Submitted to the Annals of Statistics

TWO-STEP SEMIPARAMETRIC EMPIRICAL LIKELIHOOD INFERENCE

By Francesco Bravo [‡], Juan Carlos Escanciano [*,§] and Ingrid Van Keilegom [†,¶]

University of York [‡], Universidad Carlos III de Madrid [§] and KU Leuven [¶]

In both parametric and certain nonparametric statistical models, the empirical likelihood ratio satisfies a nonparametric version of Wilks' theorem. For many semiparametric models, however, the commonly used two-step (plug-in) empirical likelihood ratio is not asymptotically distribution-free, that is, its asymptotic distribution contains unknown quantities and hence Wilks' theorem breaks down. This article suggests a general approach to restore Wilks' phenomenon in two-step semiparametric empirical likelihood inferences. The main insight consists in using as the moment function in the estimating equation the influence function of the plug-in sample moment. The proposed method is general; it leads to a chi-squared limiting distribution with known degrees of freedom; it is efficient; it does not require undersmoothing; and it is less sensitive to the first-step than alternative methods, which is particularly appealing for high-dimensional settings. Several examples and simulation studies illustrate the general applicability of the procedure and its excellent finite sample performance relative to competing methods.

1. Introduction. Since its introduction as a nonparametric likelihood alternative to likelihood-type bootstrap methods for constructing confidence regions, Owen's ([55, 56, 57]) empirical likelihood (EL henceforth) has been used extensively in both statistics and econometrics. Such popularity is justified by the appealing theoretical properties of EL confidence regions: they tend to be more concentrated in places where the density of the parameter estimator is greatest; they can be Bartlett corrected ([22] for the so-called smooth function model, [10] for exactly identified estimating equations models, [16] for exactly identified estimating equations models with nuisance parameters and [17] for over-identified estimating equations models); they do

[*]Juan Carlos Escanciano gratefully acknowledges support by the Ministerio Economia y Competitividad (Spain), ECO2017-86675-P & MDM 2014-0431, and by Comunidad de Madrid (Spain), MadEco-CM S2015/HUM-3444.

[†]Ingrid Van Keilegom acknowledges financial support from the European Research Council (2016-2021, Horizon 2020/ERC grant agreement No. 694409).

MSC 2010 subject classifications: Primary 62M10; secondary 62G10

Keywords and phrases: Empirical likelihood, Semiparametric inference, High-dimensional parameters, Wilks' phenomenon

not require estimation of scale (internal studentization) and skewness; and finally, they are range preserving and transformation respecting. Furthermore, [21] show that in linear exponential families empirical and parametric likelihood surfaces are quite close in terms of their asymptotic distribution. Specifically, the chi-squared approximations to the distributions of the empirical and likelihood ratios, as well as the asymptotic normality of their signed squared root differ in terms of order $O(n^{-1})$, where n is the sample size. See [57] for a comprehensive review of these properties and a number of applications geared mainly towards finite-dimensional statistical models.

More recently the EL method has been used in nonparametric and semiparametric models. For nonparametric models [25] considered sieve empirical likelihood for testing nonparametric hypotheses about nonparametric functions, and showed that an appropriately rescaled sieve EL ratio test has an asymptotic chi-squared calibration, with the scaling constant and degrees of freedom being independent of nuisance parameters, in other words the so-called Wilks' phenomenon ([80]) (i.e. the likelihood ratio statistic is asymptotically distribution-free and converges to a chi-squared distribution) holds for the EL. In semiparametric models [7] has shown that Wilks' Theorem also holds in certain "highly smooth" cases, see Remark 2.3 in [32] for discussion.

For semiparametric models the most popular method uses a two-step (plug-in) procedure in which the first-step estimator replaces the infinite-dimensional nuisance parameter, while in the second step the plug-in EL ratio is used to obtain inferences for the finite-dimensional parameter of interest. This two-step semiparametric EL approach has been considered by a number of authors, including [75] for partially linear models, [86] for single-index models, [89, 81, 30] for various censored regression problems, [77, 78, 79, 76, 74, 70] for various missing data problems, and [7, 48, 49, 12, 11] for other semiparametric problems. [18, 88] provide recent surveys on EL inference in the context of semiparametric regression models.

In general, the two-step semiparametric plug-in method does not yield asymptotically pivotal test statistics. Indeed, as shown in a general setting by [32], the asymptotic distribution of the resulting plug-in EL ratio is generally a weighted sum of chi-squared random variables with the weights depending (often in a complicated way) on the distribution of the data. Thus, in most situations the Wilks' phenomenon does not hold for the two-step EL ratio, so to obtain asymptotically valid EL inferences three main proposals have been put forward in the literature. The first and most common proposal is the bootstrap, as suggested for example by [74, 32]. The proposed bootstrap methods are general in nature, but they require re-estimating the

semiparametric model in each bootstrap iteration, and thus are computationally very expensive. The second proposal consists in adjusting the EL by a scale factor such that the adjusted (or rescaled) EL ratio is asymptotically pivotal. [76] proposed a specific scale factor; more general adjustments have been proposed by [86, 84, 12]. Although sometimes effective, these adjustments typically involve explicit estimation of various covariance matrices, which can be very complicated to be carried out in practice. Furthermore the internal studentization property of EL is not exploited and this can negatively affect the finite sample performance of the resulting EL statistic. The third proposal exploits that in some specific cases it is possible to modify the original estimating equation in such a way that the effect of the first-step estimation is removed. This approach has been called in the EL literature "bias-reduced or bias-corrected EL". A review of papers using this approach is provided in Section 2.3. As shown first by [91] in the context of a partially linear single-index model, this approach has the additional advantage of not requiring undersmoothing (the bias of the first-step going to zero faster than its standard deviation), much in contrast to bootstrap and adjusted based methods, but it is not clear how the method works, that is, how the modified estimating equations were obtained in the first place for the specific models considered, and how similar estimating equations could be built for other semiparametric models.

This leads us to the main contribution of this article, which is to propose a theoretical justification of "bias-corrected EL" methods in general semiparametric models. This theoretical justification includes a general construction of the method, proving Wilks' Theorem and establishing the efficiency of the procedure. The main insight consists in using as the moment function in the estimating equation the influence function of the plug-in sample moment. This entails correcting the original estimating equations based on the pathwise derivative with respect to the infinite-dimensional parameter. Pathwise differentiation arises naturally in the context of semiparametric models, and has been used extensively both in the statistical and econometric literatures; see, e.g., [40, 59, 8, 71, 51]. Our method does not require bootstrap and preserves the internal studentization property of the EL ratio. Thus, confidence regions can be computed with critical values from a standard chi-squared distribution.

There are a number of additional benefits that result from our method. The proposed modified tests are efficient (Asymptotically Maximin and Asymptotically Uniformly Most Powerful and Invariant, see Section 3.3). To our knowledge, this is the first article establishing efficiency in a two-step semiparametric testing setting for EL. We also find that, in general, with

the modified test there is no need for undersmoothing, which means that, in contrast to alternative methods, the proposed inference method is asymptotically valid with a cross-validated bandwidth for the first-step. Confidence intervals based on the new method tend to have a more accurate coverage than alternative procedures, that is also less sensitive to bandwidths. These advantages of efficiency and robustness to high-dimensional first steps do not generally hold for alternative procedures without the correction (e.g. bootstrap methods). The theoretical results above are confirmed by two Monte Carlo simulations; one in the context of average treatment effects in observational studies, and one in the context of nonlinear estimating equations with missing data.

The rest of the article is organized as follows: the next section introduces the statistical model, the method and provides some heuristic explanation as to why the proposed method works, while Section 3 presents the main results. The main results include establishing Wilks' Theorem and the efficiency for our modified estimating equation approach. Sections 4 and 5 contain, respectively, all the examples and the results of the simulations that are used to illustrate the theory and the finite sample performance of the proposed method. Section 6 is a discussion section. Section 7 contains the proofs of the main results. The Supplementary Material [13] consists of four appendices that are organized as follows. Appendix A gathers all the proofs for the examples. Appendix B proves the validity of a general numerical algorithm for estimating the pathwise derivative, Appendix C extends the main result of the paper to the case of over-identified models, and Appendix D shows an auxiliary result regarding Donsker and Glivenko-Cantelli classes. All these results are of independent interest.

