T)G_2(X, T)]^2 \\ &= \mathbb{E}[\tilde{G}_1(\hat{\alpha}_n^\top X, T) \mathbb{E}[G_2(X, T) | \hat{\alpha}_n^\top X, T]]^2 \\ &\leq \mathbb{E}[\tilde{G}_1(\hat{\alpha}_n^\top X, T)^2] \mathbb{E}[\mathbb{E}[G_2(X, T) | \hat{\alpha}_n^\top X, T]^2] \\ &= c_n \mathbb{E}[G_1(X, T)^2] \mathbb{E}[G_2(X, T)^2], \end{aligned}$$

where

$$c_n = \frac{\mathbb{E}[\mathbb{E}[G_2(X, T) | \hat{\alpha}_n^\top X, T]^2]}{\mathbb{E}[G_2(X, T)^2]} = \frac{\mathbb{E}[(F_0(\hat{\alpha}_n^\top X, T) - \mathbb{E}[F_0(\alpha_0^\top X, T) | \hat{\alpha}_n^\top X, T])^2]}{\mathbb{E}[(F_0(\hat{\alpha}_n^\top X, T) - F_0(\alpha_0^\top X, T))^2]}.$$

If  $c_n < 1$  it follows by Murphy et al. (1999) that

$$\begin{aligned} &\int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) \\ &\geq (1 - \sqrt{c_n}) (\mathbb{E}[(\hat{F}_n(\hat{\alpha}_n^\top X, T) - F_0(\hat{\alpha}_n^\top X, T))^2] + \mathbb{E}[(F_0(\hat{\alpha}_n^\top X, T) - F_0(\alpha_0^\top X, T))^2]). \end{aligned} \quad (\text{A3})$$

We now prove that there exists a  $c < 1$  such that from any subsequence  $(n_k)_{k \in \mathbb{N}}$ , there exists a subsequence  $(n_{k_l})_{l \in \mathbb{N}}$  along which  $\limsup_{l \rightarrow \infty} c_{n_{k_l}} \leq c < 1$  almost surely. This shows that  $(1 - \sqrt{c_n})^{-1} = O_P(1)$ .

To prove the claim, consider an arbitrary subsequence. For simplicity of notation, index it with  $n$ . Define  $u_n = \|\hat{\alpha}_n - \alpha_0\|$  and  $\gamma_n = (\hat{\alpha}_n - \alpha_0)/u_n$ . As  $\|\gamma_n\| = 1$  and  $S_{d-1}$



is compact,  $\gamma_n$  converges to some  $\gamma_0 \in \mathcal{S}_{d-1}$  along a subsequence. Recall that  $\hat{\alpha}_n$  converges to  $\alpha_0$  in probability. Therefore, we can extract a further subsequence along which the convergence from  $\hat{\alpha}_n$  to  $\alpha_0$  and from  $\gamma_n$  to  $\gamma_0$  happens almost surely. To make notation less cumbersome we index this subsequence again by  $n$ . Fix an event  $\omega$  in the underlying probability space such that  $\hat{\alpha}_0 \rightarrow \alpha_0$  and  $\gamma_n \rightarrow \gamma_0$ , so that we can consider  $\hat{\alpha}_n$  and  $\gamma_n$  as nonrandom.

By Assumption 3, for every  $t \in \mathbb{R}$  the map  $F_0(\cdot, t)$  is continuously differentiable on  $C_{\alpha_0}$ . Extend the function  $F_0(\cdot, t)$  such that it is bounded and continuously differentiable on  $\mathbb{R}$  and the partial derivative  $z \mapsto F_0^{(1)}(z, t)$  is bounded on  $\mathbb{R}^2$ . By Taylor's Theorem we have that for  $x \in \mathcal{X}$  and  $t \in \mathbb{R}$ ,

$$F_0(\alpha_0^\top x, t) = F_0(\hat{\alpha}_n^\top x, t) + F_0^{(1)}(\hat{\alpha}_n^\top x, t)(\alpha_0 - \hat{\alpha}_n)^\top x + o(u_n). \quad (\text{A4})$$

