Bernoulli **26**(2), 2020, 1587–1618
https://doi.org/10.3150/19-BEJ1183

Efficient estimation in single index models through smoothing splines

ARUN K. KUCHIBHOTLA[1] and ROHIT K. PATRA[2]

[1]*University of Pennsylvania, Philadelphia, USA. E-mail: arunku@upenn.edu*
[2]*University of Florida, Gainesville, USA. E-mail: rohitpatra@ufl.edu*

We consider estimation and inference in a single index regression model with an unknown but smooth link function. In contrast to the standard approach of using kernels or regression splines, we use smoothing splines to estimate the smooth link function. We develop a method to compute the penalized least squares estimators (PLSEs) of the parametric and the nonparametric components given independent and identically distributed (i.i.d.) data. We prove the consistency and find the rates of convergence of the estimators. We establish asymptotic normality under mild assumption and prove asymptotic efficiency of the parametric component under homoscedastic errors. A finite sample simulation corroborates our asymptotic theory. We also analyze a car mileage data set and a Ozone concentration data set. The identifiability and existence of the PLSEs are also investigated.

Keywords: least favorable submodel; penalized least squares; semiparametric model

1. Introduction

Consider a regression model where one observes i.i.d. copies of the predictor $X \in \mathbb{R}^d$ and the response $Y \in \mathbb{R}$ and is interested in estimating the regression function $\mathbb{E}(Y|X = \cdot)$. In nonparametric regression $\mathbb{E}(Y|X = \cdot)$ is generally assumed to satisfy some smoothness assumptions (e.g., twice continuously differentiable), but no assumptions are made on the form of dependence on X. While nonparametric models offer flexibility in modeling, the price for this flexibility can be high for two main reasons: the estimation precision decreases rapidly as d increases ("curse of dimensionality") and the estimator can be hard to interpret when $d > 1$.

A natural restriction of the nonparametric model that avoids the curse of dimensionality while still retaining some flexibility in the functional form of $\mathbb{E}(Y|X = \cdot)$ is the single index model. In single index models, one assumes the existence of $\theta_0 \in \mathbb{R}^d$ such that

$$\mathbb{E}(Y|X) = \mathbb{E}\big(Y|\theta_0^\top X\big), \quad \text{almost every (a.e.) } X,$$

where $\theta_0^\top X$ is called the index; the widely used generalized linear models (GLMs) are special cases. This dimension reduction gives single index models considerable advantages in applications when $d > 1$ compared to the general nonparametric regression model; see [20] and [4] for a discussion. The aggregation of dimension by the index enables us to estimate the conditional mean function at a much faster rate than in a general nonparametric model. Since [49], single index models have become increasingly popular in many scientific fields including biostatistics, economics, finance, and environmental science and have been deployed in a variety of settings; see [33].

Formally, in this paper, we consider the model

$$Y = m_0\big(\theta_0^\top X\big) + \epsilon, \qquad \mathbb{E}(\epsilon|X) = 0, \quad \text{a.e. } X, \tag{1}$$

where $m_0 : \mathbb{R} \to \mathbb{R}$ is called the link function, $\theta_0 \in \mathbb{R}^d$ is the index parameter, and ϵ is the unobserved mean zero error (with finite variance). We assume that both m_0 and θ_0 are unknown and are the parameters of interest. For identifiability of (1), we assume that the first coordinate of θ_0 is non-zero and

$$\theta_0 \in \Theta := \big\{\eta = (\eta_1, \ldots, \eta_d) \in \mathbb{R}^d : |\eta| = 1 \text{ and } \eta_1 \geq 0\big\} \subset S^{d-1},$$

where $|\cdot|$ denotes the Euclidean norm, and S^{d-1} is the Euclidean unit sphere in $\mathbb{R}^d$; see [4] and [7] for a similar assumption.

Most of the existing techniques for estimation in single index models can be broadly classified into two groups, namely, M-estimation and "direct" estimation. M-estimation methods involve a nonparametric regression estimator of m_0 (e.g., kernel estimator [23], Bayesian B-splines [1], regression splines [45,61,62], local-linear approximation [63,70], and penalized splines [68]) and a minimization of some appropriate criterion function (e.g., quadratic loss [62,68], robust L_1 loss [72], profiled likelihood [45], quasi-likelihood [61], modal regression [35,66], and quantile regression [63]) with respect to the index parameter to obtain an estimator of θ_0. The so-called direct estimation methods include average derivative estimators [6,21,49,53], methods based on the conditional variance of Y [64,65], dimension reduction techniques, such as sliced inverse regression [31,32], and partial least squares [69]. Another prominent direct method is a kernel-based fixed point iterative scheme to compute an efficient estimator of θ_0 [7]. In these methods one tries to directly estimate θ_0 without estimating m_0, see, for example, in [21] the authors use the estimate of the derivative of the local linear approximation to $\mathbb{E}(Y|X = \cdot)$ and not the estimate of m_0 to estimate θ_0.

In this paper, we propose an M-estimation technique based on smoothing splines to simultaneously estimate the link function m_0 and the index parameter θ_0. When θ_0 is known, (1) reduces to a one-dimensional function estimation problem and smoothing splines offer a fast and easy-to-implement nonparametric estimator of the link function – m_0 is generally estimated by minimizing a penalized least squares criterion with a (natural) roughness penalty of integrated squared second derivative [11,59]. However, in the case of single index models, the problem is considerably harder as both the link function and the index parameter are unknown and intertwined (unlike in partial linear regression model [16]).

In other words, given i.i.d. data $\{(y_i, x_i)\}_{1 \leq i \leq n}$ from model (1), we propose minimizing the following penalized loss:

$$\frac{1}{n}\sum_{i=1}^{n}\big(y_i - m\big(\theta^\top x_i\big)\big)^2 + \lambda^2 \int \big|m''(t)\big|^2 dt \quad (\lambda \neq 0) \tag{2}$$

over $\theta \in \Theta$ and 'smooth' functions m; we will make this more precise in Section 2. Here λ is known as the smoothing parameter – high values of $|\lambda|$ lead to smoother estimators. The theory developed in this paper allows for the tuning parameter λ in (2) to be data dependent. Thus, data-driven procedures such as cross-validation can be used to choose an optimal λ; see Section 5. As

opposed to average derivative methods discussed earlier [21,49], the optimization problem in (2) involves only 1-dimensional nonparametric function estimation.

To the best of our knowledge, this is the first work that uses smoothing splines in the single index paradigm, under (only) smoothness constraints. We show that the penalized least squares loss leads to a minimizer $(\hat{m}, \hat{\theta})$. We study the asymptotic properties, that is, consistency, rates of convergence, of the estimator $(\hat{m}, \hat{\theta})$ under data dependent choices of the tuning parameter λ. We show that under sub-Gaussian errors $\hat{\theta}$ is asymptotically normal and, further, under homoscedastic errors $\hat{\theta}$ achieves the optimal semiparametric efficiency bound in the sense of [3].

Ichimura [23] developed a semiparametric least squares estimator of θ_0 using kernel estimates of the link function. However, the choice of tuning parameters (e.g., the bandwidth for estimation of the link function) make this procedure difficult to implement [8,15] and its numerical instability is well documented; see, for example, [68]. To address these issues [62,68] used B-splines and penalized splines to estimate m_0, respectively. However, in their proposed procedure the practitioner is required to choose the number and placement of knots for every θ. Smoothing splines avoid the choice of number of knots and their placement. Furthermore, smoothing splines (or more generally RKHS based regression estimators) are unique in that they are defined as minimization over a Hilbert space rather than as a local average. Even though smoothing splines can be approximated by kernel regression estimators or can be seen as a linear smoother, they are obtained under global smoothness constraint. This viewpoint makes them readily usable (at least in principle) when more constraints (such as monotonicity, non-negativity, unimodality, convexity, and k-monotonicity) need to be imposed. Several works including [18,48], and [67] advocate the use of smoothing splines for this reason. The above works also propose numerical methods for computing the constrained smoothing splines estimator in the case univariate nonparametric regression; also see [9,10,38,52]. These works suggest that, in addition to the convenience in problem formulation, the proof techniques for establishing consistency and asymptotic normality of the estimator for the finite-dimensional parameter in the constrained single index model will be almost the same as those for the smooth single index model studied here.

In contrast, other regression estimators such as kernel (or Nadaraya–Watson) estimator, series expansion, and regression splines imposing almost any type of (shape) constraint requires rethinking of the methods from scratch. This difficulty has posed several interesting works that consider estimation in constrained one dimensional nonparametric regression models; see [2,14, 40], and [50]. [14] modifies the kernel regression estimator by including probability weights for the summands and choosing these weights so as to satisfy monotonicity constraints. [50] further extends this by allowing for negative weights and thus enlarging the possible set of constraints; the computation, however, becomes difficult. [40] provides specific spline basis such that monotonicity and convexity constraints on functions can be converted into simple linear inequality constraints on the coefficients. However, this explicit basis construction for other general constraints (as discussed in [67]) seems out of reach at present and the extension of these methods to the case of single index model does not follow directly from existing work.

This paper gives a systematic and rigorous study of a smoothing splines based estimator for the single index model under minimal assumptions and fills an important gap in the literature. The assumptions for m_0 in this paper are weaker than those considered in the literature. We assume that the link function has an absolutely continuous derivative as opposed to the assumed (almost) three times differentiability of m_0 [7,23,49,62]. We study the model under the assumption that $\theta \in S^{d-1}$. In contrast, when the first coordinate is assumed to be 1, the parameter space

is unbounded and consistent estimation of θ_0 requires further assumptions, see, for example, [34]. [7] points out that the assumption $\theta \in S^{d-1}$ makes the parameter space irregular and the construction of paths on the sphere is hard. In this paper, we construct paths on the unit sphere to study the semiparametric efficiency of the finite dimensional parameter and provide a closed form expression for the variance of $\hat{\theta}$; see Theorem 5.

Our exposition is organized as follows. In Section 2, we introduce some notation, formally define our estimator, and study its existence. In Section 3, we prove consistency (see Theorem 3) and provide the rates of convergence (see Theorems 2 and 4) for our estimator. We show that the estimator for θ_0 is asymptotically normal and semiparametrically efficient; see Theorem 5 in Section 4. In Section 5, we provide finite sample simulation study of the proposed estimator and compare performance with existing methods in the literature. In Section 6, we apply the methodology developed to the car mileage data and the Ozone concentration data. In Section 7, we briefly summarize the results in the paper and provide some remarks on future directions of research. Appendices A–B contain proofs of the some of the results in the paper. The proofs of the results not given in the Appendices can be found in the on-line supplementary article [26].

2. Preliminaries

Suppose that $\{(y_i, x_i)\}_{1\le i\le n}$ is an i.i.d. sample from model (1). We start with some notation. Let $\mathcal{X} \subset \mathbb{R}^d$ denote the support of X. Let D be the set of possible index values and D_0 be the set of possible index values at θ_0, i.e.,

$$D := \left\{\theta^\top x : x \in \mathcal{X}, \theta \in \Theta\right\} \quad \text{and} \quad D_0 := \left\{\theta_0^\top x : x \in \mathcal{X}\right\}. \tag{3}$$

We denote the class of all real-valued functions with absolutely continuous first derivative on D by $\mathcal{S}$, that is,

$$\mathcal{S} := \left\{m : D \to \mathbb{R} | m' \text{ is absolutely continuous}\right\}.$$

We use $\mathbb{P}$ to denote the probability of an event, $\mathbb{E}$ for the expectation of a random quantity, and P_X for the distribution of X. For $g : \mathcal{X} \to \mathbb{R}$, define

$$\|g\|^2 := \int_{\mathcal{X}} g^2 \, dP_X \quad \text{and} \quad \|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g^2(x_i).$$

Let $P_{\epsilon,X}$ denote the joint distribution of (ϵ, X) and $P_{\theta,m}$ denote the joint distribution of (Y, X) when $Y = m(\theta^\top X) + \epsilon$. In particular, P_{θ_0,m_0} denotes the joint distribution of (Y, X) when (Y, X) satisfy (1). For any function $g : I \subset \mathbb{R}^p \to \mathbb{R}$, let $\|g\|_\infty := \sup_{u\in I} |g(u)|$. Moreover, for $I_1 \subset I$, we define $\|g\|_{I_1} := \sup_{u\in I_1} |g(u)|$. For any set $I \subset \mathbb{R}$, $\varnothing(I)$ denotes the diameter of the set I. For any $a \in \mathbb{R}^d$ and $r > 0$, $B_a(r)$ denotes the Euclidean ball of radius r centered at a. The notation $a \lesssim b$ is used to express that a is less than b up to a positive constant multiple. For any function $f : \mathcal{X} \to \mathbb{R}^r$, $r \ge 1$, let $\{f_i\}_{1\le i\le r}$ denote each of the components, that is, $f(x) = (f_1(x), \ldots, f_r(x))$, $r \ge 1$ and $f_i : \mathcal{X} \to \mathbb{R}$. We define $\|f\|_{2,2} := \sqrt{\sum_{i=1}^r \|f_i\|^2}$ and $\|f\|_{2,\infty} :=$

$\sqrt{\sum_{i=1}^{r} \|f_i\|_\infty^2}$. For any real-valued function m and $\theta \in \Theta$, we define

$$(m \circ \theta)(x) := m\big(\theta^\top x\big), \quad \text{for all } x \in \mathcal{X}.$$

For any function $f : D \subset \mathbb{R} \to \mathbb{R}$ with absolutely continuous first derivative, we define the roughness penalty

$$J^2(f) := \int_D \big|f''(t)\big|^2 dt.$$

We will now introduce the penalized least square estimator (PLSE). The penalized loss for $(m, \theta) \in \mathcal{S} \times \Theta$ (and $\lambda \neq 0$) is defined as

$$\mathcal{L}_n(m, \theta; \lambda) := \frac{1}{n} \sum_{i=1}^{n} \big(y_i - m\big(\theta^\top x_i\big)\big)^2 + \lambda^2 J^2(m). \tag{4}$$

The following theorem proves the existence of a (possibly non-unique) minimizer of $(m, \theta) \mapsto \mathcal{L}_n(m, \theta; \lambda)$.[1]

Theorem 1. *For every $\lambda \neq 0$, $(m, \theta) \mapsto \mathcal{L}_n(m, \theta; \lambda)$ is a continuous function and attains its minimum on $\mathcal{S} \times \Theta$. Furthermore, there exists a measurable minimizer.*

The PLSE $(\hat{m}, \hat{\theta})$ is defined to be *any* measurable element of the set of minimizers of $\mathcal{L}_n(m, \theta; \lambda)$, that is,

$$(\hat{m}, \hat{\theta}) \in \operatorname*{arg\,min}_{(m,\theta) \in \mathcal{S} \times \Theta} \mathcal{L}_n(m, \theta; \lambda). \tag{5}$$

We suppress the dependence of $(\hat{m}, \hat{\theta})$ on λ, for notational convenience. Theorem 2.4 of [11] shows that $\hat{m}$ is a natural cubic spline with knots at $\{\hat{\theta}^\top x_i\}_{1 \le i \le n}$.