2. The Statistical Model and Method.

2.1. *Two-step semiparametric inference.* Let Z be a random vector defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and with values on $\mathcal{S}_Z \subseteq \mathbb{R}^{d_z}$, and let $\{Z_i\}_{i=1}^n$ be independent copies of Z. Assume Z satisfies the estimating equations

$$\mathbb{E}\left[g\left(Z, \theta_0, \eta_0\right)\right] = 0, \tag{2.1}$$

where $g(\cdot) : \mathcal{S}_Z \times \Theta \times \mathcal{E} \to \mathbb{R}^p$ is a vector-valued measurable known function, $\theta_0 \in \Theta \subset \mathbb{R}^p$ denotes the finite-dimensional parameter of interest, and $\eta_0 \in \mathcal{E}$ denotes the possibly infinite-dimensional nuisance parameter, taking values in a semi-metric space $\mathcal{E}$. The statistical model (2.1) is rather general, as it does not require the full specification of the distribution of Z,

albeit it does also include models that can be estimated with semiparametric maximum and quasi maximum likelihood methods, for which (2.1) may represent, respectively, the score and quasi score vector. We consider just-identified models for simplicity of notation, but our theory can be equally applied to over-identified models (i.e. number of equations larger than p, thereby extending [61] to the semiparametric case, where possibly infinite-dimensional nuisance parameters $\eta_0 \in \mathcal{E}$ are present in (2.1). Details can be found in Appendix C in the Supplementary Material, Theorem C.1.)

Under this setting, we aim to construct EL based tests or confidence regions for θ_0 using the sample $\{Z_i\}_{i=1}^n$. If $\eta_0 \in \mathcal{E}$ is known, the standard EL $(1-\alpha)$-confidence region is

$$\left\{\theta \in \Theta : -2\log EL_n(\theta, \eta_0) < \chi^2_{p,1-\alpha}\right\},$$

where $EL_n(\theta, \eta_0)$ is the likelihood ratio function

$$EL_n(\theta, \eta_0) := \max\left\{\prod_{i=1}^n np_i : p_i > 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g\left(Z_i, \theta, \eta_0\right) = 0\right\},$$

and $\chi^2_{p,\alpha}$ is the α-quantile of the chi-squared distribution with p degrees of freedom, $\alpha \in (0,1)$. In practice, η_0 is unknown and the standard two-step (plug-in) approach defines confidence regions of the form $\{\theta \in \Theta : -2\log EL_n(\theta, \widehat{\eta}) < c\}$, for a suitable constant c to be determined and a first-step consistent estimator $\widehat{\eta}$ for η_0. [32] have investigated this two-step method in a general setting, and have shown that if

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n g\left(Z_i, \theta_0, \widehat{\eta}\right) \xrightarrow{d} U \tag{2.2}$$

$$\frac{1}{n}\sum_{i=1}^n g\left(Z_i, \theta_0, \widehat{\eta}\right) g'\left(Z_i, \theta_0, \widehat{\eta}\right) \xrightarrow{\mathbb{P}} V, \tag{2.3}$$

for a non-singular matrix V (for any matrix A, A' denotes the transpose of A), then

$$-2\log EL_n(\theta_0, \widehat{\eta}) \xrightarrow{d} U'V^{-1}U, \tag{2.4}$$

provided some further regularity conditions hold. This convergence result is a generalization of the classical result by [55, 56]. The asymptotic distribution of the quadratic form $U'V^{-1}U$ is typically not chi-squared, but rather a weighted sum of chi-square random variables. To explain the discrepancy

between one-step and two-step settings, notice that a "functional" Taylor argument leads to the expansion

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(Z_i,\theta_0,\widehat{\eta}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g\left(Z_i,\theta_0,\eta_0\right)+\phi\left(Z_i,\theta_0,h_0\right)\right\}+o_{\mathbb{P}}(1), \tag{2.5}$$

where $\phi\left(Z_i,\theta_0,h_0\right)$ is the so–called pathwise derivative of $\eta\to\mathbb{E}\left[g\left(Z_i,\theta_0,\eta\right)\right]$, well explained in [71, 51], which accounts for the asymptotic impact of the first-step estimate $\widehat{\eta}$ on the sample analog of the moment $\mathbb{E}\left[g\left(Z_i,\theta_0,\eta\right)\right]$, and where h_0 may include η_0 and other nonparametric objects that may appear in the influence function as a result of "functional differentiation". Hence, if (2.5) and certain finite moment conditions hold, an application of the standard Central Limit Theorem (CLT) yields $U\overset{d}{=}N(0,\Sigma)$ in (2.2), where $\overset{d}{=}$ stands for equality in distribution, and

$$\Sigma:=\mathbb{E}\left[\left(g\left(Z,\theta_0,\eta_0\right)+\phi\left(Z,\theta_0,h_0\right)\right)\left(g\left(Z,\theta_0,\eta_0\right)+\phi\left(Z,\theta_0,h_0\right)\right)'\right]; \tag{2.6}$$

whereas a Uniform Law of Large Numbers (ULLN) yields (2.3) with $V=\mathbb{E}\left[g\left(Z,\theta_0,\eta_0\right)g'\left(Z,\theta_0,\eta_0\right)\right]$. These results imply that the limiting distribution in (2.4) is in general a weighted chi-squared distribution when $\phi\neq 0$; see [62], p.171.

2.2. *A new method: heuristics.* Let m denote the modified moment function (cf. (2.5))

$$m\left(Z,\theta_0,h_0\right):=g\left(Z,\theta_0,\eta_0\right)+\phi\left(Z,\theta_0,h_0\right),$$

and define the bias-corrected or modified EL ratio function as

$$MEL_n(\theta,h):=\max\left\{\prod_{i=1}^{n}np_i:p_i>0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_im\left(Z_i,\theta,h\right)=0\right\}.$$

Let $\widehat{h}$ be a consistent estimate of h_0 satisfying some conditions below. One of the main results of this article shows that under certain regularity conditions

$$R_{1-\alpha}:=\left\{\theta\in\Theta:-2\log MEL_n(\theta,\widehat{h})<\chi^2_{p,1-\alpha}\right\},$$

forms an asymptotically valid $(1-\alpha)-$confidence region for θ_0. This follows from the fact that the test that rejects $H_0:\theta=\theta_0$ against $H_1:\theta\neq\theta_0$ when $-2\log MEL_n(\theta_0,\widehat{h})>\chi^2_{p,1-\alpha}$ has an asymptotic level $\alpha\in(0,1)$.

To show these results, we prove in Theorem 3.1 below

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_0,\widehat{h}\right) \xrightarrow{d} N(0,\Sigma) \tag{2.7}$$

$$\frac{1}{n}\sum_{i=1}^{n} m\left(Z_i,\theta_0,\widehat{h}\right) m'\left(Z_i,\theta_0,\widehat{h}\right) \xrightarrow{\mathbb{P}} \Sigma, \tag{2.8}$$

where Σ is defined in (2.6). The key asymptotic results (2.7) and (2.8) established in this article, and the general convergence theorem in [32], imply that Wilks' phenomenon is restored, i.e.

$$-2\log MEL_n(\theta_0,\widehat{h}) \xrightarrow{d} \chi^2_p.$$

We provide now some heuristics on the validity of (2.7), and refer to Section 3 below for a formal discussion. Under certain regularity conditions, the influence function $m(Z_i,\theta,h_0)$ belongs to the orthocomplement of the tangent space of nuisance parameters, see [8]. This implies that, modulo some regularity conditions, the following invariance property holds

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i,\theta_0,\widehat{h}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i,\theta_0,h_0) + o_{\mathbb{P}}(1). \tag{2.9}$$

Intuitively, m is a projection of g, say $m=\Pi g$, and projection operators are idempotent, i.e. they satisfy $\Pi^2=\Pi$. In particular, $\Pi m = m$, which explains (2.9) and hence (2.7). The projection operator Π projects onto the orthocomplement of the tangent space of nuisance parameters, but its actual form depends on the limit of the estimator $\widehat{h}$ and the model.