Thus the numerator of  $c_n$  becomes

$$\begin{aligned} & \mathbb{E} \left[ \mathbb{E} [F_0(\hat{\alpha}_n^\top X, T) - F_0(\alpha_0^\top X, T) \mid \hat{\alpha}_n^\top X, T]^2 \right] \\ &= \mathbb{E} \left[ \mathbb{E} [F_0^{(1)}(\hat{\alpha}_n^\top X, T)(\alpha_0 - \hat{\alpha}_n)^\top X + o(u_n) \mid \hat{\alpha}_n^\top X, T]^2 \right] \\ &= \mathbb{E} \left[ \mathbb{E} [F_0^{(1)}(\hat{\alpha}_n^\top X, T)(\alpha_0 - \hat{\alpha}_n)^\top X \mid \hat{\alpha}_n^\top X, T]^2 \right] + o(u_n^2), \end{aligned}$$

as the mixed term can be controlled by

$$2o(u_n) \left| \mathbb{E} \left[ F_0^{(1)}(\hat{\alpha}_n^\top X, T)(\alpha_0 - \hat{\alpha}_n)^\top X \right] \right| = o(u_n^2).$$

This is because the partial derivative  $z \mapsto F_0^{(1)}(z, t)$  is bounded. Similarly the denominator becomes

$$\mathbb{E} \left[ (F_0(\hat{\alpha}_n^\top X, T) - F_0(\alpha_0^\top X, T))^2 \right] = \mathbb{E} \left[ (F_0^{(1)}(\hat{\alpha}_n^\top X, T)(\alpha_0 - \hat{\alpha}_n)^\top X)^2 \right] + o(u_n^2).$$

We rewrite

$$c_n = \frac{\mathbb{E} \left[ (F_0^{(1)}(\hat{\alpha}_n^\top X, T)\gamma_n^\top \mathbb{E}[X \mid \hat{\alpha}_n^\top X, T])^2 \right] + o(1)}{\mathbb{E} \left[ (F_0^{(1)}(\hat{\alpha}_n^\top X, T)\gamma_n^\top X)^2 \right] + o(1)}.$$

By Lemma 9.1 in the supplement of Balabdaoui, Durot, and Jankowski (2019) we have that  $\mathbb{E}[X \mid \hat{\alpha}_n^\top X, T] \rightarrow \mathbb{E}[X \mid \alpha_0^\top X, T]$  almost surely. By Lebesgue's dominated convergence theorem and the continuity of  $F_0^{(1)}(\cdot, t)$ , we have that

$$\begin{aligned} \limsup_{n \rightarrow \infty} c_n &= \frac{\mathbb{E} \left[ (F_0^{(1)}(\alpha_0^\top X, T)\gamma_0^\top \mathbb{E}[X \mid \alpha_0^\top X, T])^2 \right]}{\mathbb{E} \left[ (F_0^{(1)}(\alpha_0^\top X, T)\gamma_0^\top X)^2 \right]} \\ &= \frac{\gamma_0^\top \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 \mathbb{E}[X \mid \alpha_0^\top X, T] \mathbb{E}[X \mid \alpha_0^\top X, T]^\top \right] \gamma_0}{\gamma_0^\top \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 X X^\top \right] \gamma_0}. \end{aligned}$$

As  $\hat{\alpha}_n \in S_{d-1}$ , it follows that  $1 = \|\hat{\alpha}_n\| = \|\alpha_0 + u_n \gamma_n\| = \|\alpha_0\| + u_n^2 + 2u_n \langle \alpha_0, \gamma_n \rangle$  and thus  $2\langle \alpha_0, \gamma_n \rangle = -u_n \rightarrow 0$  and  $\langle \alpha_0, \gamma_0 \rangle = 0$ . Write

$$c = \sup_{\gamma \in S_{d-1} : \langle \alpha_0, \gamma \rangle = 0} \frac{\gamma^\top \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 \mathbb{E}[X | \alpha_0^\top X, T] \mathbb{E}[X | \alpha_0^\top X, T]^\top \right] \gamma}{\gamma^\top \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 X X^\top \right] \gamma}.$$

Then, we have that  $\lim_{n \rightarrow \infty} c_n \leq c$  where  $c$  does not depend on the chosen path  $\omega$ . It remains to prove that  $c < 1$ . We first expand the matrix in the denominator and get

$$\begin{aligned} \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 X X^\top \right] &= \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 \mathbb{E}[X | \alpha_0^\top X, T] \mathbb{E}[X | \alpha_0^\top X, T]^\top \right] \\ &\quad + \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 (X - \mathbb{E}[X | \alpha_0^\top X, T]) (X - \mathbb{E}[X | \alpha_0^\top X, T])^\top \right] \\ &:= A + B. \end{aligned}$$