It is easy to see that the composite population parameter $m_0 \circ \theta_0$ is identifiable. However, this does not guarantee that both m_0 and θ_0 are separately identifiable. Ichimura [23] (also see Horowitz [19], pages 12–17, and Li and Racine [33], Proposition 8.1) find sufficient conditions on the distribution/domain of X under which m_0 and θ_0 can be separately identified:

(**A0**) The function $m_0(\cdot)$ is non-constant, non-periodic, and a.e. differentiable. The first coordinate of θ_0 is positive, i.e., $\theta_{0,1} > 0$. The components of $X_1 \sim P_X$ (i.e., $X_{1,1}, \ldots, X_{1,d-1}$ and $X_{1,d}$) cannot have a perfect linear relationship. There exists an integer $d_1 \in \{1, 2, \ldots, d\}$, such that $X_{1,1}, \ldots, X_{1,d_1-1}$, and X_{1,d_1} have continuous distributions and $X_{1,d_1+1}, \ldots, X_{1,d-1}$, and $X_{1,d}$ be discrete random variables. Furthermore, there exist an open interval $\mathcal{I}$ and constant vectors $c_0, c_1, \ldots, c_{d-d_1} \in \mathbb{R}^{d-d_1}$ such that

- $c_l - c_0$ for $l \in \{1, \ldots, d - d_1\}$ are linearly independent,
- $\mathcal{I} \subset \bigcap_{l=0}^{d-d_1} \{\theta_0^\top x : x \in \mathcal{X} \text{ and } (x_{d_1+1}, \ldots x_d) = c_l\}$.

[1] See Section S.1 of the supplementary article [26] for a proof.

Ichimura [23] and Horowitz [19] prove by examples that each part of Assumption **(A0)** is necessary for identifiability of m_0 and θ_0. Further discussion on alternative identifiability assumptions when X has a Lebesgue density, we refer to Kuchibhotla et al. [27], Section 2.

3. Asymptotic analysis of the PLSE

In this section, we will list the assumptions under which we will establish consistency and find the rates of convergence of our estimators. Note that we will study $(\hat{m}, \hat{\theta})$ for any (possibly data-driven) choice of λ satisfying two rate conditions; see assumption **(A4)** below.

- **(A1)** The link function m_0 satisfies $J(m_0) < \infty$.
- **(A2)** $\mathcal{X}$, the support of X, is a compact subset of $\mathbb{R}^d$ and $\sup_{x \in \mathcal{X}} |x| \le T$.
- **(A3)** The error ϵ in model (1) is conditionally sub-Gaussian, that is, there exists $K > 0$ such that
$$\mathbb{E}\big[\exp\big(\epsilon^2/K\big)|X\big] \le 2 \quad \text{a.e. } X.$$
As stated in (1), we also assume that $\mathbb{E}(\epsilon|X) = 0$ a.e. X.
- **(A4)** The smoothing parameter λ can be chosen to be a random variable. For the rest of the paper, we denote it by $\hat{\lambda}_n$. Assume that $\hat{\lambda}_n$ satisfies the rate condition:
$$\hat{\lambda}_n^{-1} = O_p\big(n^{2/5}\big) \quad \text{and} \quad \hat{\lambda}_n = o_p\big(n^{-1/4}\big). \tag{6}$$

The assumptions deserve comments. In **(A1)** our assumption on m_0 is quite minimal – we essentially require m_0 to have an absolutely continuous derivative. Most previous works assume m_0 to be three times differentiable; see, for example, [43,49]. Note that the assumption $J(m_0) < \infty$ in combination with compact support of X implies that m_0 is bounded and we set $M_1 := \|m_0\|_\infty$. **(A2)** assumes that the support of the covariates is bounded. As the class of functions $\mathcal{S}$ is not uniformly bounded, we use assumption **(A3)** to provide control over the tail behavior of ϵ; see Chapter 8 of [56] for a discussion on this. Observe that **(A3)** allows for heteroscedastic errors. Assumption **(A4)** allows our tuning parameter to be data dependent, as opposed to a sequence of constants. This allows for data driven choices of $\hat{\lambda}_n$, such as cross-validation. We will show that for any choice of $\hat{\lambda}_n$ satisfying (6), $\hat{\theta}$ will be an asymptotically efficient estimator of θ_0. We use empirical process methods (see, e.g., [58]) to prove the consistency and to find the rates of convergence of $\hat{m} \circ \hat{\theta}$.

In Theorem 2, we show that $(\hat{m}, \hat{\theta})$ is a consistent estimator of (m_0, θ_0) and $\hat{m} \circ \hat{\theta}$ converges to $m_0 \circ \theta_0$ at rate $\hat{\lambda}_n$ (with respect to the $L_2(P_X)$-norm).

Theorem 2. *Under assumptions* **(A0)**–**(A4)**, *the PLSE satisfies* $J(\hat{m}) = O_p(1)$, $\|\hat{m}\|_\infty = O_p(1)$, *and* $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n)$.

Next, we prove the consistency of $\hat{m}$ and $\hat{\theta}$. We prove that $\hat{m}$ is consistent under the Sobolev norm, which for any set $I \subset \mathbb{R}$ and any function $g : I \to \mathbb{R}$ is defined as

$$\|g\|_I^S = \sup_{t \in I} \big|g(t)\big| + \sup_{t \in I} \big|g'(t)\big|.$$

Recall sets D and D_0 defined in (3).

Theorem 3. *Under assumptions* **(A0)**–**(A4)**,

$$\hat{\theta} \xrightarrow{P} \theta_0, \qquad \|\hat{m} - m_0\|_{D_0}^S \xrightarrow{P} 0, \quad \text{and} \quad \|\hat{m}'\|_\infty = \sup_{t \in D} |\hat{m}'(t)| = O_p(1).$$

The above result shows that not only is $\hat{m}$ consistent but its derivative $\hat{m}'$ also converges uniformly to m_0'. Proof of Theorem 2 is in Appendix A.1 and proof of Theorem 3 is given in Section S.2.2 the supplementary article [26]. We next introduce further notation and provide upper bounds on the rates of convergence of $\hat{\theta}$ and $\hat{m}$ separately.

Recall that Θ is a closed subset of $\mathbb{R}^d$ and the interior of Θ in $\mathbb{R}^d$ is the null set. Thus we will define a "local parameterization matrix" that will help us create linear perturbations of θ_0 that lie in Θ. For every real matrix $G \in \mathbb{R}^{m \times n}$, we define $\|G\|_2 := \max_{x \in S^{n-1}} |Gx|$. This is sometimes called the operator or matrix 2-norm; see, for example, page 281 of [39]. The following lemma proved[2] in Section S.2.3 of the supplementary article [26] shows that the "local parameterization matrix" as a function of θ is Lipschitz at θ_0 with respect to the operator norm.

Lemma 1. *There exists a set of matrices* $\{H_\theta \in \mathbb{R}^{d \times (d-1)} : \theta \in \Theta\}$ *satisfying the following properties*:

(a) $\xi \mapsto H_\theta \xi$ *are bijections from* $\mathbb{R}^{d-1}$ *to the hyperplanes* $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$.
(b) *The columns of* H_θ *form an orthonormal basis for* $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$.
(c) $\|H_\theta - H_{\theta_0}\|_2 \le |\theta - \theta_0|$.
(d) *For all distinct* $\eta, \beta \in \Theta \setminus \theta_0$, *such that* $|\eta - \theta_0| \le 1/2$ *and* $|\beta - \theta_0| \le 1/2$,

$$\|H_\eta^\top - H_\beta^\top\|_2 \le 8(1 + 8/\sqrt{15}) \frac{|\eta - \beta|}{|\eta - \theta_0| + |\beta - \theta_0|}.$$

Note that for each $\theta \in \Theta$, $H_\theta^\top$ is the Moore-Penrose pseudo-inverse of H_θ, for example, $H_\theta^\top H_\theta = \mathbb{I}_{d-1}$ where $\mathbb{I}_{d-1}$ is the identity matrix of order $d-1$; see Section 5.2 of [46] for a similar construction.

The following distributional assumption on X is used to find the upper bounds on the rates of convergence of $\hat{\theta}$ and $\hat{m}$ separately.

(A5) $H_{\theta_0}^\top \mathbb{E}[\mathrm{Var}(X|\theta_0^\top X)\{m_0'(\theta_0^\top X)\}^2] H_{\theta_0}$ is a positive definite matrix.

If one of the continuous covariates with a nonzero index parameter has a density (with respect to the Lebesgue measure) that is bounded away from zero (on its support) then assumption **(A5)** is satisfied. Note that **(A5)** fails if m_0 is a constant function; however a single index model is not identifiable if m_0 is constant (see **(A0)**). The following bounds (proved in Section S.2.4 of the supplementary article [26]) will help us compute the asymptotic distribution of $\hat{\theta}$ in Section 4.

Theorem 4. *Under* **(A0)**–**(A5)**, $\hat{m}$ *and* $\hat{\theta}$ *satisfy*

$$|\hat{\theta} - \theta_0| = O_p(\hat{\lambda}_n) \quad \text{and} \quad \|\hat{m} \circ \theta_0 - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n).$$

[2]Our proof is constructive.

4. Semiparametric inference

In this section, we show that $\hat{\theta}$ is asymptotically normal and is a semiparametrically efficient estimator of θ_0 under homoscedastic errors. Before going into the derivation of the limit law of $\hat{\theta}$, we need to introduce some further notation and some regularity assumptions. For every $\theta \in \Theta$, let us define $D_\theta := \{\theta^\top x : x \in \mathcal{X}\}$. Assumption **(A0)** implies that there exists $r > 0$ such that for all $\theta \in S^{d-1} \cap B_{\theta_0}(r)$ we have

$$D_\theta \subsetneq D^{(r)} := \bigcup_{\theta \in S^{d-1} \cap B_{\theta_0}(r)} D_\theta. \tag{7}$$

See Section S.3.2 of the supplementary article [26] for a proof of this. For the rest of the paper, we redefine $D := D^{(r)}$. For every $\theta \in \Theta$, define $h_\theta : D \to \mathbb{R}^d$ as

$$h_\theta(u) := \mathbb{E}\big[X | \theta^\top X = u\big]. \tag{8}$$

We use the following additional assumptions in the proof of asymptotic normality of $\hat{\theta}$.

(B1) $h_\theta(\cdot)$ is twice continuously differentiable except possibly at a finite number of points, and for every θ_1 and θ_2 in Θ,

$$\|h_{\theta_1} - h_{\theta_2}\|_\infty \le \bar{M} |\theta_1 - \theta_2|,$$

where $\bar{M}$ is a fixed finite constant.

Let $p_{\epsilon,X}$ denote the joint density (with respect to some dominating measure μ on $\mathbb{R} \times \mathcal{X}$) of (ϵ, X). Let $p_{\epsilon|X}(e, x)$ and $p_X(x)$ denote the corresponding conditional probability density of ϵ given X and the marginal density of X, respectively. We define $\sigma : \mathcal{X} \to \mathbb{R}$ by $\sigma^2(x) := \mathbb{E}(\epsilon^2 | X = x)$.

(B2) $p_{\epsilon|X}(e, x)$ is differentiable with respect to e, $\|\sigma^2(\cdot)\|_\infty < \infty$ and $\|1/\sigma^2(\cdot)\|_\infty < \infty$.

The assumptions **(B1)** and **(B2)** deserve comments. The function h_θ plays a crucial role in the construction of "least favorable" paths; see Section 4.2.2. For the functions in the path to be in $\mathcal{S}$, we use the smoothness assumptions on h_θ. In a way we need smoothness of m_0 or the distribution of X to be smooth to be able to establish semiparametric efficiency. **(B2)** gives lower and upper bounds on the variance of ϵ as we are using a un-weighted least squares method to estimate parameters in a (possibly) heteroscedastic model.

In the sequel we will use standard empirical process theory notation. For any function $f : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ and $(m, \theta) \in \mathcal{S} \times \Theta$, we define

$$P_{\theta,m} f = \int f \, d P_{\theta,m}.$$

Note that $P_{\theta,m} f$ can be a random variable if θ (or m) is random. Moreover, for any function $f : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, we define

$$\mathbb{P}_n f := \frac{1}{n} \sum_{i=1}^n f(y_i, x_i) \quad \text{and} \quad \mathbb{G}_n f := \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[f(y_i, x_i) - P_{\theta_0,m_0} f \big].$$

4.1. Efficient score

As a first step in showing that $\hat{\theta}$ is an efficient estimator, in the following we find the efficiency bound for θ_0 in model (1). To compute the score for the model, we will first consider parametric paths on Θ. For any $\eta \in \mathbb{R}^{d-1}$ and $\theta \in \Theta$, we now define a path $s \mapsto \zeta_s(\theta, \eta)$, for $s \in \mathbb{R}$ and $|s| \le |\eta|^{-1}$, as

$$\zeta_s(\theta, \eta) := \sqrt{1 - s^2|\eta|^2}\theta + sH_\theta\eta. \tag{9}$$

Note that $\theta^\top H_\theta = 0_{d-1}$ and $|H_\theta\eta| = |\eta|$ for all $\eta \in \mathbb{R}^{d-1}$. When $|s| \le 1/|\eta|$ we have $\zeta_s(\theta, \eta) \in S^{d-1}$. For every fixed $s \ne 0$, as η varies in $B_0^{d-1}(|s|^{-1})$, $\zeta_s(\theta, \eta)$ takes all values in the set $\{\beta \in S^{d-1} : \theta^\top\beta > 0\}$ and $sH_\theta\eta$ is the orthogonal projection of $\zeta_s(\theta, \eta)$ onto the hyperplane $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$.

