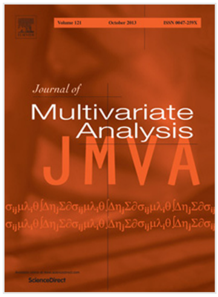
Accepted Manuscript



A new minimum contrast approach for inference in single-index models

Weiyu Li, Valentin Patilea

PII: S0047-259X(17)30198-7
DOI: http://dx.doi.org/10.1016/j.jmva.2017.03.009
Reference: YJMVA 4237

To appear in: *Journal of Multivariate Analysis*

Received date: 5 August 2016

A new minimum contrast approach for inference in single-index models

Weiyu Li[a,b,*], Valentin Patilea[a,**]

[a] CREST (Ensai), France
[b] Shandong University, China

Abstract

Semiparametric single-index models represent an appealing compromise between parametric and nonparametric approaches and have been widely investigated in the literature. The underlying assumption in single-index models is that the information carried by the vector of covariates could be summarized by a one-dimensional projection. We propose a new, general inference approach for such models, based on a quadratic form criterion involving kernel smoothing. The approach could be applied with general single-index assumptions, in particular for mean regression models and conditional law models. The covariates could be unbounded and no trimming is necessary. A resampling method for building confidence intervals for the index parameter is proposed. Our empirical experiments reveal that the new method performs well in practice.

Keywords: Conditional law, Kernel smoothing, Semiparametric regression, Single-index assumption, U-statistics

2010 MSC: 62G08, 62G20; 62J02, 62F12

1. Introduction

Modeling the relationship between one or several response variables and a vector of covariates is a common problem in statistics. Usually, one aims

*Corresponding author

**Valentin Patilea gratefully acknowledges support from the research program New Challenges for New Data of Genes and LCL.

Email addresses: liweiyu84@gmail.com (Weiyu Li), patilea@ensai.fr (Valentin Patilea)

at modeling the conditional law of responses given the covariates, or at least some characteristics of this conditional law, such as the mean, the median, the higher order moments, etc. In a parametric approach, one specifies a set indexed by a vector of parameters, i.e., a model to which the conditional law, or its characteristic of interest, is supposed to belong. Linear regression is the most prominent example. In a fully nonparametric approach, the model is specified as broadly as possible, but of course there is a price to pay for the model's complexity which is reflected in the poor accuracy of the resulting estimators. Therefore, one often looks for semiparametric approaches that strike a compromise between the accuracy that could be obtained via a parametric model and the flexibility of nonparametric specifications.

Single-index models are common semiparametric approaches that achieve such a compromise. The underlying assumption is that the information on the quantity of interest (the law of responses or some characteristics of it) carried by the covariate vector, is the same as the information carried by a one-dimensional projection of the covariate vector, the so-called index. In other words, one still considers a nonparametric approach, but only after a dimension reduction step which replaces the original covariate vector by some linear combination of its components. See, e.g., [3, 11–14, 16–18, 20, 21, 23, 27] and references therein.

Despite the extensive literature on single-index models, some technical aspects remain unsatisfactorily resolved. For instance, in most contributions, the covariates are supposed to have a bounded support. Even with bounded support covariates, trimming is usually employed to keep density estimates (usually appearing in the denominators) away from zero. In some papers, it is supposed that a pilot estimator, with suitable rate, is available. Moreover, in many contributions considering the additive regression setup, the error term is assumed to be homoscedastic.

In this paper we introduce a new general inference approach for the index in conditional models using a single-index assumption. Our approach is based on kernel smoothing and could be applied to any existing framework, under mild conditions. It does not require any of the technical conditions mentioned above.

We allow discrete and continuous covariates to be unbounded, heteroscedastic error terms to appear in the mean regression setup, and no trimming to be involved in the inference. The approach follows and extends, from a parametric to a semiparametric framework, the idea of the Smooth Minimum Distance estimation method of [19].

The paper is organized as follows. The underlying idea of the new approach is presented in Section 2. The corresponding estimators are introduced in Section 3, where their consistency and asymptotic normality are also derived. In Section 4 we propose a simple procedure for constructing confidence intervals for the index coefficients. Some empirical evidence on the performance of our inference method is provided in Section 5, using both simulated and real data examples. The simulation results indicate that our method performs well compared to existing approaches. Some technical aspects are postponed to the Appendix. Complementary proofs are provided in Supplementary Material.

2. The framework

Assume that the observations are independent copies of $(Y^\top, X^\top)^\top$ where $Y \in \mathbb{R}^d$, $d \geq 1$, denotes the random response vector and $X \in \mathbb{R}^p$, $p \geq 1$, stands for the random column vector of covariates. For mean regression, the single-index assumption means that exists a column parameter vector $\beta_0 \in \mathbb{R}^p$ such that

$$\mathrm{E}(Y \mid X) = \mathrm{E}(Y \mid X^\top \beta_0). \tag{1}$$

The scalar product $X^\top \beta_0$ is the so-called index. The direction β_0 and the nonparametric univariate (i.e., one predictor) regression $\mathrm{E}(Y \mid X^\top \beta_0)$ have to be estimated. See [6, 8, 12, 13] and references therein for a panorama of the existing estimation procedures. When applying the single-index paradigm to conditional laws of Y given X, one assumes

$$Y \perp X \mid X^\top \beta_0. \tag{2}$$

In this case, the direction defined by β_0 and the conditional law of the response Y given the index $X^\top \beta_0$, have to be estimated. See [5, 7, 10, 29] for various estimation approaches. In both situations only the direction given by β_0 is identified, so that a suitable identification condition should accompany the model assumption.

In order to formulate the problem in a general, unified way, consider $\{T_u : u \in \mathcal{U}\}$, a family of transformations of the response variable Y. The transformations T_u take values in some finite-dimensional space that could be, for instance, the space of Y or the real line. The set $\mathcal{U}$ is contained in some finite-dimensional space. Then, the general single-index model (SIM) assumption we consider is the following:

$$\exists!_{\beta_0} \, \forall_{u\in\mathcal{U}} \quad \mathrm{E}\{T_u(Y)|X\} = \mathrm{E}\{T_u(Y) \mid X^\top\beta_0\}, \tag{3}$$

where the unique β_0 is an unknown index vector which belongs to the parameter set

$$\mathcal{B} \subset \{(\beta_1, \ldots, \beta_p) : \beta_1 = 1\} \subset \mathbb{R}^p. \tag{4}$$

In particular, this framework makes it possible to take into account the two single-index assumptions presented above. Indeed, if the family of transformation contains only the identity transformation, i.e., $T_u(y) = y$ for any $u \in \mathcal{U}$, one recovers Condition (1). In contrast if $T_u(y) = \mathbf{1}(y \leq u)$ for all $u \in \mathcal{U} = \mathbb{R}^p$, then (3) becomes equivalent to Condition (2). (Here and in the following, for any v_1 and v_2 vectors of the same dimension, $v_1 \leq v_2$ stands for the componentwise inequality between v_1 and v_2.) For simplicity, hereafter we only consider Condition (3) for one of these two types of transformations T_u.

Let us assume that for any $\beta \in \mathcal{B}$, the random variable $X^\top\beta$ has a density denoted by f_β. To estimate a parameter β_0 that satisfies Condition (3), first let us define, for all $z \in \mathbb{R}$, $\beta \in \mathcal{B}$, $u \in \mathcal{U}$,

$$g_u(Y, z; \beta) = [T_u(Y) - \mathrm{E}\{T_u(Y) \mid X^\top\beta = z\}] f_\beta(z).$$

Then, Condition (3) is equivalent to the following:

$$\forall_{u\in\mathcal{U}} \;\; \mathrm{E}\{g_u(Y, X^\top\beta; \beta) \mid X\} = 0 \text{ almost surely} \;\Leftrightarrow\; \beta = \beta_0. \tag{5}$$

Next, the idea is to build a contrast function that allows the conditional moment conditions to be encompassed in one marginal (unconditional) equation. For this purpose, let $(Y_1^\top, X_1^\top)^\top$ and $(Y_2^\top, X_2^\top)^\top$ be two independent copies of $(Y^\top, X^\top)^\top$ and let ω be a real-valued integrable function defined over the space of X. Assume that ω has an integrable, strictly positive Fourier transform. For instance, one could take $\omega(x) = \exp(-\|x\|^2/2)$, $x \in \mathbb{R}^p$. As mentioned in [19], other examples are products $\omega(x) = \omega(x_1, \ldots, x_p) = \widetilde{\omega}(x_1)\cdots\widetilde{\omega}(x_p)$ with $\widetilde{\omega}$ a triangular, logistic, Student (including Cauchy), or Laplace density.