2.3. *Identifying pathwise derivatives.* The pathwise derivative $\phi(\cdot)$ in (2.5) plays a fundamental role in our method, as it is used to construct m. This section discusses the identification of $\phi(\cdot)$ in a general setting. Let F_0 denote the distribution of Z. Let L_2^0 be the subspace of measurable real-valued functions $d(Z)$ such that $\mathbb{E}(d(Z))=0$ and $\mathbb{E}(d^2(Z))<\infty$, where all expectations, unless otherwise stated, are with respect to F_0. Following [51], we denote by $\eta(F)$ the probabilistic limit of the first-step estimator $\widehat{\eta}$ when the distribution of Z is $F\in\mathcal{F}$, where $\mathcal{F}$ is a class of distributions that is unrestricted, except for regularity conditions. The precise generality of $\mathcal{F}$ is defined as follows. Let $\{F_t\}$ be a regular parametric (one-dimensional) submodel, $t\in(0,\varepsilon)\to F_t\in\mathcal{F}$, satisfying the classical mean-squared differentiability assumption with score s, i.e. as $t\downarrow 0$,

$$\int\left[\frac{dF_t^{1/2}-dF_0^{1/2}}{t}-\frac{1}{2}dF_0^{1/2}s\right]^2 = o(1).$$

The generality of $\mathcal{F}$ is that the set of scores $\{s\}$ of regular paths in $\mathcal{F}$ is linear and dense in L_2^0. Define the functional $\mu : \mathcal{F} \longrightarrow \mathbb{R}^p$

$$\mu(F) := \mathbb{E}\left[g\left(Z, \theta_0, \eta(F)\right)\right] \qquad F \in \mathcal{F}. \tag{2.10}$$

If μ is differentiable at F_0 in the sense of [71], then for any regular path $\{F_t\}$ with score $s(\cdot)$ there exists a function $\phi\left(\cdot, \theta_0, \eta(F_0)\right) \in L_2^0$ such that

$$\left.\frac{\partial \mu(F_t)}{\partial t}\right|_{t=0} = \mathbb{E}\left[\phi\left(Z, \theta_0, \eta(F_0)\right) s(Z)\right]. \tag{2.11}$$

Moreover, since the set of scores $\{s\}$ is linear and dense in L_2^0, then $\phi(\cdot, \theta_0, \eta(F_0))$ is uniquely determined from (2.11) and $\phi(\cdot, \theta_0, \eta(F_0)) \in L_2^0$. That is, $\phi\left(\cdot, \theta_0, \eta(F_0)\right)$ is the influence function of the functional $\mu\left(\cdot\right)$, an observation that was first made by [51], p. 1357.

Equation (2.11) is a functional equation in ϕ. [51] used this equation to provide expressions for ϕ when $\eta_0 \equiv \eta(F_0)$ is a regression function or a density. The literature contains numerous examples where ϕ has been explicitly computed; see [8] for a comprehensive review of many of these examples. [37] have recently suggested a smoothed version of Hampel's ([28, 29]) characterization of influence functions as Gateaux derivatives, which can be applied to $\mu(F)$ to characterize ϕ. For cases where computing ϕ explicitly is difficult, either from [51] or from [37], we propose a fully automatic numerical method to estimate ϕ and prove the validity of our bias-corrected EL with the numerically estimated influence function. See Theorem A.1 in Appendix B in the Supplementary Material, which is a new result of independent interest.

2.4. *Bias-corrected EL: A review.* The bias-corrected EL ratio was first introduced in [91] for a semiparametric partially linear single-index model. Since then, this approach has been used in other semiparametric settings, including in [87] for semiparametric regressions with longitudinal data, in [83, 85, 70] for models with missing data, and in [90, 44, 69, 82] for other semiparametric problems. Explicit recognition of the benefits of using influence functions as estimating equations to obtain chi-squared limiting distributions for EL ratio tests is given in [89, 30]. [89] considered finite dimensional nuisance parameters, and although they discussed two applications in semiparametric models, no theoretical results were given for infinite dimensional nuisance parameters. [30] proposed using a special influence function for a scalar parameter defined through an estimating equation with right censored data. Relative to this literature, the main contribution of this article is to provide a general theory of bias-corrected EL in semiparametric models. This theory involves giving a new general construction of a bias-corrected

EL, including a result with a numerically estimated influence function (Appendix B in the Supplementary Material), proving Wilks' Theorem in a general setting (Section 3.2) and establishing the efficiency of the method (Section 3.3). Some examples in Section 4 illustrate the application of the general theory. Further applications of the general theory of this article are provided in [48, 49].

3. Main Results.

3.1. *Notation.* We first elaborate further on the model introduced in (2.1). Notice that, though we do not make it explicit in (2.1), the nuisance function $h_0(\cdot)$ may contain θ_0 as an additional argument. In what follows, we suppress θ_0 in the nuisance function h_0 to save space, but it should be understood conformably, i.e. $(\theta, h) := (\theta, h(\cdot, \theta))$. We assume that a first-step nonparametric estimator $\widehat{h}(\cdot)$ for $h_0(\cdot)$ is available with certain convergence properties as specified in Assumption A below. Let $|\cdot|$ denote the Euclidean norm, i.e. $|A| := (tr(A'A))^{1/2}$, where $tr(A)$ is the trace of the matrix A. For a measurable function g of Z, define $\|g\|_\infty := \sup_{z \in \mathcal{S}_Z} |g(z)|$ and $\|g\|_r := (\mathbb{E}[|g(Z)|^r])^{1/r}$, where $\mathcal{S}_Z$ is the support of Z. The function space $\mathcal{H}$, where h_0 belongs to, is endowed with a semi-metric $\|\cdot\|_{\mathcal{H}}$. For example, $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_\infty$ or $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_r$. Since we assume consistency of $\widehat{h}$ with respect to $\|\cdot\|_{\mathcal{H}}$, we can redefine $\mathcal{H}$ as $\mathcal{H}^\delta := \{h \in \mathcal{H} : \|h - h_0\|_{\mathcal{H}} \leq \delta\}$, for an arbitrarily small $\delta > 0$. For a measurable function f we denote $\mathbb{P}f := \int f d\mathbb{P}$,

$$\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^n f(Z_i) \text{ and } \mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f).$$

Henceforth, we will use the concepts of $\mathbb{P}$-Glivenko-Cantelli and $\mathbb{P}$-Donsker classes; see, e.g., [73] for definitions. For a generic random vector Z with absolute continuous distribution we denote by f_Z its (Lebesgue) density.

3.2. *Regularity conditions and Wilk's Theorem.* This section presents the main results in a formal way under a set of "high-level"assumptions. The motivation for these high-level assumptions is to widen the applicability of the approach, while avoiding repetition. The moment function g satisfies (2.1). Having discussed methods to identify the pathwise derivative ϕ of g, we now provide regularity conditions for the validity of our results assuming knowledge of ϕ. Appendix B in the Supplementary Material relaxes this assumption and proves Wilks' Theorem with a numerically estimated influence function.

We introduce the following regularity conditions.