We now attempt to calculate the efficient score for

$$Y = m\big(\theta^\top X\big) + \epsilon \tag{10}$$

for some $(m, \theta) \in \mathcal{S} \times \Theta$ under assumptions **(A3)** and **(B2)**. The log-likelihood of the model is

$$l_{\theta,m}(y, x) = \log\big[p_{\epsilon|X}\big(y - m\big(\theta^\top x\big), x\big)p_X(x)\big].$$

Remark 1. Note that under (10), we have $\epsilon = Y - m(\theta^\top X)$. For every function $b(e, x) : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ in $L_2(P_{\epsilon,X})$, there exists an "equivalent" function $\tilde{b}(y, x) : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ in $L_2(P_{\theta,m})$ defined as $\tilde{b}(y, x) := b(y - m(\theta^\top x), x) \in L_2(P_{\theta,m})$. In this section, we use the function arguments (e, x) $(L_2(P_{\epsilon,X}))$ and (y, x) $(L_2(P_{\theta,m}))$ interchangeably.

For $\eta \in S^{d-2} \subset \mathbb{R}^{d-1}$, consider the path defined in (9). Note that this is a valid path through θ as $\zeta_0(\theta, \eta) = \theta$. The score function for this submodel (the parametric score) is

$$\left.\frac{\partial l_{\zeta_s(\theta,\eta),m}(y, x)}{\partial s}\right|_{s=0} = \eta^\top S_{\theta,m}(y, x),$$

$$\text{where } S_{\theta,m}(y, x) := -\frac{p'_{\epsilon|X}(y - m(\theta^\top x), x)}{p_{\epsilon|X}(y - m(\theta^\top x), x)}m'\big(\theta^\top x\big)H_\theta^\top x.$$

We now define a parametric submodel for the unknown nonparametric components:

$$\begin{aligned} m_{s,a}(t) &= m(t) - sa(t), \\ p_{\epsilon|X;s,b}(e, x) &= p_{\epsilon|X}(e, x)\big(1 + sb(e, x)\big), \\ p_{X;s,q}(x) &= p_X(x)\big(1 + sq(x)\big), \end{aligned} \tag{11}$$

where $s \in \mathbb{R}$, $b : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(b(\epsilon, X)|X) = 0$ and $\mathbb{E}(\epsilon b(\epsilon, X)|X) = 0$, $a \in \mathcal{S}$ such that $J(a) < \infty$ and $q : \mathcal{X} \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(q(X)) = 0$. Consider the following parametric submodel of (1),

$$s \mapsto \big(\zeta_s(\theta, \eta), m_{s,a}, p_{\epsilon|X;s,b}, p_{X;s,q}(x)\big), \tag{12}$$

where $\eta \in S^{d-2}$. Differentiating the log-likelihood of the submodel in (12) with respect to s, we get that the score along the submodel in (12) is

$$\eta^\top S_{\theta,m}(y,x) + \frac{p'_{\epsilon|X}(y-m(\theta^\top x),x)}{p_{\epsilon|X}(y-m(\theta^\top x),x)} a\big(\theta^\top x\big) + b\big(y-m\big(\theta^\top x\big),x\big) + q(x). \tag{13}$$

Newey and Stoker [43] and Ma and Zhu [36] find the following characterization of the orthogonal complement ($\Lambda^\perp$) of the nuisance tangent space Λ,

$$\Lambda^\perp = \big\{ f \in L_2(P_{\epsilon,X}) : f(e,x) = \big[g(x) - \mathbb{E}\big(g(X)|\theta^\top X = \theta^\top x\big)\big]e, \text{ for some } g : \mathcal{X} \to \mathbb{R} \big\}.$$

A derivation of the above result can also be found in Section S.3.1 of the supplementary article [26]. Now, using calculations similar those in Proposition 1 in [36], it can be shown that

$$\Pi\big(S_{\theta,m}|\Lambda^\perp\big)(y,x) = \frac{(y-m(\theta^\top x))}{\sigma^2(x)} m'\big(\theta^\top x\big) H_\theta^\top \left\{ x - \frac{\mathbb{E}(\sigma^{-2}(X)X|\theta^\top X = \theta^\top x)}{\mathbb{E}(\sigma^{-2}(X)|\theta^\top X = \theta^\top x)} \right\}, \tag{14}$$

where for any $f \in L_2(P_{\epsilon,X})$, $\Pi(f|\Lambda^\perp)$ denotes the $L_2(P_{\epsilon,X})$ projection of f onto the space $\Lambda^\perp$. $\Pi(S_{\theta,m}|\Lambda^\perp)$ is sometimes denoted by $S^{\text{eff}}_{\theta,m}$. It is important to note that the optimal estimating equation depends on $\sigma^2(\cdot)$. Since in the semiparametric model $\sigma^2(\cdot)$ is left unspecified, it is unknown. Without additional assumptions, nonparametric estimators of $\sigma^2(\cdot)$ have a slow rate of convergence to $\sigma^2(\cdot)$, especially if d is large. Thus, if we substitute $\hat{\sigma}(x)$ in the efficient score equation, the solution of the modified score equation would lead to poor finite sample performance; see [54].

To focus our presentation on the main concepts, briefly consider the case when $\sigma^2(\cdot) \equiv \sigma^2$. In this case the efficient score $\Pi(S_{\theta,m}|\Lambda^\perp)(y,x)$ is

$$\frac{1}{\sigma^2}\big(y - m\big(\theta^\top x\big)\big) m'\big(\theta^\top x\big) H_\theta^\top \big\{ x - h_\theta\big(\theta^\top x\big)\big\},$$

where $h_\theta(\theta^\top x)$ is defined in (8). Asymptotic normality and efficiency of $\hat{\theta}$ would follow if we can show that $(\hat{m}, \hat{\theta})$ satisfies the efficient score equation *approximately*, that is,

$$\mathbb{P}_n\left[\frac{1}{\sigma^2}\big(Y - \hat{m}\big(\hat{\theta}^\top X\big)\big) \hat{m}'\big(\hat{\theta}^\top X\big) H_{\hat{\theta}}^\top \big\{ X - h_{\hat{\theta}}\big(\hat{\theta}^\top X\big)\big\} \right] = o_p\big(n^{-1/2}\big)$$

and a class of functions formed by the efficient score indexed by (θ, m) in a "neighborhood" of (θ_0, m_0) satisfies some "uniformity" conditions, for example, it is a Donsker class. We formalize this notion of efficiency in Theorem 5 below.

4.2. Efficiency of $\hat{\theta}$

Theorem 5. *Assume that* (Y, X) *satisfies* (1) *and assumptions* **(A0)**–**(A5)**, **(B1)**, *and* **(B2)** *hold. Define*

$$\tilde{\ell}_{\theta,m}(y,x) := \big(y - m\big(\theta^\top x\big)\big) m'\big(\theta^\top x\big) H_\theta^\top \big\{ x - h_\theta\big(\theta^\top x\big)\big\}. \tag{15}$$

If $V_{\theta_0,m_0} := P_{\theta_0,m_0}(\tilde{\ell}_{\theta_0,m_0} S_{\theta_0,m_0}^\top)$ *is a nonsingular matrix in* $\mathbb{R}^{(d-1)\times(d-1)}$, *then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\big(0, H_{\theta_0} V_{\theta_0,m_0}^{-1} \tilde{I}_{\theta_0,m_0} \big(H_{\theta_0} V_{\theta_0,m_0}^{-1}\big)^\top\big), \tag{16}$$

where $\tilde{I}_{\theta_0,m_0} := P_{\theta_0,m_0}(\tilde{\ell}_{\theta_0,m_0} \tilde{\ell}_{\theta_0,m_0}^\top)$. *If we further assume that* $\sigma^2(\cdot) \equiv \sigma^2$ *and if the efficient information matrix,* $\tilde{I}_{\theta_0,m_0}$, *is nonsingular, then* $\hat{\theta}$ *is an efficient estimator of* θ_0, *i.e.*,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\big(0, \sigma^4 H_{\theta_0} \tilde{I}_{\theta_0,m_0}^{-1} H_{\theta_0}^\top\big). \tag{17}$$

Remark 2. Note that even if $\mathbb{E}(\epsilon^2|X) \not\equiv \sigma^2$, $\hat{\theta}$ is a consistent and asymptotically normal estimator of θ. When the constant variance assumption provides a good approximation to the truth, estimators similar to $\hat{\theta}$ have been known to have high relative efficiency with respect to the optimal semiparametric efficiency bound; see Page 94 of [54] for a discussion. When $\sigma^2(x) = V^2(\theta_0^\top x)$ for some unknown real-valued function V, we can define a weighted PLSE as

$$(\tilde{m}, \tilde{\theta}) := \operatorname*{arg\,min}_{(m,\theta)\in\mathcal{S}\times\Theta} \frac{1}{n} \sum_{i=1}^{n} \hat{w}(x_i)\big(y_i - m\big(\theta^\top x_i\big)\big)^2 + \hat{\lambda}_n^2 J^2(m),$$

where $\hat{w}(x)$ is a consistent estimator of $V^{-2}(\theta_0^\top x)$. Theorem 5 can be easily generalized to show that $\tilde{\theta}$ is an efficient estimator of θ_0 under this specific heteroscedastic structure.

Remark 3. The asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is the same as that obtained in Section 2.4 of [15] and [5] (under assumption **(A4)**). However, both require stronger smoothness assumptions on m_0 for their estimators.

Remark 4. Observe that the variance of the limiting distribution (for both the heteroscedastic and homoscedastic models) is singular. This can be attributed to the fact that Θ is a Stiefel manifold of dimension $\mathbb{R}^{d-1}$ and has an empty interior in $\mathbb{R}^d$.

4.2.1. *Proof of Theorem 5*

In the following, we give a sketch of the proof of (16). Some of the steps are proved in the following sections.

Step 1 In Theorem 6, we will show that $(\hat{m}, \hat{\theta})$ satisfy the efficient score equation *approximately*, that is,

$$\sqrt{n}\mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{m}} = o_p(1). \tag{18}$$

Step 2 In Section S.3.3 of the supplementary article [26], we prove that $\tilde{\ell}_{\hat{\theta},\hat{m}}$ is unbiased in the sense of [57], that is,

$$P_{\hat{\theta},m_0} \tilde{\ell}_{\hat{\theta},\hat{m}} = 0. \tag{19}$$

Similar conditions have appeared before in proofs of asymptotic normality of the MLE (e.g., see [22]) and the construction of efficient one-step estimators (see [24]); see Section 3 of [41] for further discussion.

Step 3 We prove

$$\mathbb{G}_n(\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}) = o_p(1) \tag{20}$$

in Theorem 7. In view of (18) and (19) an equivalent formulation of (20) is

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\tilde{\ell}_{\hat{\theta},\hat{m}} = \mathbb{G}_n \tilde{\ell}_{\theta_0,m_0} + o_p(1). \tag{21}$$

Step 4 To complete the proof of (16), it is enough to show that

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\tilde{\ell}_{\hat{\theta},\hat{m}} = \sqrt{n} V_{\theta_0,m_0} H_{\theta_0}^\top(\hat{\theta} - \theta_0) + o_p\big(\sqrt{n}|\hat{\theta} - \theta_0|\big). \tag{22}$$

A proof of slightly simplified version of (22) can be found in the proof of Theorem 6.20 of [57]. However, for the sake of completeness we give a proof of (22) in Section S.3.4 of the supplementary article [26].

Observe that (21) and (22) imply

$$\begin{aligned} &\sqrt{n} V_{\theta_0,m_0} H_{\theta_0}^\top(\hat{\theta} - \theta_0) = \mathbb{G}_n \tilde{\ell}_{\theta_0,m_0} + o_p\big(1 + \sqrt{n}|\hat{\theta} - \theta_0|\big), \\ \Rightarrow \quad &\sqrt{n} H_{\theta_0}^\top(\hat{\theta} - \theta_0) = V_{\theta_0,m_0}^{-1} \mathbb{G}_n \tilde{\ell}_{\theta_0,m_0} + o_p(1) \xrightarrow{d} V_{\theta_0,m_0}^{-1} N(0, \tilde{I}_{\theta_0,m_0}). \end{aligned} \tag{23}$$

The proof of the theorem will be complete if we can show that

$$\sqrt{n}(\hat{\theta} - \theta_0) = H_{\theta_0} \sqrt{n} H_{\theta_0}^\top(\hat{\theta} - \theta_0) + o_p(1).$$

Let $\hat{\eta}$ be the unique vector in $\mathbb{R}^{d-1}$ that satisfies the following equation:

$$\hat{\theta} = \sqrt{1 - |\hat{\eta}|^2}\theta_0 + H_{\theta_0}\hat{\eta}, \tag{24}$$

note that such an $\hat{\eta}$ will always exists as $\hat{\theta} \xrightarrow{P} \theta_0$. As $H_{\theta_0}^\top \theta_0 = 0$ and $H_{\theta_0}^\top H_{\theta_0} = \mathbb{I}_{d-1}$, pre-multiplying both sides of the previous equation by $H_{\theta_0}^\top$ we get

$$\hat{\eta} = H_{\theta_0}^\top(\hat{\theta} - \theta_0). \tag{25}$$

Substituting the above expression of $\hat{\eta}$ in (24) and subtracting θ_0 from both sides of (24) we get

$$\hat{\theta} - \theta_0 = \Big[\sqrt{1 - \big|H_{\theta_0}^\top(\hat{\theta} - \theta_0)\big|^2} - 1\Big]\theta_0 + H_{\theta_0} H_{\theta_0}^\top(\hat{\theta} - \theta_0).$$

By (23) we have that $\sqrt{n} H_{\theta_0}^\top(\hat{\theta} - \theta_0) = O_p(1)$. Moreover, note that $\sqrt{1 - x^2} - 1 = O(x^2)$, as $x \to 0$. Combining the above facts, we get