Finally, define the real-valued contrast function, for all $\beta \in \mathcal{B}$, by

$$Q(\beta) = \int_{\mathcal{U}} \mathrm{E}\{g_u(Y_1, X_1^\top\beta; \beta)^\top g_u(Y_2, X_2^\top\beta; \beta)\omega(X_1 - X_2)\}d\mu(u), \qquad (6)$$

where μ is some probability measure with support $\mathcal{U}$ considered with the Borel $\sigma-$field. As will be mentioned in the following, for the case corresponding to Assumption (2), a convenient choice is $\mu = F_Y$, where F_Y denotes the probability distribution of Y. In applications, F_Y is unknown and could be replaced by a given approximation or by the empirical distribution of the sample of Y.

The following result guarantees that the direction β_0 from Condition (5) could be identified as the unique root of the contrast Q.

Lemma 1. *Let $\mathcal{B}$ be some parameter set defined as in Eq.* (4)*. Assume that the Fourier transform of ω is strictly positive and integrable. Then $Q(\beta) \geq 0$ for all $\beta \in \mathcal{B}$. Moreover, Condition* (5) *holds true if and only if $Q(\beta_0) = 0$ and $Q(\beta) > 0$ for all $\beta \neq \beta_0$.*

The idea of our estimation approach is to build a sample-based approximation of $Q(\beta)$ and to minimize it with respect to the parameter β. Let us point out that, by the definition of the functions g_u, the covariates will be allowed to have unbounded support and no trimming will be necessary in the approximation of $Q(\beta)$.

Let us point out that, in general, one could not simply use a least-squares type contrast instead of $Q(\beta)$. For illustration, let us consider the case of a single-index assumption for the mean regression of a real-valued response, i.e.,

$Y = \mathrm{E}(Y \mid X^\top\beta_0) + \varepsilon$ and $\mathrm{E}(\varepsilon \mid X) = 0$. Then one can decompose

$$\mathrm{E}\{g_u^2(Y, X^\top\beta;\beta)\} = \mathrm{E}\left[\left\{\mathrm{E}(Y \mid X^\top\beta_0) - \mathrm{E}(Y \mid X^\top\beta)\right\}^2 f_\beta^2(X^\top\beta)\right] + \mathrm{E}\{\varepsilon^2 f_\beta^2(X^\top\beta)\}],$$

and thus it becomes clear that β_0 cannot be the minimum of $\mathrm{E}\{g_u^2(Y, X^\top\beta;\beta)\}$. Our contrast $Q(\beta)$, inspired by the Smooth Minimum Distance estimation method introduced in [19], avoids the identification problem for β_0, provided that Condition (5) holds true.

Finally, let us point out that the definition of the criterion $Q(\beta)$, and hence the estimation approach that will be described in the following, could be extended to the case of a multiple index assumption. It suffices to replace the index $X^\top\beta$ by a multiple index $X^\top B$, where B is a $p\times q-$matrix, $1 \leq q < p$, and to reconsider the construction above. For simplicity, we focus herein on single-index assumptions.

3. The estimation method

Let $(Y_1^\top, X_1^\top)^\top, \ldots, (Y_n^\top, X_n^\top)^\top$ be a random sample from $(Y^\top, X^\top)^\top$. Our estimator of β_0 is

$$\widehat\beta = \arg\min_{\beta\in\mathcal{B}} \widehat Q_n(\beta),$$

where, for each $\beta \in \mathcal{B}$,

$$\widehat Q_n(\beta) = \int_{\mathcal{U}} \left\{ \frac{1}{n^2}\sum_{i,j=1}^n \widehat g_u(Y_i, X_i^\top\beta;\beta)^\top \widehat g_u(Y_j, X_j^\top\beta;\beta)\omega_{ij} \right\} d\mu_n(u), \tag{7}$$

$\widehat g_u$ is an estimate of g_u, $\omega_{ij} = \omega(X_i - X_j)$ and μ_n is some probability measure that may depend on the sample. For a simpler presentation, in the theoretical results we will assume that μ_n is the empirical distribution of the response Y and μ is equal to F_Y, the marginal distribution of Y. For estimating g_u we use kernel smoothing and define

$$\begin{aligned}\widehat{g}_u(y,z;\beta) &= T_u(y)\widehat{f}_\beta(z) - \mathrm{E}\{\widehat{T_u(Y) \mid X^\top}\beta = z\}\widehat{f}_\beta(z)\\ &= \frac{1}{nh}\sum_{k=1}\{T_u(y) - T_u(Y_k)\}\,K\{(X_k^\top\beta - z)/h\},\end{aligned}$$

where K is a univariate kernel and h is the bandwidth. The choice of μ_n matters only in some cases, as for instance, in the case of the single-index in the distributional assumption (2). More precisely, under Assumption (2) we propose

$$\widehat{Q}_n(\beta) = \frac{1}{n}\sum_{l=1}^{n}\frac{1}{n^2}\sum_{i,j=1}^{n}\widehat{g}_{Y_\ell}(Y_i, X_i^\top\beta;\beta)\widehat{g}_{Y_\ell}(Y_j, X_j^\top\beta;\beta)\omega_{ij},$$

where

$$\widehat{g}_{Y_\ell}(Y_i, X_i^\top\beta;\beta) = \frac{1}{nh}\sum_{k=1}^{n}\{\mathbf{1}(Y_i \le Y_\ell) - \mathbf{1}(Y_k \le Y_\ell)\}\,K\{(X_k - X_i)^\top\beta/h\}.$$

In the case of Assumption (1), we propose the criterion

$$\widehat{Q}_n(\beta) = \frac{1}{n^2}\sum_{i,j=1}^{n}\widehat{g}(Y_i, X_i^\top\beta;\beta)^\top\widehat{g}(Y_j, X_j^\top\beta;\beta)\omega_{ij},$$

where

$$\widehat{g}(Y_i, X_i^\top\beta;\beta) = \frac{1}{nh}\sum_{k=1}^{n}(Y_i - Y_k)\,K\{(X_k - X_i)^\top\beta/h\}.$$

Let us comment on a common feature of the single-index estimation methods. By the nature of the model, a nonparametric estimation is involved in any semiparametric single-index estimation approach. In general, this requires controlling of small values of the nonparametric density estimators appearing in the denominators. A common practice is to suppose that the density of $X^\top\beta$ is uniformly bounded away from zero, for all $\beta \in \mathcal{B}$. Such a condition is quite unrealistic, even when X has a bounded support and a density bounded away from zero. Indeed, one may easily build a counterexample considering a bidimensional $X = (X_{(1)}, X_{(2)})^\top$ with two independent uniform random variables on $[0,1]$ and $\mathcal{B} = \{(1,\beta_2)^\top : |\beta_2| \le b\}$, for some arbitrary $b > 0$. Then, except for $\beta_2 = 0$, the random variable $X_{(1)} + \beta_2 X_{(2)}$ does not have a density bounded away from zero. The usual remedy is to trim the criterion used for estimation, i.e., to remove the observations leading to small estimated values for the density of $X^\top \beta$. The trimming may be relaxed with the sample size, i.e., the fraction of removed observations could grow more slowly than the sample size, but one still has to use complicated arguments for the asymptotics. For both situations we consider here, single-index in mean and single-index in law, the new approach we propose allows covariates to be unbounded and does not require trimming. To the best of our knowledge, our estimation method is the first to have this feature.