Assumption A: The measurable function $m(\cdot, \theta_0, h)$ is such that:

(i) Stochastic equicontinuity in h: for all sequences of numbers $\delta_n \downarrow 0$,

$$\sup_{\|h-h_0\|_{\mathcal{H}} \leq \delta_n} |\mathbb{G}_n m(\cdot, \theta_0, h) - \mathbb{G}_n m(\cdot, \theta_0, h_0)| = o_{\mathbb{P}}(1).$$

(ii) Asymptotic "no bias" condition:

$$\mathbb{P}[m(\cdot, \theta_0, \widehat{h}) - m(\cdot, \theta_0, h_0)] = o_{\mathbb{P}}(n^{-1/2}).$$

(iii) $\mathbb{P}(\widehat{h} \in \mathcal{H}^{\delta}) \to 1$, for $\delta > 0$, and $\|\widehat{h} - h_0\|_{\mathcal{H}} = o_{\mathbb{P}}(1)$.

(iv) Uniform consistency: for all $\delta_n \downarrow 0$ and for $\nu = gg'$, $\nu = g\phi'$ and $\nu = \phi\phi'$,

$$\sup_{\|h-h_0\|_{\mathcal{H}} \leq \delta_n} |\mathbb{P}_n \nu(\cdot, \theta_0, h) - \mathbb{P}_n \nu(\cdot, \theta_0, h_0)| = o_{\mathbb{P}}(1).$$

Moreover, the matrix $\Sigma = \mathbb{E}[m(Z, \theta_0, h_0) m'(Z, \theta_0, h_0)]$ is positive definite and finite.

(v) $\mathbb{P}(MEL_n(\theta_0, \widehat{h}) = 0) \to 0$ and $\max_{1 \leq i \leq n} |m(Z_i, \theta_0, \widehat{h})| = o_{\mathbb{P}}(\sqrt{n})$.

Assumption A is a high-level condition that suffices for the validity of our method. The conditions in A(i-ii) are standard in the literature; see, e.g., [15]. Assumption A(i) is implied by the $\mathbb{P}-$Donsker property of the function class $\mathcal{F} := \{m(\cdot, \theta_0, h) : h \in \mathcal{H}^{\delta}\}$; see Appendix D in the Supplementary Material for primitive conditions for this. Related high-level assumptions to the asymptotic "no bias" condition have been considered extensively in the literature; see, for example, [8] p. 396, Theorem 6.1(i) in [34], p. 557, Section 25.8 in [72], Assumption H_2 in [7], or Condition M2 in [9]. Assumptions A(iii) and A(iv) are standard in the literature on semiparametric inference. Assumption A(v) is required in [32], who discussed sufficient conditions for it to hold. Next result shows that with our method Wilk's Theorem is restored.

THEOREM 3.1. *If Assumption A holds, then*

$$-2 \log MEL_n(\theta_0, \widehat{h}) \xrightarrow{d} \chi^2_p.$$

The verification of the asymptotic "no bias" condition A(ii) may be easy due to the special properties of the model (for example in certain convex models with the efficient score as moment function), but more generally it may also require considerable effort. The following assumption suffices for A(ii) to hold.

Assumption B: For some $\delta > 0$:

(i) The map $h \to M(h) = \mathbb{E}[m(Z, \theta_0, h)]$ from $\mathcal{H}^\delta$ to $\mathbb{R}^p$ satisfies, for all $h \in \mathcal{H}^\delta$, $|M(h) - M(h_0)| \leq c \|h - h_0\|_{\mathcal{H}}^\tau$ for constants $c > 0$ and $\tau > 1$.
(ii) $\mathbb{P}(\widehat{h} \in \mathcal{H}^\delta) \to 1$ and $\|\widehat{h} - h_0\|_{\mathcal{H}} = o_{\mathbb{P}}\left(n^{-1/2\tau}\right)$.

Assumption B(i) requires sufficient smoothness in the model. This condition holds if $M(h)$ is Frechet differentiable with a zero derivative and a Hölder continuous second derivative. Frechet differentiability is often satisfied in this context, see [26]. The proof of the following Lemma is trivial, and hence omitted.

LEMMA 3.2. *Assumption B implies A(ii).*

REMARK 3.1. *Undersmoothing is not required in the conditions on the first-step $\widehat{h}$. This is shown in our Examples below using kernel estimators for $\widehat{h}$. This is important, as cross-validation and related methods that choose the optimal bandwidth for estimation of the first-step are commonly used in practice. These bandwidths are ruled out by alternative methods that do not use our correction (e.g. bootstrap methods).*

REMARK 3.2. *An extension of Theorem 3.1 to the case of a numerically estimated influence function ϕ is given in Theorem B.1 of Appendix B in the Supplementary Material. This result is convenient for situations where computing ϕ directly is too involved.*

3.3. *Efficiency.* In this section we prove the efficiency, in the sense introduced below, of the modified EL procedure. Let us denote by ψ_n the test that rejects $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ when $-2 \log MEL_n(\theta_0, \widehat{h}) > \chi^2_{p,1-\alpha}$. Consider the local alternatives $H_{1n} : \theta_n = \theta_0 + \tau/\sqrt{n}$, where $\tau \neq 0$. To investigate the asymptotic behavior of ψ_n under the local alternatives H_{1n} we need the following assumption.

Assumption C: The measurable function $m(\cdot, \theta, h)$ satisfies:

(i) Stochastic equicontinuity in θ: for all sequences of numbers $\delta_n \downarrow 0$,

$$\sup_{|\theta - \theta_0| \leq \delta_n} |\mathbb{G}_n m(\cdot, \theta, h_0) - \mathbb{G}_n m(\cdot, \theta_0, h_0)| = o_{\mathbb{P}}(1).$$

(ii) $\theta_0 \in \Theta$, with $\Theta \subset \mathbb{R}^p$ open, and $\mathbb{E}[m(Z, \theta, h_0)]$ is continuously differentiable at θ_0, with non-singular derivative.
(iii) $\mathbb{E}[m(Z, \theta, h_0) m'(Z, \theta, h_0)]$ is continuous at θ_0 and, for some $\delta > 0$,

$$\mathbb{E}\left[\sup_{\theta \in \mathcal{N}_0} |m(Z, \theta, h_0)|^{2+\delta}\right] < \infty,$$

where $\mathcal{N}_0$ is a neighborhood of θ_0.

Assumption C is standard. Note this condition allows for non-smooth moment functions m as a function of θ and h. Under Assumption C(ii), we can define $G_0 = (\partial/\partial\theta')\mathbb{E}\left[m\left(Z,\theta_0,h_0\right)\right]$. The following matrix will play a fundamental role in efficiency considerations, $B^* = G_0'\Sigma^{-1}G_0$. The first concept of efficiency used here is that of an asymptotic maximin test. We give a basic introduction to this concept as follows. Let X follow a p dimensional normal distribution with mean μ and identity variance, and let a denote a positive fixed number. A maximin test for testing $\mu = 0$ against the alternative $\mu'\mu \geq a$ is one that maximizes the minimum power $\inf_{\mu\in\mathbb{R}^p:\mu'\mu\geq a}\mathbb{E}_\mu\left[\varphi(X)\right]$ over the set of all level α tests $\varphi\left(\cdot\right)$. It is well known (see e.g. [42] , pg. 55) that the maximin test has critical region $X'X \geq \chi^2_{p,1-\alpha}$. For further details on maximin tests see [42, 67]. A test for $H_0 : \theta = \theta_0$ against $H_{1n} : \theta_n = \theta_0 + \tau/\sqrt{n}$ is asymptotic maximin when its asymptotic local power function is that of the maximin test in the limiting experiment. Our first efficiency result shows that ψ_n is asymptotic maximin.

THEOREM 3.3. *Let Assumptions A and C hold under H_{1n}. Then, the test ψ_n is asymptotic maximin for testing $H_0 : \tau = 0$ against $H_1 : \tau'B^*\tau \geq a$, for any $a > 0$.*

We establish now an efficiency result for the modified test in a semiparametric setting. Efficient tests for restrictions on a finite-dimensional parameter in regular semiparametric models have been formally defined in [20]. For multivariate null hypotheses, these authors introduce the efficiency concept of *Asymptotically Uniformly Most Powerful and Invariant* test of level α, in short AUMPI(α), see [20], Section 5. Of course, when $p = 1$, alternative definitions of efficiency, which do not require invariance, are typically used. We refer to [20] for a comprehensive discussion of these efficiency concepts. See also [38] for an illuminating application in a regression context.