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &= \sqrt{n} O_p\big(\big|H_{\theta_0}^\top(\hat{\theta} - \theta_0)\big|^2\big) + \sqrt{n} H_{\theta_0} H_{\theta_0}^\top(\hat{\theta} - \theta_0) \\ &= H_{\theta_0} \sqrt{n} H_{\theta_0}^\top(\hat{\theta} - \theta_0) + O_p\big(n^{-1/2}\big). \end{aligned}$$

Now we prove (17). Assume that $\sigma^2(\cdot) \equiv \sigma^2$. Observe that, by (14) and (15), we have

$$\begin{aligned} S_{\theta_0,m_0} &= \Pi\big(S_{\theta_0,m_0}|\Lambda^{\perp}\big) + \big(S_{\theta_0,m_0} - \Pi\big(S_{\theta_0,m_0}|\Lambda^{\perp}\big)\big) \\ &= \frac{1}{\sigma^2}\tilde{\ell}_{\theta_0,m_0} + \big(S_{\theta_0,m_0} - \Pi\big(S_{\theta_0,m_0}|\Lambda^{\perp}\big)\big). \end{aligned}$$

Thus (17) follows from (16) by observing that

$$V_{\theta_0,m_0} = P_{\theta_0,m_0}\big(\tilde{\ell}_{\theta_0,m_0} S_{\theta_0,m_0}^{\top}\big) = \frac{1}{\sigma^2}\tilde{I}_{\theta_0,m_0}.$$

4.2.2. *"Least favorable" path for m*

We will now show that (18) holds. Recall the definition (9). For any $(\theta, m) \in \Theta \times \{m \in \mathcal{S} | J(m) < \infty\}$ and $\eta \in S^{d-2}$, let $t \mapsto (\zeta_t(\theta,\eta), \xi_t(\cdot;\theta,\eta,m))$ denote a path in $\Theta \times \{m \in \mathcal{S}|J(m) < \infty\}$ that goes through (θ, m), i.e., $(\zeta_0(\theta,\eta), \xi_0(\cdot;\theta,\eta,m)) = (\theta, m)$; see (29) below for definition. Recall that $(\hat{\theta}, \hat{m})$ minimizes $\mathcal{L}_n(m, \theta, \hat{\lambda}_n)$. Hence, for every $\eta \in S^{d-2}$, the function $t \mapsto \mathcal{L}_n(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}), \zeta_t(\hat{\theta},\eta), \hat{\lambda}_n)$ is minimized at $t = 0$. In particular, if the above function is differentiable in a neighborhood of 0, then

$$\frac{\partial}{\partial t}\mathcal{L}_n\big(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}), \zeta_t(\hat{\theta},\eta), \hat{\lambda}_n\big)\Big|_{t=0} = 0. \tag{26}$$

Moreover if $(\zeta_t(\hat{\theta},\eta), \xi_t(\cdot;\hat{\theta},\eta,\hat{m}))$ satisfies

$$\begin{aligned} \frac{\partial}{\partial t}\big(y - \xi_t\big(\zeta_t(\hat{\theta},\eta)^{\top}x;\hat{\theta},\eta,\hat{m}\big)\big)^2\Big|_{t=0} &= \eta^{\top}\tilde{\ell}_{\hat{\theta},\hat{m}}(y,x), \\ \frac{\partial}{\partial t}J^2\big(\xi_t(\cdot;\hat{\theta},\eta,\hat{m})\big)\Big|_{t=0} &= O_p(1). \end{aligned} \tag{27}$$

for all $\eta \in S^{d-2}$, then we get (18) as $\hat{\lambda}_n^2 = o_p(n^{-1/2})$; see assumption (**A4**).

Observe that $\hat{\theta}$ is a consistent estimator of θ_0. As we are concerned with the path $t \mapsto \mathcal{L}_n(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}), \zeta_t(\hat{\theta},\eta), \hat{\lambda}_n)$, we will try to construct a path for any $(\theta, m) \in \{\Theta \cap B_{\theta_0}(r)\} \times \{m \in \mathcal{S}|J(m) < \infty\}$ that satisfies the above requirements. For any set $A \subset \mathbb{R}$ and any $\nu > 0$ let us define $A^{\nu} := \cup_{a\in A} B_a(\nu)$ and let ∂A denote the boundary of A. Fix $\nu > 0$. By (7), for every $\theta \in \Theta \cap B_{\theta_0}(r)$, $\eta \in S^{d-2}$, and $t \in \mathbb{R}$ sufficiently close to zero, there exists a strictly increasing function $\phi_{\theta,\eta,t} : D^{\nu} \to \mathbb{R}$ with

$$\begin{aligned} \phi_{\theta,\eta,t}(u) &= u, \quad u \in D_{\theta}, \\ \phi_{\theta,\eta,t}\big(u + \big(\theta - \zeta_t(\theta,\eta)\big)^{\top}h_{\theta}(u)\big) &= u, \quad u \in \partial D, \end{aligned} \tag{28}$$

where $h_{\theta}(u)$ and $\zeta_t(\theta,\eta)$ are defined in (8) and (9), respectively. Furthermore, we can ensure that $\phi_{\theta,\eta,t}(u)$ is infinitely differentiable for $u \in D$ and that $\frac{\partial}{\partial t}\phi_{\theta,\eta,t}|_{t=0}$ exists. Note

that $\phi_{\theta,\eta,t}(D) = D$. Moreover, $\phi_{\theta,\eta,t}$ cannot be the identity function for $t \neq 0$ if $(\theta - \zeta_t(\theta,\eta))^\top h_\theta(u) \neq 0$ for $u \in \partial D$. Now, we can define the following path through m:

$$\xi_t(u;\theta,\eta,m) := m \circ \phi_{\theta,\eta,t}\big(u + \big(\theta - \sqrt{1 - t^2|\eta|^2}\theta - t H_\theta \eta\big)^\top h_\theta(u)\big). \tag{29}$$

The function $\phi_{\theta,\eta,t}$ helps us control the partial derivative in the second equation of (27). In the following theorem (proved in Appendix B.1), we show that $(\zeta_t(\hat\theta,\eta), \xi_t(\cdot;\hat\theta,\eta,\hat m))$ is a path through $(\hat\theta,\hat m)$ and satisfies (26) and (27). Here η is the "direction" for the path $t \mapsto \zeta_t(\theta,\eta)$ and $(\eta, h_\theta(u))$ defines the "direction" for the path $t \mapsto \xi_t(\cdot;\theta,\eta,m)$.

Theorem 6. *Under assumptions* **(A0)**, **(A1)**, **(A4)**, *and* **(B1)**, $(\zeta_t(\hat\theta,\eta), \xi_t(\cdot;\hat\theta,\eta,\hat m))$ *is a valid parametric submodel, that is,* $(\zeta_t(\hat\theta,\eta), \xi_t(\cdot;\hat\theta,\eta,\hat m)) \in \Theta \times \{m \in \mathcal{S} | J(m) < \infty\}$ *for all* t *in some neighborhood of* 0. *Moreover,* $(\zeta_t(\hat\theta,\eta), \xi_t(\cdot;\hat\theta,\eta,\hat m))$ *satisfies* (27) *and* $\mathcal{L}_n(\xi_t(\cdot;\hat\theta,\eta,\hat m), \zeta_t(\hat\theta,\eta), \hat\lambda_n)$, *as function of* t, *is differentiable at* 0 *and* $\sqrt{n}\mathbb{P}_n\tilde\ell_{\hat\theta,\hat m} = o_p(1)$.

4.2.3. *Asymptotic equicontinuity of* $\tilde\ell_{\theta,m}$ *at* (θ_0, m_0)

For notational convenience, we define

$$K_1(x;\theta) := H_\theta^\top\big(x - h_\theta\big(\theta^\top x\big)\big).$$

With the above notation, from (15) we have

$$\tilde\ell_{\theta,m}(y,x) = \big(y - m\big(\theta^\top x\big)\big)m'\big(\theta^\top x\big)K_1(x;\theta).$$

Theorem 7. *Under assumptions* **(A0)**–**(A5)**, **(B1)**, *and* **(B2)**, $\mathbb{G}_n(\tilde\ell_{\hat\theta,\hat m} - \tilde\ell_{\theta_0,m_0}) = o_p(1)$.

Proof. We divide the proof Theorem 7 into two lemmas. First, observe that

$$\begin{aligned}
&\mathbb{G}_n(\tilde\ell_{\hat\theta,\hat m} - \tilde\ell_{\theta_0,m_0}) \\
&\quad= \mathbb{G}_n\big[\big(Y - \hat m\big(\hat\theta^\top X\big)\big)\hat m'\big(\hat\theta^\top X\big)K_1(X;\hat\theta) - \big(Y - m_0\big(\theta_0^\top X\big)\big)m_0'\big(\theta_0^\top X\big)K_1(X;\theta_0)\big] \\
&\quad= \mathbb{G}_n\big[\big(\epsilon + m_0\big(\theta_0^\top X\big) - \hat m\big(\hat\theta^\top X\big)\big)\hat m'\big(\hat\theta^\top X\big)K_1(X;\hat\theta) - \epsilon m_0'\big(\theta_0^\top X\big)K_1(X;\theta_0)\big] \\
&\quad= \mathbb{G}_n\big[\big(m_0\big(\theta_0^\top X\big) - \hat m\big(\hat\theta^\top X\big)\big)\hat m'\big(\hat\theta^\top X\big)K_1(X;\hat\theta)\big] \\
&\qquad+ \mathbb{G}_n\big[\epsilon\big(\hat m'\big(\hat\theta^\top X\big)K_1(X;\hat\theta) - m_0'\big(\theta_0^\top X\big)K_1(X;\theta_0)\big)\big].
\end{aligned} \tag{30}$$

The proof of Theorem 7 will be complete, if we can show that both the terms in (30) converge to 0 in probability. We begin with some definitions. Let a_n be a sequence of real numbers such that $a_n \to \infty$ as $n \to \infty$ and $a_n\|\hat m - m_0\|_{D_0}^S = o_p(1)$. We can always find such a sequence a_n,

as we have $\|\hat{m} - m_0\|_{D_0}^S = o_p(1)$ (see Theorem 3). For all $n \in \mathbb{N}$, define[3]

$$\begin{aligned}
\mathcal{C}^{m*}_{M_1,M_2,M_3} &:= \big\{m \in \mathcal{S} : \|m\|_\infty < M_1, \big\|m'\big\|_\infty < M_2, \text{and } J(m) < M_3\big\},\\
\mathcal{C}^{m}_{M_1,M_2,M_3}(n) &:= \big\{m \in \mathcal{C}^{m*}_{M_1,M_2,M_3} : a_n\|m - m_0\|_{D_0}^S \le 1\big\},\\
\mathcal{C}^{\theta}(n) &:= \big\{\theta \in \Theta \cap B_{\theta_0}(1/2) : \hat{\lambda}_n^{-1/2}|\theta_0 - \theta| \le 1\big\},\\
\mathcal{C}_{M_1,M_2,M_3}(n) &:= \big\{(m,\theta) : \theta \in \mathcal{C}^{\theta}(n) \text{ and } m \in \mathcal{C}^{m}_{M_1,M_2,M_3}(n)\big\},\\
\mathcal{C}^{*}_{M_1,M_2,M_3} &:= \big\{(m,\theta) : \theta \in \Theta \cap B_{\theta_0}(1/2) \text{ and } m \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\big\}.
\end{aligned}$$

Let us consider the first term of (30). Fix $\delta > 0$. For every fixed M_1, M_2, and M_3,

$$\begin{aligned}
&\mathbb{P}\big(\big|\mathbb{G}_n\big[\hat{m}' \circ \hat{\theta}(m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta})K_1(\cdot;\hat{\theta})\big]\big| > \delta\big)\\
&\quad\le \mathbb{P}\big(\big|\mathbb{G}_n\big[\hat{m}' \circ \hat{\theta}(m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta})K_1(\cdot;\hat{\theta})\big]\big| > \delta, (\hat{m},\hat{\theta}) \in \mathcal{C}_{M_1,M_2,M_3}(n)\big)\\
&\qquad + \mathbb{P}\big((\hat{m},\hat{\theta}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)\big)\\
&\quad\le \mathbb{P}\Big(\sup_{(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)} \big|\mathbb{G}_n\big[m' \circ \theta(m_0 \circ \theta_0 - m \circ \theta)K_1(\cdot;\theta)\big]\big| > \delta\Big)\\
&\qquad + \mathbb{P}\big((\hat{m},\hat{\theta}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)\big). \qquad (31)
\end{aligned}$$

Recall that $(\hat{m},\hat{\theta})$ is a consistent estimator of (m_0,θ_0) and $\|\hat{m}'\|_\infty$ is $O_p(1)$; see Theorem 3. Furthermore, we have that both $\|\hat{m}\|_\infty$ and $J(\hat{m})$ are $O_p(1)$ (see Theorem 2) and $\hat{\lambda}_n^{-1/2}|\hat{\theta} - \theta_0| = o_p(1)$ (see Theorem 4). Thus for any $\varepsilon > 0$, there exists M_1, M_2, and M_3 (depending on ε) such that

$$\mathbb{P}\big((\hat{m},\hat{\theta}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)\big) \le \varepsilon,$$

for all sufficiently large n. Hence, it is enough to show that for the above choice of M_1, M_2, and M_3, we have

$$\mathbb{P}\Big(\sup_{(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)} \big|\mathbb{G}_n\big[m' \circ \theta(m_0 \circ \theta_0 - m \circ \theta)K_1(\cdot;\theta)\big]\big| > \delta\Big) \le \varepsilon$$

for sufficiently large n. Lemma 2 (proved in Section S.3.5 of the supplementary article [26]) shows this. Moreover, Lemma 3 (proved in Section S.3.6 of the supplementary article) shows that the second term on the right-hand side of (30) converges to zero in probability. Thus our proof is complete. □

[3]The notations with $*$ denote the classes of functions that do not depend on n while the ones with n denote shrinking neighborhoods around (m_0,θ_0).