Let ∇_β the differential operator given by the last $p-1$ first order partial derivatives corresponding to the last $p-1$ components of β. In the case of Condition (2), let $\Sigma(\boldsymbol{\beta}_0) = 4\mathrm{E}\{\psi(Y, X; \beta_0)\psi(Y, X; \beta_0)^\top\}$,

$$J(\boldsymbol{\beta}_0) = \int_{\mathcal{U}} \mathrm{E}[\mathrm{E}\{\nabla_\beta g_u(Y_1, X_1^\top\beta_0;\beta_0) \,|\, X_1\} \\ \times \mathrm{E}\{\nabla_\beta g_u(Y_2, X_2^\top\beta_0;\beta_0) \,|\, X_2\}^\top \omega(X_1 - X_2)] d\mu(u),$$

and

$$\psi(Y_1, X_1; \beta_0) = \int_{\mathcal{U}} \mathrm{E}[\mathrm{E}\{\nabla_\beta g_u(Y, X^\top\beta_0; \beta_0) \,|\, X\} \\ \times \omega(X - X_1) \,|\, X_1] g_u(Y_1, X_1^\top\beta_0; \beta_0) d\mu(u).$$

In the case of a single-index mean regression, $g_u(y, t; \beta)$ does not depend on u, hence the integrals with respect to μ disappear from the definitions of the $(p-1)\times(p-1)-$matrices $J(\beta_0)$ and $\psi(Y, X; \beta_0)$ above. The following result describes the asymptotic behavior of the semiparametric estimator $\widehat{\beta}$. Below, $\rightsquigarrow$ denotes convergence in law and $\mathbf{0}_{p-1}$ is the null column vector in $\mathbb{R}^{p-1}$.

Proposition 2. *Let $\widehat{\beta} = \arg\min_{\beta\in\mathcal{B}} \widehat{Q}_n(\beta)$ for $\widehat{Q}_n(\beta)$ defined as in Eq.* (7) *with μ_n the empirical distribution of $Y_1, \ldots, Y_n$. Suppose that the identification Condition* (5) *holds true. Under Assumption 1, $\widehat{\beta} \to \beta_0$ in probability as $n \to \infty$. If in addition Assumption 2 holds true, then $\sqrt{n}\,(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}_p(0, V)$ as $n \to \infty$,* *where*

$$V = \begin{pmatrix} 0 & \mathbf{0}'_{p-1} \\ \mathbf{0}_{p-1} & V_{p-1} \end{pmatrix} \qquad \textit{with} \qquad V_{p-1} = J(\boldsymbol{\beta}_0)^{-1}\Sigma(\boldsymbol{\beta}_0)J(\boldsymbol{\beta}_0)^{-1},$$

Let us comment on Proposition 2. As the function ω appears in the asymptotic variance, its choice has some influence on the performance of $\widehat{\beta}$. In [19], the authors considered the case of a family of functions ω indexed by a bandwidth parameter allowed to decrease to zero. In that case, ω no longer appears in the asymptotic variance of the parametric estimator studied in [19]. However, the authors report empirical evidence of the influence of that bandwidth parameter on the mean squared error of their estimator. A detailed investigation of the influence of the choice of ω is beyond the scope of our contribution.

The first row and the first column of the matrix V in Proposition 2, are identically equal to zero. This comes from our identification condition and the parameter space defined in Eq. (4). In fact, only the subvector built with the last $p-1$ components of β_0 is estimated and the asymptotic variance of the estimator is V_{p-1}.

Given that $\widehat{\beta}$ is $\sqrt{n}$-consistent, one could derive the $\sqrt{nh}$-consistency for the conditional mean or the conditional distribution function of Y given X. These type of results are quite standard and straightforward, see, e.g., Section 2.4 in [13], and hence will be omitted.

4. Confidence intervals

The asymptotic variance of $\widehat{\beta}$ has a complicated form. To approximate the law of $\widehat{\beta}$ in small and moderate samples, we propose a resampling-based approach similar to the one used in [19]; see also [15]. The idea is to build a suitable randomly perturbed version of the criterion $\widehat{Q}_n(\beta)$ and to compute its minimum. Conditionally on the original sample, the law of this minimum is shown to be close to the law of $\widehat{\beta}$. More precisely, the steps of the procedure are as follows.

1. Generate a random sample $\xi_1, \dots, \xi_n$ from a distribution with unit mean, unit variance and finite fourth-order moment.
2. Build the randomly perturbed criterion
$$\widehat{Q}_n^*(\beta) = \int_{\mathcal{U}} \left\{ \frac{1}{n^2} \sum_{i,j=1}^{n} \widehat{g}_u(Y_i, X_i^\top \beta; \beta)^\top \widehat{g}_u(Y_j, X_j^\top \beta; \beta) \omega_{ij}^* \right\} d\mu_n(u),$$
where μ_n is the empirical distribution of the responses and $\omega_{ij}^* = \xi_i \xi_j \omega_{ij}$.
3. Define
$$\widehat{\beta}^* = \arg\min_{\beta \in \mathcal{B}} \widehat{Q}_n^*(\beta).$$

The following result establishes the asymptotic validity of this procedure. The arguments for the proof could be obtained by standard modifications of those for the proof of Proposition 2, and hence will be omitted.

Proposition 3. *Under the conditions of Proposition 2 guaranteeing the asymptotic normality of $\sqrt{n}\,(\widehat{\beta} - \beta_0)$, for any $w \in \{0\} \times \mathbb{R}^{p-1}$,*
$$\Pr\{\sqrt{n}\,(\widehat{\beta}^* - \widehat{\beta}) \leq w \mid Y_1, X_1, \dots, Y_n, X_n\} - \Pr\{\sqrt{n}\,(\widehat{\beta} - \beta_0) \leq w\} \to 0,$$
in probability as $n \to \infty$.

In practice, the conditional distribution of $\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$, given the original data set, can be estimated by repeating B times the steps 1 to 3 above, where B is some large number. For the jth independent draw of $\xi_1, \dots, \xi_n$, let $\widehat{\beta}_j^*$ denote the value obtained by minimizing $\widehat{Q}_n^*(\beta)$, $j = 1, \dots, B$. The theoretical conditional distribution of $\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$ can then be approximated by the usual empirical distribution function based on $\widehat{\beta}_1^*, \dots, \widehat{\beta}_B^*$. The conditions on the law of the ξ_i's are very mild, but the choice of this law may have some influence with small and moderate samples. Guided by our simulation experience, we suggest using the unit exponential distribution to generate the ξ_i's.

5. Empirical illustrations

We investigated the performance of our new approach to build parameter estimates and confidence intervals for single-index models through extensive simulation experiments and real data examples. The general conclusion is that our approach performs well, and sometimes much better, compared to existing approaches. In all our empirical studies we used a Gaussian kernel K.

5.1. Simulation experiments with single-index in mean models

First, we consider two setups similar to those considered in [6]: the model equation is

$$Y = (X^\top \beta_0)^2 + \varepsilon, \tag{8}$$

with a three-dimensional vector of covariates $X = (X_{(1)}, X_{(2)}, X_{(3)})^\top$, where the independent sample of $(X_{(1)}, X_{(2)})^\top$ is generated from a bivariate normal law with mean 1, unit standard deviations and correlation equal to 0.2. As for $X_{(3)}$, it is a Bernoulli random variable with parameter $p = 0.4$. The true parameter is $\beta = (\beta_{0,1}, \beta_{0,2}, \beta_{0,3})^\top = (1, 0.8, 0.5)^\top$. The first setup is a homoscedastic case where the error ε has a $\mathcal{N}(0, 0.5^2)$ law. In this case the signal-to-noise ratio, SSR/SSE, is approximately equal to 76.6. In the second setup we introduce some heteroscedasticity by considering $\varepsilon \sim \mathcal{N}[0, (X^\top\beta)^4/25]$. The value of the signal-to-noise ratio is then approximately equal to 13.