Recall the moment function $g\left(Z,\theta_0,\eta_0\right)$ satisfies (2.1) with first-steps given by η_0. To establish the optimality of the bias-corrected procedure we need to be specific about the nature of the first-steps. We follow [2] and assume the first-steps $\eta_0 = (\eta_{01}', ..., \eta_{0J}')'$ are identified by the conditional moments $\mathbb{E}\left[\rho_j(Z,\eta_{0j}(X_j))|X_j\right] = 0$, for some functions ρ_j, $j = 1, ..., J$. This setting includes many example applications as special cases. Here, $Z = (Y', X')'$ and X is the union of distinct elements of X_j, $1 \leq j \leq J$. Suppose that there is $\gamma_{0j}(X_j)$ in the mean square closure of the set of derivatives $\partial\mathbb{E}[\rho_j(Z,\eta_{0j}(F_t))|X_j]/\partial t|_{t=0}$ as F_t varies over regular parametric

models such that

(3.1)
$$\left.\frac{\partial \mathbb{E}[g(Z,\theta_0,\eta_{0j}(F_t),\eta_{0,-j})]}{\partial t}\right|_{t=0} = -\mathbb{E}[\gamma_{0j}(X_j)\,\partial E[\rho_j(Z,\eta_{0j}(F_t))|X_j]/\partial t|_{t=0}],$$

where $\eta_{0,-j}$ includes all elements of η_0 but η_{0j}. Then, from [2] the adjustment term is given by $\phi(z,\theta,h_0) = -\sum_{j=1}^{J}\gamma_{0j}(X_j)\rho_j(Z,\eta_{0j}(X_j))$. The efficiency for the modified EL procedure with pathwise derivative ϕ is shown next.

THEOREM 3.4. *Let the conditions of Theorem 3.3 in this paper and Condition 1 in [2] hold. Then, the modified EL test ψ_n is AUMPI(α).*

4. Examples. This section illustrates the general theory above with several examples. In all the examples below we assume that the corresponding variance-covariance matrix Σ in (2.6) is finite and positive definite. For any random vectors U, V and W, the notation $U \perp V|W$ will be used to indicate that U is independent of V given W. Also, $f_{U|V}$ denotes the conditional Lebesgue density of U given V.

The following notation on smooth classes of functions is used throughout the examples. Let $\mathcal{C}^q(\mathcal{X})$ be a set of smooth continuous functions on $\mathcal{X}$ endowed with the sup-norm $\|\cdot\|_\infty$, as defined in [73], p.154. That is, if $\mathcal{X}$ is a convex, bounded subset of $\mathbb{R}^d$, with non-empty interior, then for any smooth function $h : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ and some $q > 0$, let $\underline{q}$ be the largest integer smaller than q, and

$$\|h\|_{\infty,q} := \max_{|a|_1 \le \underline{q}} \sup_{x\in\mathcal{X}} |\partial_x^a h(x)| + \max_{|a|_1=\underline{\eta}} \sup_{x\neq y} \frac{|\partial_x^a h(x) - \partial_x^a h(y)|}{|x-y|^{q-\underline{q}}},$$

where $|a|_1 = \sum_i a_i$ and $\partial_x^a = \frac{\partial_x^{|a|_1}}{\partial x_1^{a_1}\ldots\partial x_d^{a_d}}$. Further, let $C_M^q(\mathcal{X})$ be the set of all continuous functions $h : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ with $\|h\|_{\infty,q} \le M$. Let $\mathcal{C}_{M,\varepsilon}^q(\mathcal{X})$ be the set of functions $f \in \mathcal{C}_M^q(\mathcal{X})$ such that $f > \varepsilon$, for some $\varepsilon > 0$.

4.1. *Mean of interval censored data.* Suppose we observe $Z = (Y, X')'$, $X = (X_1, X_2')'$, X_1 is a positive random variable, X_2 is a d_2-dimensional vector of covariates and $Y = 1(W > X_1)$. The variable W is unobserved. We are interested in inference on $\theta_0 = \mathbb{E}[W]$. The random variables W and X_1 are conditionally independent given X_2, in short $W \perp X_1|X_2$, and the support of W is $\mathcal{S}_W = [0, M]$, $M \le \infty$. This is the so-called current status model; see [27, 35, 39, 68, 5] for surveys on this model. For applications in economics see [43]. Let $\eta_0(w, x_2) := \mathbb{P}(W > w|X_2 = x_2)$ denote the

conditional survival function and note that

$$\theta_0 = -\mathbb{E}\left[\int_0^M w d\eta_0\left(w, X_2\right)\right] = \mathbb{E}\left[\int_0^M \eta_0\left(w, X_2\right) dw\right].$$

Thus, we can write the previous equality as our estimating equation with $g\left(X_2, \theta_0, \eta_0\right) = \theta_0 - \int_0^M \eta_0\left(w, X_2\right) dw$. By the conditional independence assumption, $\eta_0(x) = \mathbb{E}\left[Y|X = x\right]$, provided the support of W is contained in the support of X_1. Therefore, any consistent nonparametric estimator for a conditional mean can be used as a first-step estimator for η_0, for example, a Nadaraya-Watson (NW) kernel estimator.

Applying the pathwise derivative computation suggested in [52], pg. 1361, we obtain $\phi\left(z, \theta_0, \eta(F_0)\right) = -(y - \eta_0(x)) f_{X_2}(x_2)/f_X(x)$. Hence, our method leads to the estimating equation

$$\mathbb{E}\left[\theta_0 - \int_0^M \eta_0\left(w, X_2\right) dw - (Y - \eta_0(X))\frac{f_{X_2}\left(X_2\right)}{f_X\left(X\right)}\right] = 0.$$

That is, in this example, $m\left(Z, \theta_0, h_0\right) = \theta_0 - \int_0^M \eta_0\left(w, X_2\right) dw - (Y - \eta_0(X)) f_{X_2}(X_2)/f_X(X)$, where $h_0 = (\eta_0, f_X) \in \mathcal{H} := \mathcal{C}_1^q\left(\mathcal{S}_X\right) \times \mathcal{C}_M^q\left(\mathcal{S}_X\right)$, $q > d_x/2$, $d_x = d_2 + 1$, and $\|h_0\|_{\mathcal{H}} = \|\eta_0\|_\infty + \|f_X\|_\infty$. The nuisance parameter h_0 is estimated by a NW estimator:

$$\widehat{\eta}\left(x\right) := \frac{n^{-1}\sum_{i=1}^n Y_i K_b\left(X_i - x\right)}{\widehat{f}_X\left(x\right)}, \qquad \widehat{f}_X\left(x\right) := n^{-1}\sum_{i=1}^n K_b\left(X_i - x\right),$$

where $x \in \mathcal{S}_X := \mathcal{S}_{X_1} \times \mathcal{S}_{X_2} \subset \mathbb{R}^{d_x}$, $K_b\left(x\right) := b^{-d_x}\prod_{l=1}^{d_x} k(x_l/b)$, for some univariate bounded kernel $k\left(\cdot\right)$ with compact support, and a bandwidth parameter $b \downarrow 0$. We verify our conditions under the following assumption:

Assumption E1:

(i) We observe $Z = \left(1\left(W > X_1\right), X_1, X_2'\right)'$, where $W \perp X_1 | X_2$ and $\mathcal{S}_W = [0, M] \subset \mathcal{S}_{X_1}$.