Lemma 2. *Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, let us define two classes of functions from $\mathcal{X}$ to $\mathbb{R}^d$*

$$\mathcal{D}_{M_1,M_2,M_3}(n) := \big\{ m' \circ \theta (m_0 \circ \theta_0 - m \circ \theta) K_1(\cdot;\theta) : (m,\theta) \in \mathcal{C}_{M_1,M_2,M_3}(n) \big\},$$
$$\mathcal{D}^*_{M_1,M_2,M_3} := \big\{ m' \circ \theta (m_0 \circ \theta_0 - m \circ \theta) K_1(\cdot;\theta) : (m,\theta) \in \mathcal{C}^*_{M_1,M_2,M_3} \big\}.$$

$\mathcal{D}_{M_1,M_2,M_3}(n)$ is a Donsker class and

$$\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} \|f\|_{2,\infty} \le 2TM_2\big(a_n^{-1} + TM_2\hat{\lambda}_n^{1/2}\big) =: D_{M_1,M_2,M_3}(n). \tag{32}$$

Moreover, $J_{[]}(\gamma, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,2}) \lesssim \gamma^{1/2}$, where for any class of functions $\mathcal{F}$, $J_{[]}$ is the entropy integral (see e.g., Page 270, [58]) defined as

$$J_{[]}\big(\delta, \mathcal{F}, \|\cdot\|_{2,2}\big) := \int_0^\delta \sqrt{\log N_{[]}\big(t, \mathcal{F}, \|\cdot\|_{2,2}\big)}\, dt.$$

Finally, we have

$$\mathbb{P}\Big(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta \Big) \to 0 \quad \text{as } n \to \infty.$$

Lemma 3. *Let us define $U_{\theta,m} : \mathcal{X} \to \mathbb{R}^{d-1}$, $U_{\theta,m}(x) := m'(\theta^\top x) K_1(x;\theta)$. Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, let us define*

$$\mathcal{W}_{M_1,M_2,M_3}(n) := \big\{ U_{\theta,m} - U_{\theta_0,m_0} : (m,\theta) \in \mathcal{C}_{M_1,M_2,M_3}(n) \big\},$$
$$\mathcal{W}^*_{M_1,M_2,M_3} := \big\{ U_{\theta,m} - U_{\theta_0,m_0} : (m,\theta) \in \mathcal{C}^*_{M_1,M_2,M_3} \big\}.$$

Then $\mathcal{W}_{M_1,M_2,M_3}(n)$ is a Donsker class such that

$$\sup_{f \in \mathcal{W}_{M_1,M_2,M_3}(n)} \|f\|_{2,\infty} \le \big[2T^{3/2} M_3 \hat{\lambda}_n^{1/4} + 2Ta_n^{-1} + M_2(2T + \bar{M})\hat{\lambda}_n^{1/2} \big] =: W_{M_1,M_2,M_3}(n).$$

Moreover, $J_{[]}(\gamma, \mathcal{W}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,2}) \lesssim \gamma^{1/2}$. Hence, as $n \to \infty$, we have

$$\mathbb{P}\big(\big| \mathbb{G}_n \big[\epsilon (U_{\hat{\theta},\hat{m}} - U_{\theta_0,m_0}) \big] \big| > \delta \big) \to 0. \tag{33}$$

5. Simulation study

To investigate the finite sample performance of $(\hat{m}, \hat{\theta})$, we carry out several simulation experiments. We also compare the finite sample performance of the proposed estimator with the EFM estimator (estimating function method [7]), the EDR estimator (effective dimension reduction [21]), and the estimator proposed in [62] (denoted by WY). [7] compares the performance of the EFM estimator to existing estimators such as the refined minimum average variance estimator

(rMAVE) [65] and the EDR estimator and argues that EFM has improved overall performance compared to existing estimators. Thus we do not include the rMAVE estimator in our simulation study. The code to compute the EDR estimates can be found in the R package `EDR`. Moreover, the authors of [7] and [62] kindly provided us with the R codes to evaluate the EFM and the WY estimators, respectively. The codes used to implement our procedure are available in the `simest` package in R; see Kuchibhotla and Patra [25]. In what follows, we chose the penalty parameter $\hat{\lambda}_n$ for the PLSE through generalized cross validation, that is, choose $\hat{\lambda}_n$ by minimizing $\mathcal{T} : \mathbb{R} \to \mathbb{R}$

$$\mathcal{T}(\lambda) := \frac{Q_n(\hat{m}_\lambda, \hat{\theta}_\lambda)}{1 - n^{-1}\operatorname{trace}(A(\lambda))},$$

where

$$Q_n(m, \theta) := \frac{1}{n}\sum_{i=1}^{n}\big(y_i - m\big(\theta^\top x_i\big)\big)^2 \quad \text{and} \quad (\hat{m}_\lambda, \hat{\theta}_\lambda) := \operatorname*{arg\,min}_{(m,\theta)\in\mathcal{S}\times\Theta} \mathcal{L}_n(m, \theta; \lambda),$$

and $A(\lambda)$ is the *hat* matrix for $\hat{m}_\lambda$ (see, e.g., Sections 3.2 and 3.3 of [11] for a detailed description of $A(\lambda)$ and its connection to the generalized cross validation procedure); see [51] for an extensive discussion on why the generalized cross validation is an attractive choice for choosing the penalty parameter in the single index model. We choose $\hat{\lambda}_n$ by minimizing $\mathcal{T}(\cdot)$ over a grid of values that satisfy assumption **(A4)**. For all the other methods considered in the paper, we have used the suggested values of tuning parameters. In the following, we consider three different data generating mechanisms. The codes used for the simulation examples can be found at http://stat.ufl.edu/~rohitpatra/research. In the rest of the section, we use GCV to denote the PLSE to stress the fact that λ is chosen via the generalized cross validation procedure.

5.1. A simple model

We start with a simple model. Assume that $(X_1, X_2) \in \mathbb{R}^2$, $(X_1, X_2) \sim \text{Uniform}[-2, 2] \times [0, 1]$, $\epsilon \sim N(0, 0.5^2)$, and

$$Y = \big(\theta_0^\top X\big)^2 + \epsilon, \quad \text{where } \theta_0 = (1, -1)/\sqrt{2}. \tag{34}$$

Observe that for this example, $H_{\theta_0}^\top = [1, 1]/\sqrt{2}$ (see Section S.2.3 of the supplementary article [26]) and the analytic expression of the efficient information is

$$\tilde{I}_{\theta_0, m_0} = 4\operatorname{Var}(\epsilon)\mathbb{E}\big(\theta_0^\top X H_{\theta_0}^\top\big[X - \mathbb{E}\big(X|\theta_0^\top X\big)\big]\big)^2 = 4\operatorname{Var}(\epsilon)\mathbb{E}\big|\big(\theta_0^\top X\big)^2\big[H_{\theta_0}^\top \operatorname{Var}\big(X|\theta_0^\top X\big)H_{\theta_0}\big]\big|.$$

Using the above expression, we calculated the asymptotic variance of $\sqrt{n}(\hat{\theta}_1 - \theta_{0,1})$ to be 0.328. Figure 1 contains the quantile-quantile plot of the four estimators considered in this section based on 500 replications of random samples from (34) with $n = 500$.

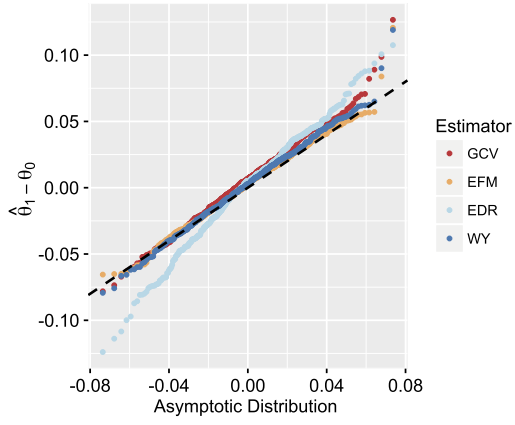

Figure 1. Quantile quantile plot of $\hat{\theta}_1 - \theta_0$ from 500 replications with the true asymptotic distribution of the $\hat{\theta}_1 - \theta_{0,1}$ on the X-axis when we have 500 i.i.d. samples from (34).

5.2. Dependent covariates

We now consider a simulation scenario where covariates are dependent and the predictor $X \in \mathbb{R}^6$ contains discrete components. More precisely, $(X_1, \ldots, X_6)$ is generated according to the following law: $(X_1, X_2) \sim \text{Uniform}[-1, 1]^2$, $X_3 := 0.2X_1 + 0.2(X_2 + 2)^2 + 0.2Z_1$, $X_4 := 0.1 + 0.1(X_1 + X_2) + 0.3(X_1 + 1.5)^2 + 0.2Z_2$, $X_5 \sim \text{Bernoulli}(\exp(X_1)/\{1 + \exp(X_1)\})$, and $X_6 \sim \text{Bernoulli}(\exp(X_2)/\{1 + \exp(X_2)\})$. Here Z_1 and Z_2 are two independent Uniform$[-1, 1]$ random variables independent of (X_1, X_2). Finally, we let

$$Y = \left(\theta_0^\top X\right)^2 + \epsilon, \tag{35}$$

where θ_0 is $(1.3, -1.3, 1, -0.5, -0.5, -0.5)/\sqrt{5.13}$. In the following, we consider three different scenarios based on different error distributions:

(**2.1**) $\epsilon \sim N(0, 1)$, (Homoscedastic, Gaussian Error)
(**2.2**) $\epsilon | X \sim N(0, \log(2 + (X^\top \theta_0)^2))$, (Heteroscedastic, Gaussian Error)
(**2.3**) $\epsilon | \xi \sim (-1)^\xi \text{Beta}(2, 3)$, where $\xi \sim \text{Ber}(0.5)$. (Homoscedastic, Non-Gaussian Error)

The results of our simulations based 500 replications with sample size $n = 200$ from (35) for each of the above three error distributions is displayed in Figure 2. The first two rows of Figure 2 show the box plots of L_1 and L_2 loss of GCV, EFM, EDR, and WY estimators of θ_0. In the third row of Figure 2, we display the box plot of the in-sample $L_2(\mathbb{P}_n)$ loss ($\|\hat{m} \circ \theta_0^\top - m_0 \circ \theta_0^\top\|_n$) for the GCV and the WY estimators. EFM and EDR are not included because they do not provide estimators for the link function.[4]

Gu and Yang [12] show that for any root-n consistent estimator $\hat{\theta}$ of the index estimator, the kernel (or Nadaraya–Watson) regression estimator on the data $\{(\hat{\theta}^\top X_i, Y_i), 1 \le i \le n\}$ is

[4]Our proposed method and [62] provide estimators for both the link function and the index parameter.

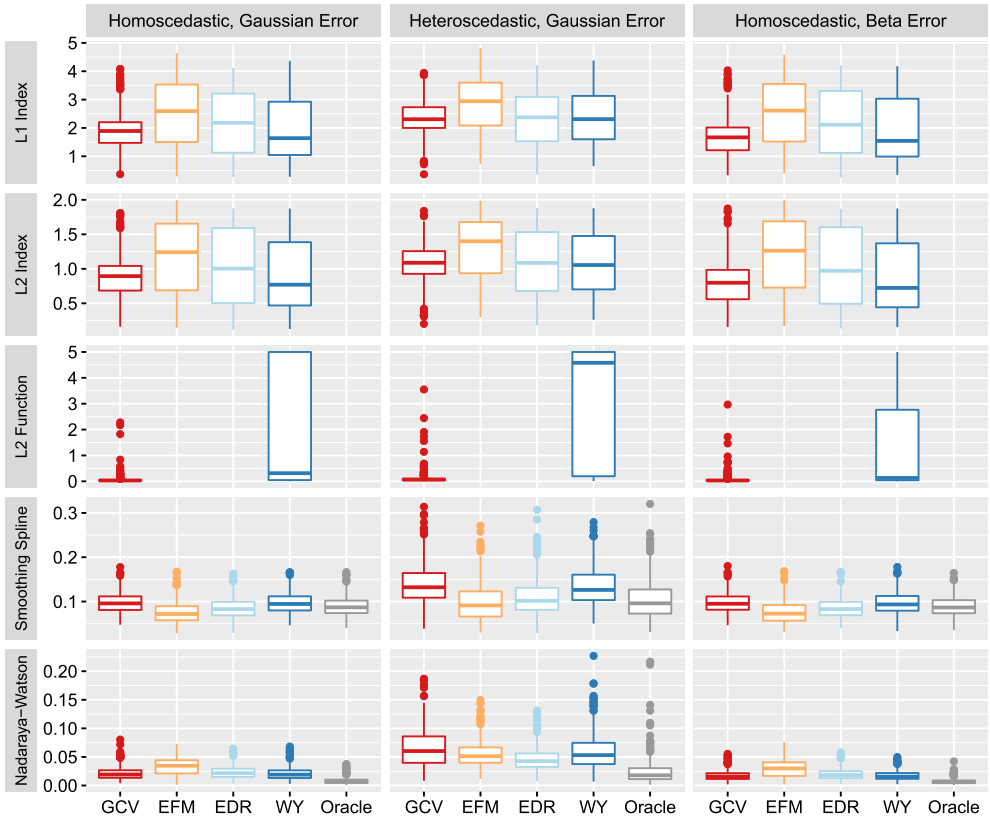

Figure 2. Box plots (over 500 replications) of various errors based on 200 observations from models (**2.1**), (**2.2**), and (**2.3**) in the left, the middle, and the right columns, respectively. First two rows display L_1 and L_2 errors of estimates of θ_0. The third row corresponds to $\|\hat{m} \circ \theta_0 - m_0 \circ \theta_0\|_n$ for the estimators proposed in Section 2 and [62]. The fourth and fifth rows corresponds to $\|\tilde{m} \circ \theta_0 - m_0 \circ \theta_0\|_n$ for one-dimensional smoothing splines and Nadaraya–Watson estimators based on the estimated index $\{(\hat{\theta}^\top X_i, Y_i), 1 \le i \le n\}$, respectively. In the fourth and fifth rows, "Oracle" refers to the estimator of link function based on true index $\{(\theta_0^\top X_i, Y_i), 1 \le i \le n\}$.

asymptotically indistinguishable from the kernel estimator based on $\{(\theta_0^\top X_i, Y_i), 1 \le i \le n\}$. This oracle type property led us to compute the estimators of the link function based the data $\{(\hat{\theta}^\top X_i, Y_i), 1 \le i \le n\}$ for GCV, EFM, EDR, and WY.[5] The "Oracle" in Figure 2 refers to the estimator of link function based on the true θ_0. The plot of the error (see fifth row of Figure 2) in the estimation of the link function based on the Nadaraya–Watson estimator provides numerical confirmation of the oracle property proved in [12]; we used the `np` package [17] to compute the bandwidth choice for the nonparametric regression estimator. We also estimate the one-dimensional

[5]Recall that GCV, EFM, EDR, and WY are all root-n consistent.

link function based on smoothing splines[6] applied to the data $\{(\hat{\theta}^\top X_i, Y_i), 1 \le i \le n\}$; see fourth row of Figure 2. The results of [12] do not directly imply a similar oracle type phenomenon for smoothing splines based estimators. However, the fourth row of Figure 2 provides some numerical evidence for this oracle type property for the smoothing splines estimators. The proof of the oracle type property developed in [12] crucially uses the smoothness of the Nadaraya–Watson estimator (as a function of the index) and we have not been able to extend it to the case of smoothing splines estimators in single index models.[7]

The relative poor performance of EDR, EFM, and WY in estimating θ_0 can possibly be attributed to the dependency between covariates. Scenarios **(2.1)** and **(2.2)** are similar to simulation settings considered in [34] and [36], respectively. The codes to compute the estimator proposed in [34] were not available to us.