Note that  $\gamma_0^\top A \gamma_0$  equals the numerator in the expression of  $c$ . Consider some  $\gamma \in S_{d-1}$  with  $\langle \alpha_0, \gamma \rangle = 0$ . Define the  $2 \times d$  matrix  $A_0$  to have first row equal to  $\alpha_0^\top$  and second row equal  $\gamma^\top$  and  $Z = (Z_1, Z_2) = A_0 X$ . Since  $X$  has a density that is positive on  $\mathcal{X}$ , the variable  $Z$  admits a density that is positive on the set  $\mathcal{Z} := \{A_0 x : x \in \mathcal{X}\}$ , which has nonempty interior. Then,

$$\begin{aligned} \gamma^\top \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 (X - \mathbb{E}[X | \alpha_0^\top X, T]) (X - \mathbb{E}[X | \alpha_0^\top X, T])^\top \right] \gamma \\ = \mathbb{E} \left[ F_0^{(1)}(\alpha_0^\top X, T)^2 (\gamma^\top X - \mathbb{E}[\gamma^\top X | \alpha_0^\top X, T])^2 \right], \end{aligned}$$

is equal to zero if and only if  $\gamma^\top X = \mathbb{E}[\gamma^\top X | \alpha_0^\top X, T]$  almost surely or equivalently  $Z_2 = \mathbb{E}[Z_2 | Z_1]$  almost surely. This would mean that the distribution of  $Z$  is concentrated on a one-dimensional subspace. This contradicts the fact that the density of  $Z$  with respect to the Lebesgue measure is positive on  $\mathcal{Z}$ . It follows that  $\gamma^\top B \gamma > 0$  and thus  $c < 1$ . This proves the claim. In integral notation, it follows from (A3) that

$$\begin{aligned} &\int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) \\ &\geq (1 - \sqrt{c_n}) \left( \int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\hat{\alpha}_n^\top x, t))^2 d\mathbb{P}^X dQ(t) \right. \\ &\quad \left. + \int_{\mathcal{X} \times \mathbb{R}} (F_0(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) \right) \\ &\geq (1 - \sqrt{c_n}) \int_{\mathcal{X} \times \mathbb{R}} (F_0(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X dQ(t) \\ &= (1 - \sqrt{c_n}) \int_{\mathcal{X} \times \mathbb{R}} \left( F_0^{(1)}(\hat{\alpha}_n^\top x, t) (\alpha_0 - \hat{\alpha}_n)^\top x + o(u_n) \right)^2 d\mathbb{P}^X dQ(t) \\ &\geq c' \|\hat{\alpha}_n - \alpha_0\|^2 \inf_{\beta \in S_{d-1}} \int_{\mathcal{X} \times \mathbb{R}} (\beta^\top x)^2 d\mathbb{P}^X(x) dQ(t), \end{aligned}$$

for some  $c' > 0$  by the previous observations, for  $n$  large enough. Note that the infimum above is strictly positive and achieved for some  $\beta$ , as the function

$\beta \mapsto \int_{\mathcal{X} \times \mathbb{R}} (\beta^\top x)^2 \mathbb{P}^X(x) dQ(t)$  is continuous,  $S_{d-1}$  is compact and the density  $p_X$  is bounded away from zero. Thus, there exists  $K > 0$  such that

$$\|\hat{\alpha}_n - \alpha_0\|^2 \leq K \int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X dQ(t) = O_P(n^{-2/3})$$

for large  $n$  and almost surely.

We turn to the second part. Recall that the density of  $\hat{\alpha}_n^\top X$  is bounded from below by  $\underline{q} > 0$ , so

$$\begin{aligned} \int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\hat{\alpha}_n^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) &\geq \underline{q} \int_{C_{\hat{\alpha}_n} \times \mathbb{R}} (\hat{F}_n(z, t) - F_0(z, t))^2 dz dQ(t) \\ &\geq \underline{q} \int_{\mathbb{R}} \int_{\underline{c} + v_n}^{\bar{c} - v_n} (\hat{F}_n(z, t) - F_0(z, t))^2 dz dQ(t) \end{aligned} \quad (\text{A5})$$

with probability tending to one for  $n \rightarrow \infty$ , using the definition of  $v_n$  and that  $\|\hat{\alpha}_n - \alpha_0\| = O_P(n^{-1/3})$ . The left-hand side of (A5) can be bounded from above by

$$\begin{aligned} \int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\hat{\alpha}_n^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) &\leq 2 \int_{\mathcal{X} \times \mathbb{R}} (\hat{F}_n(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) \\ &\quad + 2 \int_{\mathcal{X} \times \mathbb{R}} (F_0(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t). \end{aligned}$$