Our estimator $\widehat{\beta}$ depends on the bandwidth h. Here we select h from an equidistant grid $\{0.03, 0.06, \ldots, 0.30\}$ such that the loss $\widehat{Q}(\widehat{\beta})$ is minimal. The simulation results, based on 500 replicates with a sample of size $n = 50$, are shown in Table 1. We report the elementary descriptive statistics, mean, median and standard deviation, and the absolute estimation error (AEE) defined as $|\beta_{0,2} - \widehat{\beta}_2| + |\beta_{0,3} - \widehat{\beta}_3|$. We also include the results obtained from the EFM approach proposed by [6], adjusted by a final Fisher scoring step, as could be found in the codes kindly provided by the authors. Moreover we report the benchmark results obtained by the nonlinear least squares method (NLS) in the homoscedastic case and by the weighted nonlinear least squares method (WNLS) in the heteroscedastic case. With these parametric estimation approaches, the conditional mean and the conditional variance are known up to the parameter β_0. The results show that our method performs well compared to EFM. Its performance is not quite as satisfactory in the homoscedastic case, but remains

Table 1: Single-index in mean. Simulation results for the estimators of β_0 obtained from 500 replicates generated using Model (8).

		Homoscedastic case, $n = 50$				Heteroscedastic case, $n = 50$		
		NLS	Ours	EFM		WNLS	Ours	EFM
$\beta_{0,2} = 0.8$	Median	0.7992	0.8025	0.7994	Median	0.7987	0.8018	0.7965
	Mean	0.7991	0.8030	0.7993	Mean	0.7972	0.8096	0.8070
	StD	0.0163	0.0607	0.0376	StD	0.0168	0.0935	0.1244
	MSE	0.0002	0.0014	0.0036	MSE	0.0003	0.0088	0.0155
$\beta_{0,3} = 0.5$	Median	0.5001	0.4981	0.5000	Median	0.4987	0.4996	0.4995
	Mean	0.4996	0.5025	0.4997	Mean	0.4982	0.5057	0.5072
	StD	0.0164	0.0477	0.0390	StD	0.0117	0.0712	0.0988
	MSE	0.0002	0.0022	0.0008	MSE	0.0001	0.0050	0.0098
AEE[1]		0.0252	0.0836	0.0532	AEE[1]	0.0204	0.1251	0.1664

[1] AEE stands for the absolute estimation error.

marginally preferable to EFM in the heteroscedastic case. As expected, the parametric approaches are more accurate.

Next, we consider a third setup inspired by the empirical study presented in [21]. The law of the six covariate vector $X = (X_{(1)}, \dots, X_{(6)})^\top$ is constructed as follows:

1) $X_{(1)}, X_{(2)}, e_1$ and e_2 are independent, standard normal random variables.
2) $X_{(3)} = 0.2X_{(1)} + 0.2(X_{(2)} + 2)^2 + 0.2e_1$ and $X_{(4)} = 0.1 + 0.1(X_{(1)} + X_{(2)}) + 0.3(X_{(1)} + 1.5)^2 + 0.2e_2$.
3) Given $X_{(1)}$ and $X_{(2)}$, generate $X_{(5)}$ and $X_{(6)}$ independently as Bernoulli variables with respective success probabilities $\exp(X_1)/\{1+\exp(X_1)\}$ and $\exp(X_2)/\{1 + \exp(X_2)\}$.

Let $\beta_0 = (1.3, -1.3, 1, -0.5, 0.5, -0.5)^\top/1.3$. The response Y is obtained as

$$Y = \sin(2X^\top\beta_0) + 2\exp(X^\top\beta_0) + \varepsilon, \tag{9}$$

Table 2: Single-index in mean. Simulation results for the estimators of β_0 obtained from 500 replicates from the model (9).

		Mean	StD	Median	MSE
$\beta_2 = -1$	WNLS	-1	0.004	-1	1.6092×10^{-5}
	Ours	-1.012	0.038	-1.012	0.0016
	EFM	-3.955	3.723	-4.205	22.5622
$\beta_3 \approx 0.769$	WNLS	0.769	0.006	0.769	3.4043×10^{-5}
	Ours	0.777	0.033	0.776	0.0011
	EFM	3.181	3.003	3.196	0.148091
$\beta_4 \approx -0.385$	WNLS	-0.384	0.003	-0.385	8.0883×10^{-6}
	Ours	-0.380	0.012	-0.380	0.0001
	EFM	-1.223	2.249	-0.661	5.7518
$\beta_5 \approx 0.385$	WNLS	0.385	0.007	0.385	5.0449×10^{-5}
	Ours	0.390	0.017	0.390	0.0003
	EFM	1.193	1.295	1.373	2.3268
$\beta_6 \approx -0.385$	WNLS	-0.385	0.007	-0.384	5.0710×10^{-5}
	Ours	-0.388	0.015	-0.387	0.0002
	EFM	-1.355	1.233	-1.455	2.4593
AEE[1]	WNLS: 0.0213		Ours: 0.0923		EFM: 10.0191

[1] AEE stands for the absolute estimation error.

210 where $\varepsilon \sim \mathcal{N}[0, \ln\{2 + (X^\top \beta_0)^2\}]$. Again, we compare our method with EFM and WNLS. The results presented in Table 2 are obtained from 500 replicates with samples of size $n = 100$. Again, the bandwidth h is chosen by minimizing the loss $\widehat{Q}_n(\widehat{\beta})$ over the grid $\{0.05, 0.1, 0.15\}$. The EFM approach produces very poor results, while our method provides accurate estimates, with performances close to that of the WNLS estimates. The very good accuracy of the WNLS estimators could be explained by the construction of the setup, which yields a value of the signal-to-noise ratio close to 2700.

5.2. Simulation experiments with single-index in law models

Three setups with responses having a single-index conditional law are considered. First,

$$Y = X^\top \beta_0 + \varepsilon, \tag{10}$$

with X a trivariate normal random vector with mean zeros, standard deviations equal to 1 and pairwise correlations equal to 0.2, and a Cauchy distribution error term. Next, following [21], we consider

$$Y = \sin(2X^\top \beta_0) + 2\exp(X^\top \beta_0) + \varepsilon \tag{11}$$

and

$$Y = \sin(2X^\top \beta_0) + 2\exp(X^\top \beta_0) + \sqrt{\ln(2 + X^\top \beta_0)}\, \varepsilon, \tag{12}$$

where the error ε has a normal distribution and the vector of covariates $X = (X_{(1)}, X_{(2)}, X_{(3)})^\top$ is generated as follows:

1) $X_{(1)}$ and e_1 are independent standard normal random variables.
2) $X_{(2)} = 0.3 + 0.2X_{(1)} + 0.1(X_{(1)} + 1.5)^2 - 0.3e_1^2$.
3) given $X_{(1)}$, $X_{(3)}$ is a Bernoulli variable with probability $\exp(X_{(1)})/\{1 + \exp(X_{(1)})\}$.

In Examples (10)–(12), the real parameter value is $\beta_0 = (1, 0.8, -0.5)^\top$. The simulation results are based on 200 replicates with samples of size $n = 50$ being reported in Table 3. Our method is compared with the maximum likelihood estimation (MLE), the method (PLISE) from [5] and the method proposed in [21] denoted as Eff. The bandwidth h is selected as the minimum of the loss $\widehat{Q}(\widehat{\beta})$ on the grid $\{0.01, \ldots, 0.05\}$. In the conditional Cauchy responses cases, our method performs much better than PLISE and slightly better than Eff. The poor behavior of the MLE is probably related to the multiple local maxima of a Cauchy likelihood, a well known problem in classical statistics; see, e.g., [24]. In the conditional Gaussian examples, our method seems to outperform the semiparametric competitors with respect to almost all the indicators we provide (mean, median, standard deviation and absolute estimation error).