5.3. High dimensional covariates

For the final simulation scenario, we consider a setting similar to that of Example 4 in Cui et al. [7], Section 3.2. We consider d-variate covariates for $d = 10, 50$, and 100. For each d, we assume that $X \sim \text{Uniform}[0,5]^d$, $\epsilon \sim N(0, 0.2^2)$, $\theta_0 = (2, 1, \mathbf{0}_{d-2})^\top/\sqrt{5}$, and have $n = 400$ observations from the following model:

$$Y = \sin\big(aX^\top\theta_0\big) + \epsilon, \quad \text{where } a = \pi/2, 3\pi/4, \text{and } 3\pi/2. \tag{36}$$

Note that here a higher value of a represents a more oscillating link function. Figure 3 summarizes the finite sample performance of the estimators over 500 replications. The performance of all the estimators worsen as the a increases. When a is $\pi/2$ or $3\pi/4$, GCV significantly outperforms the estimators considered in the simulation study. The IQR bars for the GCV in the first

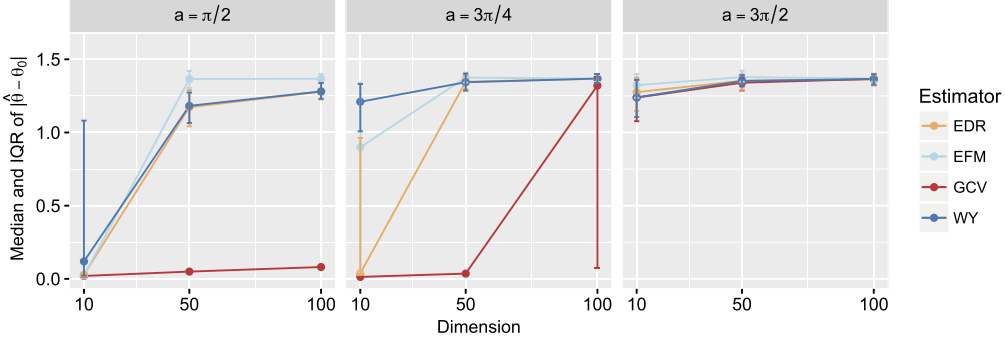

Figure 3. The quartiles of $|\hat{\theta} - \theta_0|$ from 500 replications for $n = 400$ from (36).

[6]We used `smooth.spline` command in `R` and choose λ by the GCV procedure proposed in [11].
[7]This is a very interesting research direction and we plan to study it in the near future.

two panels of Figure 3 are not visible because they are very small (relative to the scale of the plot).

6. Real data analysis

6.1. Car mileage data

In this sub-section, we model the mileages (Y) of 392 cars using the covariates (X): displacement (D), weight (W), acceleration (A), and horsepower (H); see http://lib.stat.cmu.edu/datasets/cars.data for the data set. For our data analysis, we have scaled and centered each of covariates to have mean 0 and variance 1. To compare the prediction capabilities of the linear model to that of the single index model for this data set, we randomly split the data set into a training set of size 260 and a test set of size 132 and compute the prediction error for both the linear model fit and the single index model fit. The average prediction error over 1000 such random splits was 4.3 for the linear model fit and 3.8 for the single index model fit. The results indicate that the single index model is a better fit.

In the left panel of Figure 4, we have the scatter plot of $\{(\hat{\theta}^\top x_i, y_i)\}_{i=1}^{392}$ overlaid with the plot of $\hat{m}(\hat{\theta}^\top x)$. In Table 1, we display the estimates of θ_0 based on the methods considered in the paper. The MAVE, the EFM estimator, and the GCV give similar estimates while the EDR gives a different estimate of the index parameter.

6.2. Ozone concentration data

For the second real data example, we study the relationship between Ozone concentration (Y) and three meteorological variables (X): radiation level (R), wind speed (W), and temperature

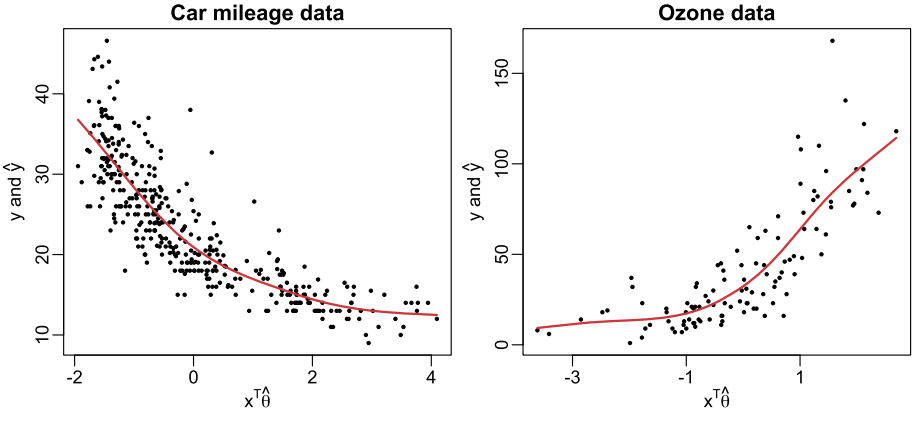

Figure 4. Scatter plots of $\{(x_i^\top \hat{\theta}, y_i)\}_{i=1}^n$ overlaid with the plots of $\hat{m}$ (in solid red line) for the two real data sets considered. Left panel: the car mileage data (Section 6.1); right panel: Ozone concentration data (Section 6.2).

Table 1. Estimates of θ_0 for the data sets in Sections 6.1 and 6.2

	Car mileage data				Ozone data		
Method	D	W	A	H	R	W	T
GCV	0.48	0.18	0.11	0.85	0.32	−0.62	0.71
EFM	0.44	0.18	0.13	0.87	0.29	−0.60	0.75
EDR	0.33	0.11	0.15	0.93	0.22	−0.64	0.73
rMAVE	0.48	0.17	0.17	0.84	0.31	−0.58	0.75

(T). The data consists of 111 days of complete measurements from May to September, 1973, in New York city. The data set can be found in the `EnvStats` package in R. [68] fit a linear model, an additive model, and a fully nonparametric model and conclude that the single index model fits the data best. To fit a single index model to the data [68] fix 10 knots and fit cubic penalized splines to the data. It should be noted that observations from consecutive days are not independent. However in our analysis, we have ignored this dependence; see [68] for a similar analysis. The right panel of Figure 4 shows the scatter plot of $\hat{\theta}^\top X$ and Y overlaid with the plot of $\hat{m}(\hat{\theta}^\top X)$. As in the previous example, we have scaled and centered each of the covariates such that they have mean 0 and variance 1. We see that all the considered methods in the paper give similar estimates for θ_0; see Table 1.

7. Concluding remarks

In this paper, we propose a simple penalized least squares based estimator $(\hat{m}, \hat{\theta})$ for the unknown link function, m_0, and the index parameter, θ_0, in the single index model under mild smoothness assumptions on m_0. We prove that $\hat{m}$ is rate optimal (for the given smoothness) and $\hat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal. Moreover under homoscedastic errors, we show that $\hat{\theta}$ is efficient in the sense of [3]. We have developed the R package `simest` to compute the proposed estimators. We observe that the PLSE has superior finite sample performance compared to most competing methods.

Several interesting future directions follow. Estimation and inference adapting to the smoothness of the link function is an interesting direction. [29] proposes an estimator for the single index model that adapts to the smoothness of the true link function, but the estimator depends on true (unknown) density of X and requires independence between ϵ and X; also see [28]. In the context of one dimensional smoothing splines Györfi et al. [13], Chapter 21, consider adaptation to smoothness using complexity regularization and extension of such a procedure to the case of single index model is an interesting direction of future research. In line with the recent literature on high-dimensional asymptotics in the single index model [30,60,71], it would be interesting to prove analogues of our results in a finite sample setting and under sparsity inducing regularization of the index parameter. [30,47] consider variable selection in the single index model via a (additional) SCAD penalty (on the index parameter) on local linear and regression splines based estimation methods, respectively. They suggest that for the single index model, SCAD based variable selection methods have better performance when compared to LASSO based methods

studied in [60,71]. Variable selection by incorporating a SCAD penalty on (4) is an exciting direction of research and we plan to pursue this in the near future.

Appendix A: Proofs of results in Section 3

We start with two useful lemmas concerning the properties of functions in $\mathcal{S}$.

Lemma 4 (Lemma 3.6 of [42]). *Let $m \in \{g \in \mathcal{S} : J(g) < \infty\}$. Then $|m'(s) - m'(s_0)| \le J(m)|s - s_0|^{1/2}$ for every $s, s_0 \in D$.*

Lemma 5. *Let $m \in \{g \in \mathcal{S} : J(g) < \infty$ and $\|g\|_\infty \le M\}$, where M is a finite constant. Then*

$$\|m'\|_\infty \le 2M/\varnothing(D) + \big(1 + J(m)\big)\varnothing(D)^{1/2},$$

where $\varnothing(D)$ is the diameter of D. Moreover if $\varnothing(D) < \infty$, then

$$\|m'\|_\infty \le C\big(1 + J(m)\big),$$

where C is a finite constant depending only on M and $\varnothing(D)$.

Proof. Fix $s_0 \in D$. Integrating the inequality

$$-J(m)|t - s_0|^{1/2} \le m'(t) - m'(s_0) \le J(m)|t - s_0|^{1/2}$$

with respect to t, we get

$$\big|m(s) - m(s_0) - m'(s_0)(s - s_0)\big| \le J(m)\varnothing(D)^{3/2},$$

where $\varnothing(D)$ is the diameter of D. Since $\|m\|_\infty \le M$, we get that

$$\big|m'(s_0)(s - s_0)\big| \le 2M + J(m)\varnothing(D)^{3/2}.$$

If we choose s such that $|s - s_0| = \varnothing(D)/2$, then we have

$$\|m'\|_\infty \le 2M/\varnothing(D) + \big(1 + J(m)\big)\varnothing(D)^{1/2}.$$

The rest of the lemma follows by choosing $C = 2M/\varnothing(D) + \varnothing(D)^{1/2}$. □

A.1. Proof of Theorem 2

Our proof of Theorem 2 is along the lines of the proofs of Lemma 3.1 in [37] and Theorem 10.2 in [56]. Since $(\hat{m}, \hat{\theta})$ minimizes $Q_n(m, \theta) + \hat{\lambda}_n^2 J^2(m)$, we have

$$Q_n(\hat{m}, \hat{\theta}) + \hat{\lambda}_n^2 J^2(\hat{m}) \le Q_n(m_0, \theta_0) + \hat{\lambda}_n^2 J^2(m_0). \tag{37}$$

Observe that by definition of $Q_n(m,\theta)$, we have that (37) implies

$$\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n^2+\hat{\lambda}_n^2 J^2(\hat{m})=\frac{2}{n}\sum_{i=1}^n\epsilon_i\big(\hat{m}\big(\hat{\theta}^\top x_i\big)-m_0\big(\theta_0^\top x_i\big)\big)+\hat{\lambda}_n^2 J^2(m_0).$$

To find the rate of convergence of $\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n$ we will try to find upper bounds for $\sum_{i=1}^n\epsilon_i(\hat{m}(\hat{\theta}^\top x_i)-m_0(\theta_0^\top x_i))$ in terms of $\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n$ (modulus of continuity); see Section 1 of [55] for a similar proof technique. To be able to find such a bound, we first study the behavior of $\hat{m}\circ\hat{\theta}$. Observe that by Cauchy–Schwarz inequality we have

$$\begin{aligned}
&Q_n(m_0,\theta_0)-Q_n(\hat{m},\hat{\theta})\\
&=\frac{2}{n}\sum_{i=1}^n\epsilon_i\big(\hat{m}\big(\hat{\theta}^\top x_i\big)-m_0\big(\theta_0^\top x_i\big)\big)-\frac{1}{n}\sum_{i=1}^n\big(\hat{m}\big(\hat{\theta}^\top x_i\big)-m_0\big(\theta_0^\top x_i\big)\big)^2\\
&\le\left(\frac{4}{n}\sum_{i=1}^n\epsilon_i^2\right)^{1/2}\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n-\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n^2. \qquad (38)
\end{aligned}$$

Note that by **(A3)**, $(1/n)\sum_{i=1}^n\epsilon_i^2=O(1)$ almost surely. On the other hand, since $(\hat{m},\hat{\theta})$ minimizes $Q_n(m,\theta)+\hat{\lambda}_n^2 J^2(m)$, we have

$$Q_n(m_0,\theta_0)-Q_n(\hat{m},\hat{\theta})\ge\hat{\lambda}_n^2\big(J^2(\hat{m})-J^2(m_0)\big)\ge-\hat{\lambda}_n^2 J^2(m_0)\ge o_p(1), \qquad (39)$$

as $\hat{\lambda}_n=o_p(1)$. Combining (38) and (39), we have

$$\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n^2\le\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n O_p(1)+o_p(1).$$

Thus, we have $\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n=O_p(1)$. We also have $\|\hat{m}\circ\hat{\theta}\|_n=O_p(1)$ as $\|m_0\circ\theta_0\|_\infty<\infty$.