The first term is bounded  $O_P(n^{-2/3})$  by Theorem 1 and the second term can be handled due to the fact that the absolute value of the partial derivative  $F_0^{(1)}(z, t)$  is bounded by  $K := \sup_{t \in \text{supp}(Q)} K_t$ ; this yields

$$\begin{aligned} \int_{\mathcal{X} \times \mathbb{R}} (F_0(\hat{\alpha}_n^\top x, t) - F_0(\alpha_0^\top x, t))^2 d\mathbb{P}^X(x) dQ(t) &\leq K^2 \int_{\mathcal{X} \times \mathbb{R}} ((\alpha_0 - \hat{\alpha}_n)^\top x)^2 d\mathbb{P}^X(x) dQ(t) \\ &\leq K^2 R^2 \|\alpha_0 - \hat{\alpha}_n\|^2 \\ &= O_P(n^{-2/3}). \end{aligned}$$

■

## A.5 Identifiability

The identifiability result in this section is a direct adaptation of Theorem 5.1 of Balabdaoui, Durot, and Jankowski (2019).

**Proposition 3.** *Assume  $\mathcal{X} \subset \mathbb{R}^d$  is convex and has at least one interior point. Furthermore, assume  $X$  has a density with respect to the Lebesgue measure which is strictly positive on  $\mathcal{X}$ . Suppose that for each  $t \in \text{supp}(Q)$  the function  $F_0(\cdot, t)$  is left-continuous (or right-continuous), non constant and does not have discontinuity points on the boundary of  $C_{\alpha_0}$ . Then  $(F_0, \alpha_0)$  is identifiable.*

*Proof.* We will prove the left-continuous case; the right-continuous case can be treated with the same arguments. Consider pairs  $(F, \alpha), (H, \beta) \in \mathcal{F}$  having the property that for each  $t \in \text{supp}(Q)$ , the functions  $F(\cdot, t)$  on  $C_\alpha$  and  $H(\cdot, t)$  are left-continuous on  $C_\beta$ ,

non constant and do not have discontinuity points on the boundary of their domain. Assume

$$F(\alpha^T x, t) = H(\beta^T x, t)$$

for  $\mathbb{P}^X$  almost all  $x \in \mathbb{R}^d$ . Fix  $t_0 \in \mathbb{R}$  and define  $f = F(\cdot, t_0)$  and  $h = H(\cdot, t_0)$ . By assumption we have  $f(\alpha^T x) = h(\beta^T x)$  for almost every  $x \in \mathcal{X}$ . As  $f, h$  are left-continuous, this holds for all points in the interior of  $\mathcal{X}$ . If we prove  $\alpha = \beta$  we can follow that  $f = h$  on the interior of  $C_\alpha = C_\beta$ . As there are no discontinuity points on the boundary,  $f = h$  holds everywhere on  $C_\alpha$  and finally, so  $F = H$  on  $C_\alpha \times \mathbb{R}$ . Therefore, it suffices to show  $\alpha = \beta$ .

As  $\mathcal{X}$  is convex, for  $L > 0$  small enough we can find an open ball  $B_L$  of radius  $L$  contained in  $\mathcal{X}$  such that  $x \mapsto f(\alpha^T x)$  is non constant and

$$f(\alpha^T x) = h(\beta^T x) \tag{A6}$$

for every  $x \in B_L$ . Without loss of generality, we assume that  $B_L$  is centered at the origin — if necessary, replace  $f(z)$  with  $f(z - \alpha^T x_0)$  and  $h(z)$  with  $h(z - \beta^T x_0)$ , where  $x_0$  is the center of a ball with the desired properties. We first show  $\beta \in \{\alpha, -\alpha\}$  and then  $\beta \neq -\alpha$ .