(ii) $f_X\left(x\right)$, $f_{X_2}(x_2)/f_X(x)$ and $\eta_0\left(x\right)$ are r times continuously differentiable in $x = (x_1, x_2)$, with uniformly bounded derivatives (including zero derivatives), where r is as in (iii) below. Moreover, $\inf_{x \in \mathcal{S}_X} f_X\left(x\right) > 0$, $\mathbb{E}[|f_{X_2}(X_2)/f_X(X)|^{2+\delta}] < \infty$, $h_0 \in \mathcal{H}^\delta$ and $\mathbb{P}(\widehat{h} \in \mathcal{H}^\delta) \to 1$, for some $\delta > 0$.

(iii) The kernel function $k : \mathbb{R} \to \mathbb{R}$ is bounded, symmetric, and satisfies the following conditions: $\int k\left(t\right) dt = 1$, $\int t^l k\left(t\right) dt = 0$ for $l = 1, \ldots, r-1$, and $\int \left|t^r k\left(t\right)\right| dt < \infty$ for some $r \geq 2$; and for some $v > 1$, $|k(t)| \leq C\left|t\right|^{-v}$ for $|t| > L$, $0 < L < \infty$.

(iv) The deterministic sequence of positive numbers $b \equiv b_n$ satisfies: (a) $b_n \to 0$ and $b_n^{2d_x} n / \log n \to \infty$; and (b) $n b_n^{4r} \to 0$.

Primitive conditions for $\mathbb{P}(\widehat{h} \in \mathcal{H}^\delta) \to 1$ have been given in [50, 24]. Note that undersmoothing is not required, that is, we require $nb_n^{4r} \to 0$ rather than the typical $nb_n^{2r} \to 0$. Assumption E1 is sufficient for Assumptions A,B and C, as the following Proposition shows.

Proposition E1. *Under Assumption E1, the conclusions of Theorem 3.1, Theorem 3.3 and Theorem 3.4 hold for this example.*

4.2. *Average treatment effect.* There is an extensive literature on the measurement and evaluation of treatment effects in observational studies. We use the potential outcome notation of [64]. Let D be the treatment indicator, Y_1 be the outcome under treatment and Y_0 be the outcome without treatment. We only observe $Z = (Y, D, X')'$, where $Y = Y_1 \cdot D + Y_0 \cdot (1 - D)$ and X is a d_x-dimensional vector of covariates. We assume the treatment is unconfounded, i.e. (Y_1, Y_0) is independent of D, conditional on X. One parameter of interest is the average treatment effect (ATE) $\theta_0 = \mathbb{E}[Y_1 - Y_0]$. Define the propensity score $\eta_0(X) := \mathbb{E}[D|X]$, which is assumed to be bounded away from zero and one. Then, it is known that under unconfoundedness the ATE is given by $\theta_0 = \mathbb{E}[YD/\eta_0(X) - Y(1-D)/\{1-\eta_0(X)\}]$. See [63]. This representation suggests the two-step estimator

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i D_i}{\widehat{\eta}(X_i)} - \frac{Y_i (1 - D_i)}{1 - \widehat{\eta}(X_i)} \right],$$

where $\widehat{\eta}$ is a consistent estimator of the propensity score. [31] derived the influence function for $\widehat{\theta}$ and provided sufficient conditions for the asymptotic normality of $\sqrt{n}(\widehat{\theta} - \theta_0)$ when $\widehat{\eta}$ is a series Logit estimator. In particular, they showed that, with $\mu_j(X) = \mathbb{E}[Y(j)|X]$ $(j = 0, 1)$ denoting the conditional mean for potential outcomes, the pathwise derivative due to the estimation of the propensity score η_0 is given by

$$\phi(Z, \theta_0, h_0) = (D - \eta_0(X)) \left(\frac{\mu_1(X)}{\eta_0(X)} + \frac{\mu_0(X)}{1 - \eta_0(X)} \right), \tag{4.1}$$

where $h_0 = (\eta_0, \mu_0, \mu_1) \in \mathcal{H} := \bar{\mathcal{C}}^q_{1,\varepsilon}(\mathcal{S}_X) \times \mathcal{C}^q_M(\mathcal{S}_X) \times \mathcal{C}^q_M(\mathcal{S}_X)$, and $\bar{\mathcal{C}}^q_{1,\varepsilon}(\mathcal{S}_X)$ is the subspace of functions $f \in \mathcal{C}^q_1(\mathcal{S}_X)$ such that $\varepsilon < f < 1 - \varepsilon$, for some ε, $0 < \varepsilon < 1$, and $\|h_0\|_{\mathcal{H}} = \|\eta_0\|_\infty + \|\mu_0\|_\infty + \|\mu_1\|_\infty$. The extra nuisance parameters μ_0 and μ_1 can also be estimated by suitable kernel estimators, after noticing that by unconfoundedness, $\mu_1(X) = \mathbb{E}[YD|X]/\eta_0(X)$ and

similarly $\mu_0(X) = \mathbb{E}[Y(1-D)|X]/(1-\eta_0(X))$. Therefore, our method suggests inference based on the modified estimating equation

$$\mathbb{E}\left[\theta_0 - \frac{YD}{\eta_0(X)} + \frac{Y(1-D)}{1-\eta_0(X)} + (D-\eta_0(X))\,\iota(X)\right] = 0,$$

where $\iota(x) := \mu_1(x)/\eta_0(x) + \mu_0(x)/[1-\eta_0(x)]$. We verify here our conditions for this example when $\widehat{h} = (\widehat{\eta}, \widehat{\mu}_0, \widehat{\mu}_1)$, where

$$\widehat{\eta}(x) := \frac{n^{-1}\sum_{i=1}^n D_i K_b(X_i - x)}{n^{-1}\sum_{i=1}^n K_b(X_i - x)},$$
$$\widehat{\mu}_1(x) := \frac{n^{-1}\sum_{i=1}^n Y_i D_i K_b(X_i - x)}{n^{-1}\sum_{i=1}^n D_i K_b(X_i - x)},$$
$$\widehat{\mu}_0(x) := \frac{n^{-1}\sum_{i=1}^n Y_i (1-D_i) K_b(X_i - x)}{n^{-1}\sum_{i=1}^n (1-D_i) K_b(X_i - x)}.$$

We require the following assumption.

Assumption E2:

(i) We observe $Z = (Y, D, X')'$, where $Y = Y_1 \cdot D + Y_0 \cdot (1-D)$ and $(Y_1, Y_0) \perp D|X$.

(ii) $f_X(x)$, $\iota(x)$ and $\eta_0(x)$ are r times continuously differentiable in x, with uniformly bounded derivatives (including zero derivatives), where r is as in E1(iii). Moreover, $\inf_{x\in\mathcal{S}_X} f_X(x) > 0$, $\mathbb{E}\left[|Y|^{2+\delta}\right] < \infty$, $\mathbb{E}\left[|\iota(X)|^{2+\delta}\right] < \infty$, $h_0 \in \mathcal{H}^\delta$ and $\mathbb{P}(\widehat{h} \in \mathcal{H}^\delta) \to 1$, for some $\delta > 0$.

Proposition E2. *Under Assumptions E1(iii-iv) and E2, the conclusions of Theorem 3.1, Theorem 3.3 and Theorem 3.4 hold for this example.*

4.3. *Estimating equations with missing data.* Consider inference based on the p estimating equations $\mathbb{E}[s(X, W, \theta_0)] = 0$, where X is a d_x-dimensional random vector that is always observed and W is a d_w-dimensional random vector that is only observed when $D = 1$ and not observed otherwise ($D = 0$). That is, the data we observe is a random sample of $Z = (X', W'D, D)'$. We assume missingness at random, i.e., W is independent of D, conditional on X. [74] proposed EL inference based on nonparametric imputation in this general setting. See also [14] for semiparametric efficiency calculations. The nonparametric imputation has an impact on the asymptotic distribution of the EL ratio test, and its limiting distribution is a weighted chi-squared, cf. [74]. Here, we apply our method to obtain a version of Wilks' Theorem in this general setting for missing data.