Table 3: Single-index in conditional law. Simulation results from 500 replicates with Models (10)–(12).

Model (10), conditional Cauchy response, $n = 50$					
		MLE	PLISE	Eff	Ours
$\beta_2 = 0.8$	Median	0.7924	0.8245	0.7956	0.8020
	Mean	0.9063	0.9647	0.7971	0.8082
	StD	0.5872	0.7635	0.1268	0.1113
	MSE	0.3544	0.6072	0.0160	0.0123
$\beta_3 = -0.5$	Median	-0.5169	-0.5351	-0.5426	-0.5405
	Mean	-0.5621	-0.5801	-0.5414	-0.5452
	StD	0.3459	0.3883	0.0807	0.0879
	MSE	0.1229	0.1564	0.0082	0.0097
AEE[1]		0.6119	0.7295	0.1700	0.1585
Model (11), conditional Gaussian, homoscedastic response, $n = 50$					
		MLE	PLISE	Eff	Ours
$\beta_2 = 0.8$	Median	0.8020	0.8156	0.8116	0.7977
	Mean	0.8002	0.8193	0.8161	0.7989
	StD	0.0231	0.1634	0.1598	0.1158
	MSE	0.0005	0.0269	0.0256	0.0133
$\beta_3 = -0.5$	Median	-0.5003	-0.5015	-0.5116	-0.5105
	Mean	-0.4998	-0.4998	-0.4998	-0.5166
	StD	0.0094	0.0713	0.0982	0.0720
	MSE	0.00008	0.0050	0.0096	0.0054
AEE[1]		0.0243	0.1797	0.1944	0.1411
Model (12), conditional Gaussian, heteroscedastic response, $n = 50$					
		MLE	PLISE	Eff	Ours
$\beta_2 = 0.8$	Median	0.8010	0.8131	0.8197	0.7922
	Mean	0.8007	0.8203	0.8278	0.7887
	StD	0.0250	0.1752	0.2171	0.1072
	MSE	0.0006	0.0309	0.0477	0.0115
$\beta_3 = -0.5$	Median	-0.5102	-0.5023	-0.4974	-0.5004
	Mean	-0.5000	-0.5022	-0.4952	-0.5114
	StD	0.0109	0.0734	0.1022	0.0676
	MSE	0.0001	0.0053	0.0104	0.0046
AEE[1]		0.0272	0.1894	0.2229	0.1373

[1] AEE stands for the absolute estimation error.

Table 4: Empirical level and empirical length for the componentwise resampling-based confidence intervals in Models (8)–(10): sample size $n = 50$ and 199 samples $\xi_1, \ldots, \xi_n$ generated.

	Model (8) with $\mathcal{N}(0, 0.25)$ errors			
	90% resampling CI		95% resampling CI	
	Length	Level	Length	Level
$\beta_2 = 0.8$	0.1592	92	0.2054	96.5
$\beta_3 = 0.5$	0.1785	89	0.2301	96
	Model (10) with Cauchy errors			
	90% resampling CI		95% resampling CI	
	Length	Level	Length	level
$\beta_2 = 0.8$	0.2822	89.5	0.3713	96
$\beta_3 = -0.5$	0.1923	90	0.2538	96

5.3. *Confidence intervals*

Next, we use the idea described in Section 4 to build confidence intervals for the components of β in Models (8) and (10). We consider 200 samples of size $n = 50$ and for each sample we generated 199 independent random samples $\xi_1, \ldots, \xi_n$ from a unit exponential distribution and computed the criteria $\widehat{Q}^*(\beta)$. The 90% and 95% confidence intervals obtained with the optimal values $\widehat{\beta}^*$, are presented in Table 4. The level is quite accurate and the intervals have reasonable length, indicating that our simulation-based procedure for building confidence intervals is quite effective.

5.4. *Real data applications*

The investigation of the finite-sample performance of our semiparametric approach is completed by two applications using real data.

The first example is the New York air quality data set considered in [4]. It contains the measurements of daily ozone concentration (*ozone*), wind speed (*wind*), daily maximum temperature (*temp*), and solar radiation level (*solar*) on

Table 5: The estimator $\widehat{\beta}$ and the componentwise resampling-based confidence intervals (RCI) (levels 0.9 and 0.95) for New York air quality data: single-index mean regression model.

Variable	Coefficient estimate	0.9 RCI	0.95 RCI
$temp$	-6.0144	$(-6.3278, -5.7520)$	$(-6.4813, -5.6591)$
$wind^2$	-3.1942	$(-3.2430, -2.9008)$	$(-3.2914, -2.8394)$
$solar^2$	-1.3832	$(-1.5161, -1.1607)$	$(-1.6557, -1.0798)$
$wind * temp$	-0.5791	$(-0.7472, -0.1098)$	$(-0.8292, -0.0554)$
$temp * solar$	1.5339	(1.3256, 1.7995)	(1.2714, 1.8581)

111 successive days from May to September 1973 in the New York metropolitan area. The response variable is *ozone*, with empirical mean 42.0991 and empirical variance 1107.29. [1,17,28] considered a single-index mean regression model for this data set, while [5] fitted a single-index in law model. Here we consider the covariate vector X whose components are the variables $wind$, $temp$, $wind^2$, $solar^2$, $wind * temp$, and $temp * solar$. We then consider the single-index in mean assumption. The coefficient of $wind$ is set equal to 1. The single-index assumption was checked using the test proposed by [22] with bootstrap critical values and the p-value was 0.403. The estimate of the direction β and the componentwise confidence intervals are given in Table 5. The plot of the estimated link function is provided in Figure 1. The mean absolute deviation is

$$\frac{1}{111}\sum_{i=1}^{111}\left|ozone_i - \widehat{\mathrm{E}}(ozone_i|X_i^\top\widehat{\beta})\right| = 17.9248.$$

To estimate the parameter β, we select the bandwidth by minimization of the loss $Q(\widehat{\beta})$ on a grid $\{0.01, \ldots, 0.09\}$. Given the estimate $\widehat{\beta}$, we build the adjusted values by univariate smoothing of the response given $X_i^\top\widehat{\beta}$ with a bandwidth selected by least-squares cross-validation.

The second real data example illustrates the single-index in law model. We consider data on the employees' salaries in the Fifth National Bank of Springfield; see [2]. There are 208 observations in the data set and every observation

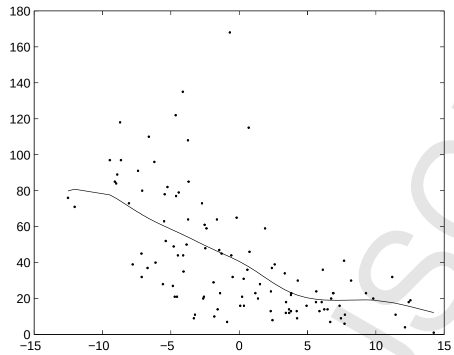


Figure 1: The estimated link function for New York air quality data set.

contains eight variables: *Education* (a categorical variable with five education levels), *Grade* (a categorical variable with six job levels), *Year1* (years of work experience at Fifth National), *Age* (employee's current age), *Year2* (years of work experience at another bank prior to working at Fifth National), *Gender* ('female'=1, 'male'=0), *PC Job* (a categorical variable depending on whether the job is computer related, 'yes'=1, 'no'=0), *Salary* (annual salary, the response variable).

As in [9], we delete the observations with *Age* over 60 or working experience $Year1 + Year2$ over 30 and this results in a subsample of 199 observations. Following [21], we also drop the variable *Education*, set the coefficient of *Grade* equal to 1 and let *Grade* take values from 1 to 6.