Assume for a contradiction that  $\beta \notin \{\alpha, -\alpha\}$ . Then  $\alpha$  and  $\beta$  are linearly independent and by the Cauchy-Schwarz inequality for  $v = \beta - \alpha$ , it holds  $v^T \alpha = \beta^T \alpha - 1 < 0$  and  $v^T \beta > 0$ . Using the monotonicity of  $f$  and  $h$  it follows that

$$\begin{aligned} f(z) &= f(\alpha^T(z\alpha)) = h(\beta^T(z\alpha)) = h(\alpha^T(z\alpha) + v^T(z\alpha)) \geq h(z), \\ h(z) &= h(\beta^T(z\beta)) = f(\alpha^T(z\beta)) = f(\beta^T(z\beta) - v^T(z\beta)) \geq f(z), \end{aligned}$$

for each  $z \in [0, L)$  and so  $f(z) = h(z)$  on  $[0, L)$ . By the same arguments one shows  $f(-z) = h(-z)$  on  $[0, L)$ , and so  $f = h$  on  $(-L, L)$ . Hence, for  $x \in B_L$  we have

$$f(\alpha^T x) = f(\beta^T x). \tag{A7}$$

Since  $x \mapsto f(\alpha^T x)$  is non-constant on  $B_L$ , there exists a point  $b \in (-L, L)$  of strict decrease, so one of the following two conditions must hold,

$$f(b) > f(b + \epsilon), \quad \epsilon \in (0, L - b); \tag{A8}$$

$$f(b - \epsilon) > f(b), \quad \epsilon \in (0, L + b). \tag{A9}$$

The ball  $B_L$  can be chosen in such a way that  $b \neq 0$ . In the case (A8), if  $b > 0$  we can choose  $\epsilon$  small enough such that for  $x := (b + \epsilon)\beta$  it holds  $x \in B_L$  and  $\alpha^T x \leq b$ , since  $\alpha^T \beta < 1$ . Then, we have

$$f(\alpha^T x) \geq f(b) > f(b + \epsilon) = f(\beta^T x),$$

which contradicts (A7). If  $b < 0$  we let  $x = b\alpha$  and choose  $\epsilon$  sufficiently small such that  $b + \epsilon < 0$  and  $\beta^T x = b\beta^T \alpha \geq b + \epsilon$ . Then,

$$f(\alpha^T x) = f(b) > f(b + \epsilon) \geq f(\beta^T x),$$

which contradicts (A7), again. The second case, (A9), can be proven with similar ideas. Namely, if  $b < 0$  choose  $x = (b - \epsilon)\beta$  and  $\epsilon$  small enough such that  $\alpha^T \beta(b - \epsilon) \geq b$ . Then,

$$f(\beta^T x) = f(b - \epsilon) \geq f(b) \geq f(\alpha^T x)$$

which contradicts (A7). If  $b > 0$  choose  $x = b\alpha$  and  $\epsilon$  small enough such that  $b\alpha^T \beta \leq b - \epsilon$ . Then,

$$f(\beta^T x) \geq f(b - \epsilon) \geq f(b) = f(\alpha^T x)$$

which contradicts (A7). This proves  $\beta \in \{-\alpha, \alpha\}$ .

Finally, we assume for a contradiction that  $\beta = -\alpha$ . For  $z \in [0, L]$  we have

$$f(z) = f(\alpha^T(z\alpha)) = h(\beta(z\alpha)) = h(-z),$$

by (A6). With the same argument one shows

$$h(z) = h(\beta(z\beta)) = f(\alpha(z\beta)) = f(-a).$$

Thus by monotonicity of  $h$  we have for  $z \in [0, L]$ ,

$$f(z) = h(-z) \geq h(z) = f(-z),$$

and so  $f(z) = f(-z)$  on  $[0, L]$ . As  $f$  is also nonincreasing, we conclude that  $f$  is constant on  $(-L, L)$ , a contradiction. Consequently,  $\alpha = \beta$  and Proposition 3 follows. ■

## A.6 Proof of Lemma 1

*Proof.* Replacing  $Y_1, \dots, Y_n$  by  $f(Y_1), \dots, f(Y_n)$  in (4) and the fact that  $1\{Y_i \leq Y_j\} = 1\{f(Y_i) \leq f(Y_j)\}$  almost surely for  $i, j = 1, \dots, n$  imply  $L_n(\mathbb{P}_n^Y; \tilde{F}_{n, \hat{\alpha}_n}, \hat{\alpha}_n) = L_n(\mathbb{P}_n^{f(Y)}; \tilde{F}_{n, \hat{\alpha}_n}, \hat{\alpha}_n)$ , which also yields the statement about the minimizers in (i). Part (ii) holds by definition of  $t_i, i = 1, \dots, n, \tilde{F}_{n, \hat{\alpha}_n}$ , and  $\tilde{F}_0$ . ■