We modify the approach of [74] and consider the estimating equation

$$g\left(Z,\theta,\eta_0\right)=Ds\left(X,W,\theta\right)+\left(1-D\right)\frac{q_0\left(X,\theta\right)}{p_0\left(X\right)}, \tag{4.2}$$

where $\eta_0=(q_0',p_0)'$, $q_0\left(X,\theta\right):=E\left[Ds(X,W,\theta)|X\right]$ and $p_0\left(X\right):=E\left[D|X\right]$ are the nuisance parameters. This approach is slightly different from the one in [74, 70], who proposed a nonparametric imputation method by sampling from a smoothed nonparametric estimator of the distribution of W given X and $D=0$. Inference with this nonparametric imputation may be sensitive to the number of draws performed. Our approach overcomes this problem by imputing directly s and treating the imputation as a nuisance parameter in our semiparametric model. As shown in [74], our method is strictly more efficient than that based on imputing W with a finite number of draws, with the efficiency gap between these two procedures going to zero as the number of draws goes to infinity. Nevertheless, our main contribution in this example is not the nonparametric imputation of s, but rather obtaining distribution-free semiparametric EL inference without undersmoothing.
[74], Lemma 1, provided sufficient conditions under which (2.5) holds with

$$\phi\left(Z,\theta_0,h_0\right)=D\left(s\left(X,W,\theta_0\right)-\frac{q_0\left(X,\theta_0\right)}{p_0\left(X\right)}\right)\frac{1-p_0(X)}{p_0(X)}.$$

Therefore, our method suggests doing inference with the estimating moment

$$m\left(Z,\theta_0,h_0\right)=\frac{D}{p_0\left(X\right)}s\left(X,W,\theta_0\right)+\left(1-\frac{D}{p_0\left(X\right)}\right)\frac{q_0\left(X,\theta_0\right)}{p_0\left(X\right)}.$$

We propose to estimate $h_0=\eta_0=(q_0',p_0)'\in\mathcal{H}:=\mathcal{C}_M^q\left(\mathcal{S}_X\right)\times\cdots\times\mathcal{C}_M^q\left(\mathcal{S}_X\right)\times\mathcal{C}_{1,\varepsilon}^q\left(\mathcal{S}_X\right)$, by the NW kernel estimators

$$\begin{aligned}\widehat{q}\left(x,\theta\right) &:= \frac{1}{n}\sum_{i=1}^n\frac{D_is(X_i,W_i,\theta)K_b\left(X_i-x\right)}{n^{-1}\sum_{j=1}^nK_b\left(X_j-x\right)}\\ \widehat{p}\left(x\right) &:= \frac{1}{n}\sum_{i=1}^n\frac{D_iK_b\left(X_i-x\right)}{n^{-1}\sum_{j=1}^nK_b\left(X_j-x\right)}.\end{aligned} \tag{4.3}$$

The following assumption is sufficient for Theorem 3.1 in this example. Sufficient conditions for Theorem 3.3 and Theorem 3.4 to hold for this example can be straightforwardly established, but we do not consider them for the sake of space.

Assumption E3:

(i) We observe $Z := (X', W'D, D)'$ with $W \perp D|X$.
(ii) $f_X(x)$, $q_0(x,\theta)$ and $p_0(x)$ are r times continuously differentiable in x, with uniformly bounded derivatives (including zero derivatives), where r is as in E1(iii). Moreover, $\inf_{x\in\mathcal{S}_X} f_X(x) > 0$, $h_0 \in \mathcal{H}^\delta$ and $\mathbb{P}(\widehat{h} \in \mathcal{H}^\delta) \to 1$, for some $\delta > 0$.

Proposition E3. *Under Assumptions E1(iii-iv) and E3, the conclusion of Theorem 3.1 holds for this example.*

4.4. *Censored quantile regression.* Consider a censored quantile regression model $Q_{T|X}(\tau|X) = \inf\{t : \mathbb{P}(T \le t|X) \ge \tau\} = X'\theta_0$, where T is (a possible monotone transformation of) the survival time, X is a vector of covariates, and $X'\theta_0$ contains an intercept.

The data consist of $Z_i = (Y_i, X_i', \Delta_i)'$, which are i.i.d. copies of the vector $Z = (Y, X', \Delta)'$, where $Y = T \wedge C$ is the observed survival time, $\Delta = I(T \le C)$ is the censoring indicator, and C is the censoring time, which is assumed to be conditionally independent of T given X. As in [41] we take X one-dimensional, and we consider the estimating equation

$$g(Z, \theta_0, \eta_0) = X\left[\frac{I(Y - X'\theta_0 \ge 0)}{\eta_0(X'\theta_0|X)} - (1-\tau)\right],$$

where $\eta_0(\cdot|X) = \mathbb{P}(C > \cdot|X)$ is the unknown conditional survival function of the censoring variable C given X. The nuisance parameter η_0 is estimated by the conditional (local) Kaplan-Meier estimator ([6])

$$\widehat{\eta}(t|x) = \prod_{Y_i \le t, \Delta_i = 0} \Big(1 - \frac{W_i(x, b_n)}{\sum_{j=1}^n I(Y_j \ge Y_i) W_j(x, b_n)}\Big),$$

where $W_i(x, b_n) = k_b(X_i - x)/\sum_{j=1}^n k_b(X_j - x)$ is the standard Nadaraya Watson kernel, k is a one-dimensional density function, $k_b(\cdot) = k(\cdot/b)/b$ and $b \equiv b_n$ is a bandwidth. It follows from Theorem 3.2 in [23] that

$$\widehat{\eta}(t|x) - \eta_0(t|x) = -\frac{\eta_0(t|x)}{f_X(x)} \frac{1}{n} \sum_{i=1}^n k_b(X_i - x)\xi(Y_i, \Delta_i, t|x) + R_n(t|x), \tag{4.4}$$

where $\sup_x \sup_{t \le \tau_x} |R_n(t|x)| = O_{\mathbb{P}}((nb_n)^{-3/4}(\log n)^{3/4}) = o_{\mathbb{P}}(n^{-1/2})$ provided $nb_n^3(\log n)^{-3} \to \infty$, $\tau_x < \inf\{t : H(t|x) = 1\}$ and

$$\xi(y, \delta, t|x) = -\int_{-\infty}^{y\wedge t} \frac{dH_c(s|x)}{(1 - H(s|x))^2} + \frac{I(y \le t, \delta = 0)}{1 - H(y|x)},$$

with $H(t|x) = \mathbb{P}(Y \le t|X = x)$ and $H_c(t|x) = \mathbb{P}(Y \le t, \Delta = 0|X = x)$. We will assume that $\inf_x(1-H(x'\theta_0|x)) > 0$, and hence we can choose $\tau_x = x'\theta_0$.

Using the Hajek-projection for U-statistics with kernel depending on n (see e.g. Lemma 3.1 in [60]) it can be easily shown that

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\{g(Z_i,\theta_0,\widehat{\eta}) - g(Z_i,\theta_0,\eta_0)\} \\ & = (1-\tau)n^{-1}\sum_{i=1}^{n} X_i\xi(Y_i,\Delta_i,X_i'\theta_0|X_i) + o_{\mathbb{P}}(n^{-1/2}). \end{aligned}$$

This suggests that the pathwise derivative is given by $\phi(Z,\theta_0,h_0) = (1-\tau)X\xi(Y,\Delta,X'\theta_0|X)$, where $h_0(t|x) = (H(t|x), H_c(t|x), \eta_0(t|x))'$, or for general θ and $h = (h_1,h_2,h_3)'$,

$$\phi(Z,\theta,h) = (1-\tau)X\left[-\int_{-\infty}^{Y\wedge X'\theta}\frac{dh_2(s|X)}{(1-h_1(s|X))^2} + \frac{I(Y\le X'\theta,\Delta=0)}{1-h_1(Y|X)}\right],$$

and hence

$$\begin{aligned} m(Z,\theta,h) = X\Big[&\frac{I(Y-X'\theta\ge 0)}{h_3(X'\theta|X)} - (1-\tau) \\ &+ (1-\tau)\Big\{-\int_{-\infty}^{Y\wedge X'\theta}\frac{dh_2(s|X)}{(1-h_1(s|X))^2} + \frac{I(Y\le X'\theta,\Delta=0)}{1-h_1(Y|X)}\Big\}\Big]. \end{aligned}$$