The single-index assumption for the conditional law was checked using the test proposed by [22] with asymptotic critical values and the p-value was 0.166. The estimator $\widehat{\beta}$ obtained by our approach is reported in Table 6, together with the resampling-based confidence intervals.

In contrast with the results reported by [21], we found a significant negative coefficient for *Gender*. This could be explained by the negative correlation between *Gender* and *Grade*. For instance, there is no female with *Grade* = 6 in our working sample. In Figures 2– 4, we show the estimates of the values of the conditional distribution functions and of the empirical distribution function

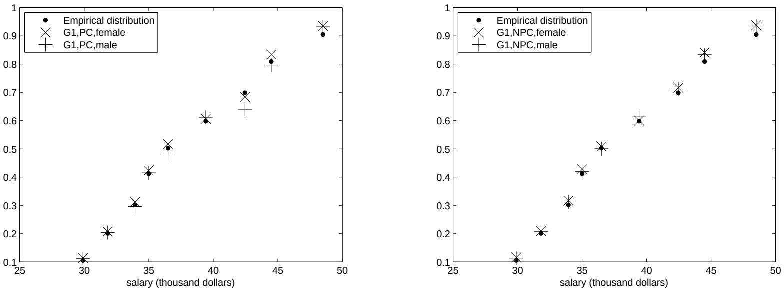
Table 6: The estimator $\widehat{\beta}$ and the componentwise resampling-based confidence intervals (RCI) (levels 0.9 and 0.95) for Fifth National Bank of Springfield salary data: single-index in law model.

Variable	Estimation	0.9 RCI	0.95 RCI
$Year1$	0.5135	(0.4853,0.5627)	(0.4783 ,0.5910)
Age	0.7271	(0.6894, 0.7530)	(0.6754, 0.7755)
$Year2$	0.0862	(0.04879, 0.1038)	(0.0364 , 0.1140)
$Gender$	-0.7831	$(-0.8395, -0.6578)$	$(-0.8677, -0.6339)$
$PCJob$	0.6899	(0.6751,0.9341)	(0.6521, 0.9983)

for ten values of the response. The response values were determined as the empirical deciles of the observed responses. We plot the kernel estimates of the conditional distribution functions for three different job levels (*Grade* = 1, 3 and 5, respectively). For each of the three job levels, we compute the estimates of the conditional distribution given $X = x$ for four different values of x. These values x correspond to all the possible outcomes for the variables *Gender* and *PC Job*. The components corresponding to the covariates *Year1*, *Age*, *Year2*, are set equal to the average values of the subsamples obtained with the given job level, and with *PCJob = 1* or 0, for each gender. For each value of the conditional distribution function estimated by kernel smoothing, we selected the bandwidth by least-squares cross-validation. In most cases, the figures reveal little difference between the distribution functions for female and male, which confirms the usual conclusion found in the literature, i.e., there is no evidence in the Fifth National data set that the female employees are discriminated against; see, e.g., [9].

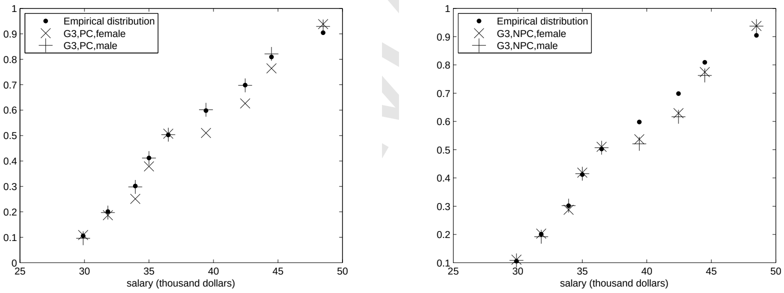
6. Appendix

Assumption 1. *The following conditions hold true.*



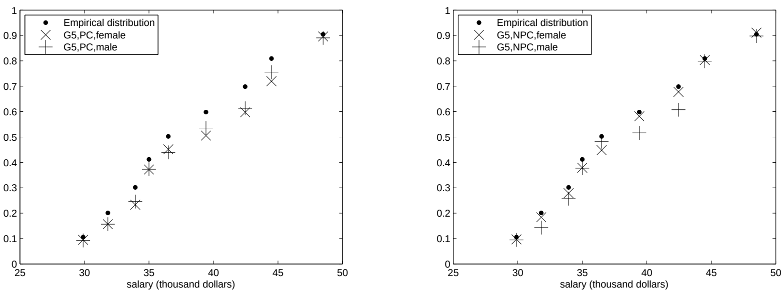
(a) the job is computer related (b) the job is not computer related

Figure 2: The conditional distribution function of the Fifth National Bank salary data set for $Grade = 1$: $Year1$, Age, $Year2$ take the sample mean value given $Grade = 1$ and the values of $Gender$ and $PC\ Job$



(a) the job is computer related (b) the job is not computer related

Figure 3: The same plots as in Figure 2 in the case $Grade = 3$.



(a) the job is computer related (b) the job is not computer related

Figure 4: The same plots as in Figure 2 in the case $Grade = 5$.

1. *The observations* $(Y_1^\top, X_1^\top)^\top, \dots, (Y_n^\top, X_n^\top)^\top$ *are independent copies of* $(Y^\top, X^\top)^\top \in \mathbb{R}^d \times \mathbb{R}^p$.
2. *The parameter set is* $\mathcal{B} = \{1\} \times \mathcal{B}'$ *and* $\mathcal{B}' \subset \mathbb{R}^{p-1}$ *is a compact set. The vector* $\beta_0 \in \mathcal{B}$ *satisfying Condition* (5) *is the unique element* $\mathcal{B}$*. For any* $\beta \in \mathcal{B}$*, the random variable* $X^\top \beta$ *has a density* f_β *such that* $\sup_{\beta \in \mathcal{B}} \sup_{z \in \mathbb{R}} f_\beta(z) < \infty$.
3. *We have* $\sup_{u \in \mathcal{U}} \sup_{\beta \in \mathcal{B}} \sup_{z \in \mathbb{R}} |\mathrm{E}\{T_u(Y) \mid X^\top \beta = z\}| f_\beta(z) < \infty$ *and*
$$\lim_{\delta \to 0} \sup_{\beta \in \mathcal{B}} \sup_{z \in \mathbb{R}} |f_\beta(z + \delta) - f_\beta(z)| = 0,$$
as well as
$$\lim_{\delta \to 0} \sup_{u \in \mathcal{U}} \sup_{\beta \in \mathcal{B}} \sup_{z \in \mathbb{R}} \big| [\mathrm{E}\{T_u(Y) \,|\, X^\top \beta = \cdot\} f_\beta](z + \delta) \\ - [\mathrm{E}\{T_u(Y) \mid X^\top \beta = \cdot\ \} f_\beta])(z) \big| = 0.$$
4. *The family of transformations* $\{T_u : u \in \mathcal{U}\}$ *is a VC-class (or Euclidean) for an envelope with finite moment of order* $4 + \rho$ *for some* $\rho > 0$.
5. *The value* $\boldsymbol{\beta}_0$ *is a well-separated point of minimum for* $Q(\beta)$ *defined in Eq.* (6) *with* $\omega(x) = \exp(-\|x\|^2/2)$ *and* μ *equal to the distribution* F_Y *of the observations* Y*, i.e., for any* $\varepsilon > 0$, $\inf_{\beta \in \mathcal{B}, \|\beta - \beta_0\| \geq \varepsilon} Q(\beta) > Q(\beta_0)$.
6. *The kernel* K *is a univariate integrable function with bounded variation. The bandwidth* h *satisfies the condition* $h + n^{-1} h^{-2} \to 0$.