We will now use the Sobolev embedding theorem to get a bound on $\|\hat{m}\|_\infty$ in terms of $J(\hat{m})$.

Lemma 6 (Sobolev embedding theorem, Page 85, [44]). *Let $m: I\to\mathbb{R}$ ($I\subset\mathbb{R}$ is an interval) be a function such that $J(m)<\infty$. We can write*

$$m(t)=m_1(t)+m_2(t),$$

with $m_1(t)=\beta_1+\beta_2 t$ and $\|m_2\|_\infty\le J(m)\varnothing(I)$.

Thus, by the above lemma, we can find functions $\hat{m}_1$ and $\hat{m}_2$ such that

$$\hat{m}(t)=\hat{m}_1(t)+\hat{m}_2(t),$$

where $\hat{m}_1=\hat{\beta}_1+\hat{\beta}_2 t$, and $\|\hat{m}_2\|_\infty\le J(\hat{m})\varnothing(D)$. Then

$$\frac{\|\hat{m}_1\circ\hat{\theta}\|_n}{1+J(m_0)+J(\hat{m})}\le\frac{\|\hat{m}\circ\hat{\theta}\|_n}{1+J(m_0)+J(\hat{m})}+\frac{\|\hat{m}_2\circ\hat{\theta}\|_n}{1+J(m_0)+J(\hat{m})}=O_p(1). \qquad (40)$$

Let us define

$$\mathbb{A}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \varphi_\theta(X_i)\varphi_\theta^\top(X_i) \quad \text{and} \quad A(\theta) := \int \varphi_\theta(x)\varphi_\theta(x)^\top \, dP_X(x),$$

where $\varphi_\theta(x) := (1, \theta^\top x)^\top$. Furthermore, we denote the smallest eigenvalues of $\mathbb{A}_n(\theta)$ and $A(\theta)$ by $\vartheta_n(\theta)$ and $\vartheta(\theta)$ respectively. Since Θ is a bounded subset of $\mathbb{R}^d$, by the Glivenko–Cantelli theorem, we have

$$\sup_{\theta\in\Theta} \big|\vartheta_n(\theta) - \vartheta(\theta)\big| = o_p(1).$$

Let $\vartheta_0 := \min_{\theta\in\Theta} \vartheta(\theta)$. By assumption **(A0)** and and the fact that $|\theta| = 1$, we have $\det(A(\theta)) = \theta^\top \operatorname{Var}(X)\theta$ and $\inf_{\theta\in\Theta} \det(A(\theta)) > 0$. It follows that $\vartheta_0 > 0$ and

$$\begin{aligned}
\|\hat{m}_1 \circ \hat{\theta}\|_n^2 &= (\hat{\beta}_1, \hat{\beta}_2)\mathbb{A}_n(\theta)(\hat{\beta}_1, \hat{\beta}_2)^\top \\
&\geq \vartheta_n(\hat{\theta})\big(\hat{\beta}_1^2 + \hat{\beta}_2^2\big) \\
&= \big[\vartheta_n(\hat{\theta}) - \vartheta(\hat{\theta})\big]\big(\hat{\beta}_1^2 + \hat{\beta}_2^2\big) + \vartheta(\hat{\theta})\big(\hat{\beta}_1^2 + \hat{\beta}_2^2\big) \\
&\geq o_p\big(\hat{\beta}_1^2 + \hat{\beta}_2^2\big) + \vartheta_0\big(\hat{\beta}_1^2 + \hat{\beta}_2^2\big) \\
&\geq o_p\big(\hat{\beta}_1^2 + \hat{\beta}_2^2\big) + \vartheta_0 \max(\hat{\beta}_1, \hat{\beta}_2)^2.
\end{aligned}$$

Thus by (40) we have

$$\frac{\max(\hat{\beta}_1, \hat{\beta}_2)}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{41}$$

Moreover, since D is a bounded set, by (41) we have $\|\hat{m}_1\|_\infty/(1 + J(m_0) + J(\hat{m})) = O_p(1)$. Combining this with Lemma 6, we get

$$\frac{\|\hat{m}\|_\infty}{1 + J(m_0) + J(\hat{m})} \leq \frac{\|\hat{m}_1\|_\infty}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2\|_\infty}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{42}$$

Now define the class of functions

$$\mathcal{B}_C := \left\{ \frac{m \circ \theta - m_0 \circ \theta_0}{1 + J(m_0) + J(m)} : m \in \mathcal{S}, \theta \in \Theta, \text{ and } \frac{\|m\|_\infty}{1 + J(m_0) + J(m)} \leq C \right\}.$$

Observe that by (42), we can find a C_ε such that

$$\mathbb{P}\left(\frac{\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0}{1 + J(m_0) + J(\hat{m})} \in \mathcal{B}_{C_\varepsilon} \right) \geq 1 - \varepsilon, \quad \forall n. \tag{43}$$

The following lemma in [56] gives a upper bound for $\sum_{i=1}^n \epsilon_i g(x_i)$, in terms of entropy of the class of functions g.

Lemma 7 (Lemma 8.4, [56]). *Suppose $\mathcal{G}$ be a class of functions. If $\log N_{[]}(\delta, \mathcal{G}, \|\cdot\|_\infty) \le A\delta^{-\alpha}$, $\sup_{g\in\mathcal{G}} \|g\|_n \le R$, and ϵ satisfies assumption (A3), for some constants $0 < \alpha < 2$, A, and R. Then for some constant c, we have for all $T \ge c$,*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i g(x_i)|}{\|g\|_n^{1-\frac{\alpha}{2}}} \ge T \right) \le c \exp\left[\frac{-T^2}{c^2} \right].$$

Lemma 8, proved in Section S.2.1 of the supplementary article [26], finds the bracketing number for the class of functions $\mathcal{B}_C$.

Lemma 8. *For every fixed positive M_1, M_2, and C, we have*

$$\log N\big(\delta, \mathcal{B}_C, \|\cdot\|_\infty\big) \lesssim \delta^{-1/2}.$$

In the view of (43), Lemmas 7 and 8 allow us to conclude

$$\frac{(1/n)\sum_{i=1}^n \epsilon_i (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))}{\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4}(1 + J(m_0) + J(\hat{m}))^{1/4}} = O_p\big(n^{-1/2}\big). \tag{44}$$

Together, (39) and (44) imply

$$\begin{aligned}
&\hat{\lambda}_n^2\big(J^2(\hat{m}) - J^2(m_0)\big) \\
&\quad\le Q_n(m_0,\theta_0) - Q_n(\hat{m},\hat{\theta}) \\
&\quad= \frac{2}{n}\sum_{i=1}^n \big(y_i - m_0\big(\theta_0^\top x_i\big)\big)\big(\hat{m}\big(\hat{\theta}^\top x_i\big) - m_0\big(\theta_0^\top x_i\big)\big) - \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2 \\
&\quad\le \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4}\big(1 + J(m_0) + J(\hat{m})\big)^{1/4} O_p\big(n^{-1/2}\big) - \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2. \qquad (45)
\end{aligned}$$

We will now consider two cases.

Case 1: Suppose $J(\hat{m}) > 1 + J(m_0)$. By the proof of Theorem 10.2 of [57] with $\nu = 2$ and $\alpha = 1/2$, we have that

$$J(\hat{m}) = O_p\big(n^{-1/2}\big)\hat{\lambda}_n^{-5/4} \quad \text{and} \quad \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n = O_p\big(n^{-1/2}\big)\hat{\lambda}_n^{-1/4}.$$

However, by assumption (A3), we have that $\hat{\lambda}_n^{-1} = O_p(n^{2/5})$. Hence, the conclusion follows.

Case 2: When $J(\hat{m}) \le 1 + J(m_0)$, (45) implies,

$$\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2 \le \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4}\big(1 + J(m_0)\big)^{1/4} O_p\big(n^{-1/2}\big) + \hat{\lambda}_n^2 J^2(m_0).$$

Therefore, it follows that either

$$\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n \le \big(1 + J(m_0)\big)^{1/5} O_p\big(n^{-2/5}\big) = O_p(\hat{\lambda}_n)$$

or

$$\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n \le O_p(1)\hat{\lambda}_n J(m_0) = O_p(\hat{\lambda}_n)J(m_0).$$

Thus, we have that $J(\hat{m}) = O_p(1)$, $\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|_n = O_p(\hat{\lambda}_n)$, and, by (42), $\|\hat{m}\|_\infty = O_p(1)$. To find the rates of convergence of $\|\hat{m}\circ\hat{\theta}-m_0\circ\theta_0\|$, we use the following lemma.

Lemma 9 (Lemma 5.16, [56]). *Suppose $\mathcal{G}$ is a class of uniformly bounded functions and for some $0<\nu<2$,*

$$\sup_{\delta>0}\delta^{\nu}\log N_{[\,]}\big(\delta,\mathcal{G},\|\cdot\|_\infty\big)<\infty.$$

Then for every given $\alpha>0$ there exists a constant $C>0$ such that

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{g\in\mathcal{G},\|g\|>Cn^{-1/(2+\nu)}}\left|\frac{\|g\|_n}{\|g\|}-1\right|>\alpha\right)=0.$$

Appendix B: Proofs of results in Section 4

B.1. Proof of Theorem 6

We will first show that $\xi_t(u;\theta,\eta,m)$ is a valid submodel. Note that $\phi_{\theta,\eta,0}(u+(\theta-\theta)^\top h_\theta(u)) = u$, $\forall u\in D$. Hence,

$$\xi_0\big(\theta^\top x;\theta,\eta,m\big) = m\circ\phi_{\theta,\eta,0}\big(\theta^\top x\big) = m\big(\theta^\top x\big).$$

Now we will prove that $J^2(\xi_t(\cdot;\theta,\eta,m))<\infty$. Let us define

$$\psi_{\theta,\eta,t}(u) := \phi_{\theta,\eta,t}\big(u+\big(\theta-\zeta_t(\theta,\eta)\big)^\top h_\theta(u)\big),$$

then $\xi_t(u;\theta,\eta,m) = m\circ\psi_{\theta,\eta,t}(u)$ Observe that

$$\begin{aligned}
&J^2\big(\xi_t(\cdot;\theta,\eta,m)\big)\\
&\quad=\int_D\big|\xi_t''(u;\theta,\eta,m)\big|^2\,du\\
&\quad=\int_D\big[m''\circ\psi_{\theta,\eta,t}(u)\psi'_{\theta,\eta,t}(u)^2+m'\circ\psi_{\theta,\eta,t}(u)\psi''_{\theta,\eta,t}(u)\big]^2\,du\\
&\quad=\int_D\big[m''(u)\big(\psi'_{\theta,\eta,t}\circ\psi^{-1}_{\theta,\eta,t}(u)\big)^2+m'(u)\psi''_{\theta,\eta,t}\circ\psi^{-1}_{\theta,\eta,t}(u)\big]^2\frac{du}{\psi'_{\theta,\eta,t}\circ\psi^{-1}_{\theta,\eta,t}(u)},
\end{aligned}$$

where $\psi'_{\theta,\eta,t}(u) = \frac{\partial}{\partial u}\psi_{\theta,\eta,t}(u)$. Thus, we have that $J^2(\xi_t(\cdot;\theta,\eta,m)) = O(1)$ whenever $J(m) = O(1)$, $\|m\|_\infty = O(1)$, and t in a small neighborhood of 0 (as $\psi_{\theta,\eta,t}(\cdot)$ is a strictly increasing

function when t is small). Next, we evaluate $\partial \xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)/\partial t$ to help with the calculation of the score function for the submodel $\{\zeta_t(\theta,\eta),\xi_t(\cdot;\theta,\eta,m)\}$. Note that

$$\begin{aligned}
&\frac{\partial}{\partial t}\xi_t\big(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m\big)\\
&\quad=\frac{\partial}{\partial t}m\circ\phi_{\theta,\eta,t}\big(\zeta_t(\theta,\eta)^\top x+\big(\theta-\zeta_t(\theta,\eta)\big)^\top h_\theta\big(\zeta_t(\theta,\eta)^\top x\big)\big)\\
&\quad=m'\circ\phi_{\theta,\eta,t}\big(\zeta_t(\theta,\eta)^\top x+\big(\theta-\zeta_t(\theta,\eta)\big)^\top h_\theta\big(\zeta_t(\theta,\eta)^\top x\big)\big)\\
&\qquad\times\Big[\dot\phi_{\theta,\eta,t}\big[\zeta_t(\theta,\eta)^\top x+\big[\theta-\zeta_t(\theta,\eta)\big]^\top h_\theta\big(\zeta_t(\theta,\eta)^\top x\big)\big]\\
&\qquad+\phi'_{\theta,\eta,t}\big[\zeta_t(\theta,\eta)^\top x+\big(\theta-\zeta_t(\theta,\eta)\big)^\top h_\theta\big(\zeta_t(\theta,\eta)^\top x\big)\big]\frac{\partial\zeta_t(\theta,\eta)}{\partial t}^\top\big[x\\
&\qquad+\big(\theta-\zeta_t(\theta,\eta)\big)^\top h'_\theta\big(\zeta_t(\theta,\eta)^\top x\big)x-h_\theta\big(\zeta_t(\theta,\eta)^\top x\big)\big]\Big],
\end{aligned}$$

where $\dot\phi_{t,\theta}(u)=\partial\phi_{\theta,\eta,t}(u)/\partial t$. We will now show that the score function of the submodel $\{t,\xi_t(\cdot;,\theta,\eta,m)\}$ is $\tilde\ell_{\theta,m}(y,x)$. Using the facts that $\phi'_{\theta,\eta,t}(u)=1$ and $\dot\phi_{\theta,\eta,t}(u)=0$ for all $u\in D$ (follows from the definition (28)) and $\partial\zeta_t(\theta,\eta)/\partial t=(-2t/\sqrt{1-t^2|\eta|^2})\theta+H_\theta\eta$, we get

$$\begin{aligned}
&\frac{\partial}{\partial t}\big(y-\xi_t\big(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m\big)\big)^2\Big|_{t=0}\\
&\quad=-2\big(y-\xi_t\big(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m\big)\big)\frac{\partial\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)}{\partial t}\Big|_{t=0}\\
&\quad=-2\big(y-m\big(\theta^\top x\big)\big)m'\big(\theta^\top x\big)\eta^\top H_\theta^\top\big(x-h_\theta\big(\theta^\top x\big)\big).
\end{aligned}$$