Let us introduce some notation. Let $\widetilde{X} \in \mathbb{R}^{p-1}$ be the $(p-1)$-dimensional vector of the last components of X. Below, $(\widetilde{X})_r$ (resp. $(\widetilde{X}\widetilde{X}^\top)_{rq}$) denotes the rth components (resp. the $rq-$entry) of the vector $\widetilde{X}$ (resp. matrix $\widetilde{X}\widetilde{X}^\top$). If A is a matrix with real entries, $\|A\| = \sqrt{\mathrm{trace}(A^\top A)}$. In the following, ∂_z (resp. ∂^2_{zz}) denotes the first (resp. second) order derivative with respect to z.

Assumption 2. *The following conditions hold true.*

1. *There exists* $a > 0$ *such that* $\mathrm{E}\{\exp(a\|X\|)\} < \infty$.
2. *The subvector* β_0 *built with the last* $p-1$ *components belong to the interior of* $\mathcal{B}'$*, where* $\mathcal{B} = \{1\} \times \mathcal{B}'$.

3. $\sup_{z\in\mathbb{R}} \mathrm{E}(\|\widetilde{X}\|^4 \mid X^\top\beta_0 = z) f_{\beta_0}(z) < \infty$

4. *For all* $u \in \mathcal{U}$, *the functions* $z \mapsto \mathrm{E}\{T_u(Y) \mid X^\top\beta_0 = z\}$, $z \mapsto f_{\beta_0}(z)$, $z \mapsto \mathrm{E}\{(\widetilde{X})_r \mid X^\top\beta_0 = z\}$, *and* $z \mapsto \mathrm{E}\{(\widetilde{X}\widetilde{X}^\top)_{rq} \mid X^\top\beta_0 = z\}$ *are four times continuously differentiable and the derivatives up to order four are bounded. The fourth order derivative are Lipschitz functions. The Lipschitz constant is independent of* u *in the case of the fourth-order derivative of* $\mathrm{E}\{T_u(Y) \mid X^\top\beta_0 = z\}$.

5. $\sup_{u\in\mathcal{U}} \sup_z \mathrm{E}\{T_u^2(Y) \mid X^\top\beta_0 = z\} < \infty$.

6. *Let* A *be the set of values* $u \in \mathcal{U}$ *such that*

$$\mathrm{var}\left[\{\widetilde{X} - \mathrm{E}(\widetilde{X} \mid X^\top\beta_0)\}\partial_z[\mathrm{E}\{T_u(Y) \mid \cdot\}](X^\top\beta_0)\right] \tag{13}$$

is positive definite. Then $F_Y(A) > 0$.

7. *Let* $z \mapsto \lambda_\beta(z; u)$ *denote any of the four functions at point* (4) *above, considered for each* $\beta \in \mathcal{B}$, *and their derivatives up to the second order. Then, the family of functions* $\{\lambda_\beta : \beta \in \mathcal{B}, u \in \mathcal{U}\}$ *is a VC-class (or Euclidean) for an envelope having a finite moment of order 8. Moreover, for any sequence* $b_n \to 0$,

$$\sup_{\|\beta-\beta_0\|\le b_n} \sup_{z\in\mathbb{R}} \sup_{z\in\mathbb{R}} |\lambda_\beta(z;u) - \lambda_{\beta_0}(z;u)| \to 0.$$

8. *The kernel* K *is a symmetric and twice continuously differentiable univariate density with the second order derivative with bounded variation. Moreover, for* $\kappa = 1, 2$, $\int_{\mathbb{R}} |K^{(\kappa)}(u)|du < \infty$, *where* $K^{(\kappa)}$ *denotes the* κ*th derivative of* K.

9. $nh^4 \to 0$ *and* $nh^{3+a} \to \infty$ *for some* $a \in (0, 1)$.

6.1. Proofs

Proof of Lemma 1. Let $\mathcal{F}[\omega](v) = \int_{\mathbb{R}^p} e^{-2\pi i x^\top v}\omega(x)dx$, $u \in \mathbb{R}^p$, denote the Fourier Transform of ω. If $\mathcal{F}[\omega]$ is integrable, by the Inverse Fourier Transform formula and Fubini's Theorem, we can write

$$
\begin{aligned}
Q(\beta) &= \int_{\mathcal{U}} \mathrm{E}\{\omega(X_1 - X_2) g_u(Y_1, X_1^\top \beta; \beta)^\top g_u(Y_2, X_2^\top \beta; \beta)\} d\mu(u) \\
&= \int_{\mathcal{U}} \mathrm{E}\left\{ g_u(Y_1, X_1^\top \beta; \beta)^\top g_u(Y_2, X_2^\top \beta; \beta) \int_{\mathbb{R}^p} e^{2\pi i (X_1 - X_2)^\top v} \mathcal{F}[\omega](v) dv \right\} d\mu(u) \\
&= \int_{\mathcal{U}} \int_{\mathbb{R}^p} \left\| \mathrm{E}\left[\mathrm{E}\left\{ g_u(Y, X^\top \beta; \beta) \mid X \right\} e^{2\pi i X^\top v} \right] \right\|^2 \mathcal{F}[\omega](v) dv d\mu(u).
\end{aligned}
$$

By the fact that $\mathcal{F}[\omega]$ is positive, one has $Q(\beta) \geq 0$ for all $\beta \in \mathcal{B}$. Using also the uniqueness of the Fourier Transform, one can deduce that

$$
Q(\beta) = 0 \;\Leftrightarrow\; \mathrm{E}\left\{ g_u(Y, X^\top \beta; \beta) \mid X \right\} = 0 \text{ almost surely, for } \mu\text{−almost all } u \in \mathcal{U}.
$$

The conclusion of the lemma follows from the definition of the functions g_u and the transformation T_u. ■

Proof of Proposition 2. The proof of the asymptotic normality is quite lengthy and hence will be provided in the Supplementary Material. Concerning the consistency, by Assumptions 1–5, β_0 is a well-separated point of minimum for $Q(\beta)$. Thus, it suffices to prove that

$$
\sup_{\beta \in \mathcal{B}} \left| \widehat{Q}_n(\beta) - Q(\beta) \right| = o_{\mathbb{P}}(1). \tag{14}
$$

See, e.g., Theorem 5.7 in [25]. For this purpose, let us simplify notation and write $\widehat{g}_{u,i}(\beta)$ instead of $\widehat{g}_u(Y_i, X_i^\top \beta; \beta)$. By Lemma 4,

$$
\begin{aligned}
\sup_{\beta \in \mathcal{B}} \sup_{u \in \mathcal{U}} \Bigg| & \frac{1}{n^2} \sum_{i,j=1}^{n} \widehat{g}_u(Y_i, X_i^\top \beta; \beta)^\top \widehat{g}_u(Y_j, X_j^\top \beta; \beta) \omega_{ij} \\
& - \frac{1}{n^2} \sum_{i,j=1}^{n} g_u(Y_i, X_i^\top \beta; \beta)^\top g_u(Y_j, X_j^\top \beta; \beta) \omega_{ij} \Bigg| = o_{\mathbb{P}}(1).
\end{aligned}
$$

Next, by the uniform law of large numbers for Glivenko–Cantelli classes of functions (see, e.g., Theorem 19.4 in [25]), we deduce

$$
\begin{aligned}
\sup_{\beta \in \mathcal{B}} \sup_{u \in \mathcal{U}} \Bigg| & \frac{1}{n^2} \sum_{i,j=1}^{n} \widehat{g}_u(Y_i, X_i^\top \beta; \beta)^\top \widehat{g}_u(Y_j, X_j^\top \beta; \beta) \omega_{ij} \\
& - \mathrm{E}\{g_u(Y_1, X_1^\top \beta; \beta)^\top g_u(Y_2, X_2^\top \beta; \beta) \omega_{12}\} \Bigg| = o_{\mathbb{P}}(1).
\end{aligned}
$$

From this, it follows that

$$\sup_{\beta\in\mathcal{B}}\left|\widehat{Q}(\beta)-\int_{\mathcal{U}}\mathrm{E}\{g_u(Y_1,X_1^\top\beta;\beta)^\top g_u(Y_2,X_2^\top\beta;\beta)\omega_{12}\}d\mu_n(u)\right|=o_{\mathbb{P}}(1).$$