Observe that $(\hat m,\hat\theta)$ minimizes the penalized loss function in (5) and $\xi_0(\zeta_0(\hat\theta,\eta)^\top x;\hat\theta,\eta,\hat m)=\hat m(\hat\theta^\top x)$, where $\zeta_t(\hat\theta,\eta)=\sqrt{1-t^2|\eta|^2}\hat\theta+sH_{\hat\theta}\eta$. Hence, for every $\eta\in\mathbb{R}^{d-1}$, the function

$$t\mapsto\frac{1}{n}\sum_{i=1}^n\big(y_i-\xi_t\big(\zeta_t(\hat\theta,\eta)^\top x;\hat\theta,\eta,\hat m\big)\big)^2+\hat\lambda_n^2\int_D\Big|\frac{\partial^2}{\partial u^2}\xi_t(u;\hat\theta,\eta,\hat m)\Big|^2\,du \tag{46}$$

on a some small neighborhood of 0 (that depends on η) is minimized at $t=0$. Moreover, using some tedious algebra it can be shown that $J^2(\xi_t(\cdot;\theta,\eta,m))$ is differentiable and

$$\frac{\partial}{\partial t}J^2\big(\xi_t(\cdot;\theta,\eta,m)\big)\Big|_{t=0}\lesssim\int_D\big|m''(p)\big|^2\,dP.$$

This we have that the function in (46) is differentiable at $t = 0$. Conclude that, for all $\eta \in \mathbb{R}^{d-1}$ we have

$$\eta^\top \mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{m}} - \hat{\lambda}_n^2 \frac{\partial J^2(\xi_t(\cdot;\theta,\eta,m))}{\partial t}\bigg|_{t=\hat{\theta}} = 0.$$

The proof of the theorem is now complete as $\hat{\lambda}_n^2 = o_p(n^{-1/2})$.

Acknowledgements

We would like to thank Bodhisattva Sen for many helpful discussions and for his help in writing this paper. We would also like to thank Promit Ghosal and David A. Hirshberg for helpful discussions. Finally, we would like thank the anonymous referees, associate editor, and Editor for their work in refereeing the paper. Their suggestions led to an improved paper.

Supplementary Material

Supplement to "Efficient estimation in single index models through smoothing splines" (DOI: 10.3150/19-BEJ1183SUPP; .pdf). The supplement contains proofs not provided in the main text or the appendices.

References

[1] Antoniadis, A., Grégoire, G. and McKeague, I.W. (2004). Bayesian estimation in single-index models. *Statist. Sinica* **14** 1147–1164. MR2126345

[2] Beresteanu, A. (2004). Nonparametric estimation of regression functions under restrictions on partial derivatives. Technical report.

[3] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer. MR1623559

[4] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489. MR1467842 https://doi.org/10.2307/2965697

[5] Chang, Z., Xue, L. and Zhu, L. (2010). On an asymptotically more efficient estimation of the single-index model. *J. Multivariate Anal.* **101** 1898–1901. MR2651964 https://doi.org/10.1016/j.jmva.2010.02.005

[6] Chaudhuri, P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression. *Ann. Statist.* **25** 715–744. MR1439320 https://doi.org/10.1214/aos/1031833670

[7] Cui, X., Härdle, W.K. and Zhu, L. (2011). The EFM approach for single-index models. *Ann. Statist.* **39** 1658–1688. MR2850216 https://doi.org/10.1214/10-AOS871

[8] Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference* **136** 730–769. MR2181975 https://doi.org/10.1016/j.jspi.2004.09.006

[9] Dontchev, A.L., Qi, H.-D., Qi, L. and Yin, H. (2002). A Newton method for shape-preserving spline interpolation. *SIAM J. Optim.* **13** 588–602. MR1951036 https://doi.org/10.1137/S1052623401393128

[10] Elfving, T. and Andersson, L.-E. (1988). An algorithm for computing constrained smoothing spline functions. *Numer. Math.* **52** 583–595. MR0945101 https://doi.org/10.1007/BF01400893
[11] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. *Monographs on Statistics and Applied Probability* **58**. London: CRC Press. MR1270012 https://doi.org/10.1007/978-1-4899-4473-3
[12] Gu, L. and Yang, L. (2015). Oracally efficient estimation for single-index link function with simultaneous confidence band. *Electron. J. Stat.* **9** 1540–1561. MR3376116 https://doi.org/10.1214/15-EJS1051
[13] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. New York: Springer. MR1920390 https://doi.org/10.1007/b97848
[14] Hall, P. and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29** 624–647. MR1865334 https://doi.org/10.1214/aos/1009210683
[15] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171 https://doi.org/10.1214/aos/1176349020
[16] Härdle, W. and Liang, H. (2007). Partially linear models. In *Statistical Methods for Biostatistics and Related Fields* 87–103. Berlin: Springer. MR2376405 https://doi.org/10.1007/978-3-540-32691-5_5
[17] Hayfield, T. and Racine, J.S. (2008). Nonparametric econometrics: The np package. *J. Stat. Softw.* **27**.
[18] Henderson, D.J. and Parmeter, C.F. (2009). Imposing economic constraints in nonparametric regression: Survey, implementation, and extension. In *Nonparametric Econometric Methods*. *Adv. Econom.* **25** 433–469. Bingley: Emerald Group Publ, Ltd. MR3495788 https://doi.org/10.1108/S0731-9053(2009)0000025016
[19] Horowitz, J.L. (1998). *Semiparametric Methods in Econometrics*. *Lecture Notes in Statistics* **131**. New York: Springer. MR1624936 https://doi.org/10.1007/978-1-4612-0621-7
[20] Horowitz, J.L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. *Springer Series in Statistics*. New York: Springer. MR2535631 https://doi.org/10.1007/978-0-387-92870-8
[21] Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333 https://doi.org/10.1214/aos/1009210681
[22] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24** 540–568. MR1394975 https://doi.org/10.1214/aos/1032894452
[23] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981 https://doi.org/10.1016/0304-4076(93)90114-K
[24] Klaassen, C.A.J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15** 1548–1562. MR0913573 https://doi.org/10.1214/aos/1176350609
[25] Kuchibhotla, A.K. and Patra, R.K. (2016). simest: Single index model estimation with constraints on link function. R package version 0.6.
[26] Kuchibhotla, A.K. and Patra, R.K. (2020). Supplement to "Efficient estimation in single index models through smoothing splines." https://doi.org/10.3150/19-BEJ1183SUPP.
[27] Kuchibhotla, A.K., Patra, R.K. and Sen, B. (2017). Efficient estimation in convex single index models. ArXiv e-prints.
[28] Lepski, O. and Serdyukova, N. (2013). Adaptive estimation in the single-index model via oracle approach. *Math. Methods Statist.* **22** 310–332. MR3146598 https://doi.org/10.3103/S1066530713040030
[29] Lepski, O. and Serdyukova, N. (2014). Adaptive estimation under single-index constraint in a regression model. *Ann. Statist.* **42** 1–28. MR3161459 https://doi.org/10.1214/13-AOS1152
[30] Li, J., Li, Y. and Zhang, R. (2017). B spline variable selection for the single index models. *Statist. Papers* **58** 691–706. MR3686846 https://doi.org/10.1007/s00362-015-0721-z
[31] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

[32] Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136 https://doi.org/10.1214/aos/1176347254

[33] Li, Q. and Racine, J.S. (2007). *Nonparametric Econometrics*: *Theory and Practice*. Princeton, NJ: Princeton Univ. Press. MR2283034

[34] Li, W. and Patilea, V. (2017). A new minimum contrast approach for inference in single-index models. *J. Multivariate Anal.* **158** 47–59. MR3651372 https://doi.org/10.1016/j.jmva.2017.03.009

[35] Liu, J., Zhang, R., Zhao, W. and Lv, Y. (2013). A robust and efficient estimation method for single index models. *J. Multivariate Anal.* **122** 226–238. MR3189320 https://doi.org/10.1016/j.jmva.2013.08.007

[36] Ma, Y. and Zhu, L. (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 305–322. MR3021389 https://doi.org/10.1111/j.1467-9868.2012.01040.x

[37] Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014–1035. MR1447739 https://doi.org/10.1214/aos/1069362736

[38] Meegaskumbura, R. (2011). Control theoretic smoothing splines with derivative constraints. Ph.D. thesis.

[39] Meyer, C. (2000). *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: SIAM. MR1777382 https://doi.org/10.1137/1.9780898719512

[40] Meyer, M.C. (2008). Inference using shape-restricted regression splines. *Ann. Appl. Stat.* **2** 1013–1033. MR2516802 https://doi.org/10.1214/08-AOAS167

[41] Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168 https://doi.org/10.2307/2669386

[42] Murphy, S.A., van der Vaart, A.W. and Wellner, J.A. (1999). Current status regression. *Math. Methods Statist.* **8** 407–425. MR1735473

[43] Newey, W.K. and Stoker, T.M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica* **61** 1199–1223. MR1234794 https://doi.org/10.2307/2951498

[44] Oden, J.T. and Reddy, J.N. (2012). *An Introduction to the Mathematical Theory of Finite Elements*. New York: Wiley Interscience. MR0461950

[45] Park, H., Petkova, E., Tarpey, T. and Ogden, R.T. (2020). A single-index model with multiple-links. *J. Statist. Plann. Inference* **205** 115–128. MR4011626 https://doi.org/10.1016/j.jspi.2019.05.008

[46] Patra, R.K., Seijo, E. and Sen, B. (2018). A consistent bootstrap procedure for the maximum score estimator. *J. Econometrics* **205** 488–507. MR3813528 https://doi.org/10.1016/j.jeconom.2018.04.001

[47] Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *J. Statist. Plann. Inference* **141** 1362–1379. MR2747907 https://doi.org/10.1016/j.jspi.2010.10.003

[48] Pešta, M. and Hlávka, Z. (2017). Shape constrained regression in Sobolev spaces with application to option pricing. In *Analytical Methods in Statistics*. *Springer Proc. Math. Stat.* **193** 123–157. Cham: Springer. MR3639790

[49] Powell, J.L., Stock, J.H. and Stoker, T.M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117 https://doi.org/10.2307/1913713

[50] Racine, J.S., Parmeter, C.F. and Du, P. (2009). Constrained nonparametric kernel regression: Estimation and inference. Working paper.

[51] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge: Cambridge Univ. Press. MR1998720 https://doi.org/10.1017/CBO9780511755453

[52] Shen, J. and Lebair, T.M. (2015). Shape restricted smoothing splines via constrained optimal control and nonsmooth Newton's methods. *Automatica J. IFAC* **53** 216–224. MR3318591 https://doi.org/10.1016/j.automatica.2014.12.040

[53] Stoker, T.M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481. MR0868152 https://doi.org/10.2307/1914309
[54] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. New York: Springer. MR2233926
[55] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343 https://doi.org/10.1214/aos/1176347632
[56] van de Geer, S.A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge: Cambridge Univ. Press. MR1739079
[57] van der Vaart, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1999). *Lecture Notes in Math*. **1781** 331–457. Berlin: Springer. MR1915446
[58] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256
[59] Wahba, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. Philadelphia, PA: SIAM. MR1045442 https://doi.org/10.1137/1.9781611970128
[60] Wang, K. and Lin, L. (2014). New efficient estimation and variable selection in models with single-index structure. *Statist. Probab. Lett.* **89** 58–64. MR3191462 https://doi.org/10.1016/j.spl.2014.02.019
[61] Wang, L. and Cao, G. (2018). Efficient estimation for generalized partially linear single-index models. *Bernoulli* **24** 1101–1127. MR3706789 https://doi.org/10.3150/16-BEJ873
[62] Wang, L. and Yang, L. (2009). Spline estimation of single-index models. *Statist. Sinica* **19** 765–783. MR2514187
[63] Wu, T.Z., Yu, K. and Yu, Y. (2010). Single-index quantile regression. *J. Multivariate Anal.* **101** 1607–1621. MR2610735 https://doi.org/10.1016/j.jmva.2010.02.003
[64] Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137. MR2328530 https://doi.org/10.1017/S0266466606060531
[65] Xia, Y., Tong, H., Li, W.K. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. MR1924297 https://doi.org/10.1111/1467-9868.03411
[66] Yang, J., Tian, G., Lu, F. and Lu, X. (2020). Single-index modal regression via outer product gradients. *Comput. Statist. Data Anal.* **144** 106867, 14. MR4019835 https://doi.org/10.1016/j.csda.2019.106867
[67] Yatchew, A. and Bos, L. (1997). Nonparametric least squares regression and testing in economic models. *J. Quant. Econ.* **13** 81–131.
[68] Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97** 1042–1054. MR1951258 https://doi.org/10.1198/016214502388618861
[69] Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36** 1649–1668. MR2435451 https://doi.org/10.1214/07-AOS529
[70] Zhou, L., Lin, H., Chen, K. and Liang, H. (2019). Efficient estimation and computation of parameters and nonparametric functions in generalized semi/non-parametric regression models. *J. Econometrics* **213** 593–607. MR4023924 https://doi.org/10.1016/j.jeconom.2019.06.005
[71] Zhu, L.-P., Qian, L.-Y. and Lin, J.-G. (2011). Variable selection in a class of single-index models. *Ann. Inst. Statist. Math.* **63** 1277–1293. MR2830860 https://doi.org/10.1007/s10463-010-0287-4
[72] Zou, Q. and Zhu, Z. (2014). M-estimators for single-index model using B-spline. *Metrika* **77** 225–246. MR3157984 https://doi.org/10.1007/s00184-013-0434-z

Received May 2019 and revised October 2019