Next, by the uniform law of large numbers for Glivenko–Cantelli classes of functions,

$$\sup_{\beta\in\mathcal{B}}\left|\int_{\mathcal{U}}\mathrm{E}\{g_u(Y_1,X_1^\top\beta;\beta)^\top g_u(Y_2,X_2^\top\beta;\beta)\omega_{12}\}d\mu_n(u)-Q(\beta)\right|=o_{\mathbb{P}}(1).$$

Gathering facts, we deduce that the uniform convergence in Eq. (14) holds true, and thus $\widehat{\beta}$ is consistent in probability. ■

Lemma 4. *Under Assumption 1,*

$$\sup_{1\le i\le n}\sup_{u\in\mathcal{U}}\sup_{\beta\in\mathcal{B}}\left\|\widehat{g}_u(Y_i,X_i^\top\beta;\beta)-g_u(Y_i,X_i^\top\beta;\beta)\right\|=o_{\mathbb{P}}(1).$$

Proof of Lemma 4. The result follows from the following two properties:

$$\sup_{1\le i\le n}\sup_{\beta\in\mathcal{B}}\left\|\widehat{f_\beta(X_i^\top\beta)}-f_\beta(X_i^\top\beta)\right\|=o_{\mathbb{P}}(1)$$

and

$$\sup_{1\le i\le n}\sup_{u\in\mathcal{U}}\sup_{\beta\in\mathcal{B}}\left\|\widehat{\mathrm{E}\{T_u(Y_i)\mid X_i^\top\beta\}f_\beta(X_i^\top\beta)}\right.$$
$$\left.-\mathrm{E}\{T_u(Y_i)\mid X_i^\top\beta\}f_\beta(X_i^\top\beta)\right\|=o_{\mathbb{P}}(1).\quad(15)$$

Since the first property is a particular case of the second, we only provide the justification for Eq. (15). The latter property is a direct consequence of the following statements:

$$\sup_{z\in\mathbb{R}}\sup_{u\in\mathcal{U}}\sup_{\beta\in\mathcal{B}}\left|\mathrm{E}[T_u(Y)K\{(X^\top\beta-z)/h\}]\right.$$
$$\left.-\mathrm{E}\left[T_u(Y)\mid X^\top\beta=z\right]f_\beta(z)\right|=o(1)\quad(16)$$

and

$$\sup_{z\in\mathbb{R}}\sup_{u\in\mathcal{U}}\left|\frac{1}{nh}\sum_{k=1}^n T_u(Y_k)K\{(X_k^\top\beta-z)/h\}\right.$$
$$\left.-\mathrm{E}\left[T_u(Y)K\{(X^\top\beta-z)/h\}\right]\right|=o_{\mathbb{P}}(1).\quad(17)$$

Statement (16) follows by a standard change of variables and Assumptions 1–3. For the uniform convergence in Eq. (17), it suffices, for instance, to use the Maximal Inequality; see Theorem 3.1 in [26]. In that result, it suffices to take p sufficiently large to ensure that $(4p-2)/(p-1) \leq 4+\rho$, with ρ from Assumptions 1–4, and apply the maximal inequality with $\delta = h^{1/2}$. Deduce that

$$\sup_{z\in\mathbb{R}} \sup_{u\in\mathcal{U}} \left| \frac{1}{nh} \sum_{k=1}^{n} T_u(Y_k) K\{(X_k^\top \beta - z)/h\} - \mathrm{E}\left[T_u(Y) K\{(X^\top \beta - z)/h\}\right] \right| = O_{\mathbb{P}}(n^{-1/2} h^{-1/2} \ln^{1/2} n) = o_{\mathbb{P}}(1).$$

Now the proof is complete. ■

References

References

[1] A. Anestis, G. Gérard, W.M. Ian, Bayesian estimation in single-index models, Statistica Sinica 14 (2004) 1147–1164.

[2] S.C. Albright, W.L. Winston, C.J. Zappe, Data Analysis and Decision Making with Microsoft Excel, Duxbury Press, Pacific Grove, CA, 1999.

[3] R.J. Carroll, J.Q. Fan, M.P. Wand, Generalized partially linear single-index models, J. Amer. Statist. Assoc. 92 (1997) 477–489.

[4] J.M. Chambers, W.S. Cleveland, B. Kleiner, P.A. Tukey, Graphical Methods for Data Analysis, Wadsworth, Belmont, CA, 1983.

[5] C.T. Chiang, M.Y. Huang, New estimation and inference procedures for a single-index conditional distribution model, J. Multivariate Anal. 111 (2012) 271–285.

[6] X. Cui, W. Härdle, L. Zhu, The EFM approach for single-index models, Ann. Statist. 39 (2011) 1658–1688.

[7] M. Delecroix, W. Härdle, M. Hristache, Efficient estimation in conditional single-index regression, J. Multivariate Anal. 86 (2003) 213–226.

[8] M. Delecroix, M. Hristache, V. Patilea, On semiparametric M-estimation in single-index regression, J. Statist. Plan. Inference 136 (2006) 730–769.

[9] J. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, Ann. Statist. 32 (2004) 928–961.

[10] P. Hall, Q. Yao, Approximating conditional distribution functions using dimension reduction, Ann. Statist. 33 (2005) 1404–1421.

[11] W. Härdle, P. Hall, H. Ichimura, Optimal smoothing in single-index models, Ann. Statist. 21 (1993) 157–178.

[12] M. Hristache, A. Juditsky, V. Spokoiny, Direct estimation of the index coefficient in a single-index model, Ann. Statist. 29 (2001) 595–917.

[13] J.L. Horowitz, Semiparametric and nonparametric methods in econometrics, Springer, New York, 2009.

[14] H. Ichimura, Semiparametric least squares (SLS) and weighted SLS estimation of single index models, J. Econometrics 58 (1993) 71–120.

[15] Z. Jin, Z. Ying, L.J. Wei, A simple resampling method by perturing the minimand, Biometrika 88 (2001) 381–390.

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

[17] E. Kong, Y. Xia, Variable selection for the single-index model, Biometrika 94 (2007) 217–229.

[18] E. Kong, Y. Xia, A single-index quantile regression model and its estimation, Econometric Theory 28 (2012) 730–768.

[19] P. Lavergne,V. Patilea, Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory, J. Econometrics 177 (2013) 47–59.

[20] H. Liang, X. Liu, R. Li, C.L. Tsai, Estimation and testing for partially linear single-index models, Ann Statist. 38 (2010) 3811–3836.

[21] Y. Ma, L, Zhu, Efficient estimation in sufficient dimension reduction, Ann. Statist. 41 (2013) 250–268.

[22] S. Maistre, V. Patilea, Nonparametric model checks of single-index assumptions, arXiv:1410.4934[math.ST] (2014)

[23] J.L. Powell, J.H. Stock, T.M. Stoker, Semiparametric estimation of index coefficients, Econometrica 51 (1989) 1403–1430.

[24] J.A. Reeds, Asymptotic number of roots of Cauchy location likelihood equations, Ann. Statist. 13 (1985) 775–784.

[25] A.D. van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.

[26] A.D. van der Vaart, J.A. Wellner, A local maximal inequality under uniform entropy, Electron. J. Statist. 5 (2011) 192–203.

[27] X. Yin, R.D. Cook, Dimension reduction for the conditional kth moment in regression, J. R. Statist. Soc. B 64 (2002) 159–175.

[28] Y. Yu, D. Ruppert, Penalized spline estimation for partially linear single index models, J. Amer. Statist. Assoc. 97 (2002) 1042–1054.

[29] J. Zhang, Z. Feng, P. Xu, Estimating the conditional single-index error distribution with a partial linear mean regression, Test 24 (2015) 61–83.