The functions h_1,h_2,h_3 are supposed to belong to the space $\mathcal{G}$, defined by

$$\begin{aligned} \mathcal{G} \;=\; \{&g:\mathcal{S}_X\times\mathbb{R}\to[0,1] : g(x,\cdot)\in\mathcal{B}_M \text{ for all } x\in\mathcal{S}_X, \\ &\text{and } g(\cdot,t)\in\mathcal{C}_M^q(\mathcal{S}_{X,t}), \text{ for all } t\in\mathbb{R}\big\}, \end{aligned}$$

where $q \ge 1+\delta$ for some small $\delta > 0$, $\mathcal{B}_M = \{f : \mathbb{R}\to[0,1] : f$ has variation bounded by $M\}$, and $\mathcal{S}_{X,t} = \{x\in\mathcal{S}_X : t \le x'\theta_0\}$. Define $\mathcal{H} = \{(h_1,h_2,h_3)' : h_j\in\mathcal{G}, j=1,\dots,3\}$. We equip $\mathcal{H}$ with the semi-norm $\|h\|_{\mathcal{H}} = \sum_{j=1}^{3}\sup_{x\in\mathcal{S}_X}\sup_{t\le x'\theta_0}|h_j(t|x)|$ for $h = (h_1,h_2,h_3)'$. Finally, let

$$\widehat{H}(t|x) = \sum_{i=1}^{n} W_i(x,b_n)I(Y_i\le t), \quad \widehat{H}_c(t|x) = \sum_{i=1}^{n} W_i(x,b_n)I(Y_i\le t,\Delta_i=0).$$

The following assumption is sufficient for Theorem 3.1 in this example.

Assumption E4:

(i) We observe $Z = (Y,X',\Delta)'$, where $Y = T\wedge C$, $\Delta = I(T\le C)$, and $C \perp T|X$.

(ii) The distribution function F_X of X is three times continuously differentiable on the interior of $\mathcal{S}_X$, and $\inf_{x\in\mathcal{S}_X} f_X(x) > 0$.

(iii) The distribution functions $H(t|x)$ and $H_c(t|x)$ are continuous in (x,t), their first and second partial derivatives with respect to x exist, and they are continuous and uniformly bounded in (x,t). Moreover, $\inf_{x\in\mathcal{S}_X}(1-H(x'\theta_0|x)) > 0$, and there exist continuous and non-decreasing functions L_1, L_2 and L_3 with $L_j(-\infty) = 0$ and $L_j(\infty) < \infty$ $(j = 1,2,3)$, such that for all $x \in \mathcal{S}_X$ and for all $t_1, t_2 \in (-\infty,\infty)$,

$$\Big|H(t_1|x) - H(t_2|x)\Big| \le \Big|L_1(t_1) - L_1(t_2)\Big|$$
$$\Big|\frac{\partial}{\partial x}H(t_1|x) - \frac{\partial}{\partial x}H(t_2|x)\Big| \le \Big|L_2(t_1) - L_2(t_2)\Big|$$
$$\Big|\frac{\partial}{\partial x}H_c(t_1|x) - \frac{\partial}{\partial x}H_c(t_2|x)\Big| \le \Big|L_3(t_1) - L_3(t_2)\Big|.$$

(iv) The kernel function k is a symmetric probability density function with compact support, satisfying $\int t^l k(t)dt = 0$ for $l = 1,\dots,r-1$ and $\int |t^r k(t)|dt < \infty$ for some $r \ge 2$. Moreover, k is twice continuously differentiable.

(v) The deterministic sequence of positive numbers $b \equiv b_n$ satisfies $nb_n^{3+2\delta}(\log n)^{-1} \to \infty$ and $nb_n^5(\log n)^{-1} = O(1)$, where $\delta > 0$ is as in the definition of the class $\mathcal{G}$.

Proposition E4. *Under Assumption E4, the conclusion of Theorem 3.1 holds for this example.*

5. Monte Carlo Results. In this section we illustrate the finite sample properties of the proposed method using the average treatment effect (ATE) and the missing data examples.

5.1. *Average treatment effect.* We consider testing and constructing confidence intervals for the ATE parameter $\theta_0 = \mathbb{E}[Y_1 - Y_0]$, using the same design as that used by [36], where $Y_0 = 2X + \eta$, $Y_1 = Y_0 + \theta_0$, and $D = I(X\beta_0 + \varepsilon > 0)$ with both η and ε independent $N(0,1)$, and X is a $U[-1/2, 1/2]$ random variable. Notice that β_0 controls the range of the propensity score and it affects considerably the asymptotic variance of the ATE estimator. In the simulations we specify $\theta_0 \in \{-2, 0\}$, $\beta_0 \in \{1,2,3\}$, the sample sizes are $n = 100$ and $n = 300$, and $\eta_0(\cdot)$, $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are estimated with a leave-one-out kernel estimator with bandwidths b chosen as the design's theoretical optimal ones, see [36] for details[1]. The tables and figures below are

[1]We have also considered bandwidths chosen with least squares cross-validation. The results of the simulations are qualitatively very similar to those reported below, hence are

based on 1000 replications. The tables report the finite sample size (at the 5% and 10% significance level) of the test for the null hypothesis $H_0 : \theta = \theta_0$ using a Wald statistic based on the estimator of [31] (Wald), its bootstrapped version (Boot), the adjusted EL ratio (AEL), the modified EL ratio based on the pathwise derivative (4.1) (MEL) and modified EL ratio (MELN) based on the numerical approximation of $\phi(\cdot)$ using (B.8) in Appendix B in the Supplementary Material, which is given by $\hat{\phi}_i = \widehat{\delta}_i - n^{-1}\sum_{j=1}^{n}\widehat{\delta}_j$, where

$$\widehat{\delta}_i := \frac{-1}{nt}\sum_{j=1}^{n}\left[\frac{Y_jD_j}{\widehat{\eta}^b_{ti}(X_j)} - \frac{Y_j(1-D_j)}{1-\widehat{\eta}^b_{ti}(X_j)} - \frac{Y_jD_j}{\widehat{\eta}(X_j)} + \frac{Y_j(1-D_j)}{1-\widehat{\eta}(X_j)}\right], \tag{5.1}$$

$$\widehat{\eta}^b_{ti}(x) = \widehat{\eta}^b_{2ti}(x)/\widehat{\eta}^b_{1ti}(x),\ \widehat{\eta}^b_{1ti}(x) = \widehat{f}_X(x) + tK_b(x - X_i),$$

$$\widehat{\eta}^b_{2ti}(x) = \widehat{\eta}(x)\widehat{f}_X(x) + tD_iK_b(x - X_i).$$

The value of t used in these simulations for the numerical approximation is 0.08. Unreported results with other values of t show that inferences are not sensitive to t (we have experimented with several values of t between 0.01 and 0.3 and the obtained results are qualitatively the same). The bootstrap estimator is computed as in [45] using 500 replications and using the design's optimal bandwidths, whereas the adjusted EL ratio is based on the statistic $-2\widehat{\rho}\log EL_n(\theta_0, \widehat{h}) \xrightarrow{d} \chi^2_1$, with the estimated adjustment

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}\left(\frac{Y_iD_i}{\widehat{\eta}(X_i)} - \frac{Y_i(1-D_i)}{1-\widehat{\eta}(X_i)}\right)^2}{\sum_{i=1}^{n}\left(\frac{Y_iD_i}{\widehat{\eta}(X_i)} - \frac{Y_i(1-D_i)}{1-\widehat{\eta}(X_i)} - (D_i - \widehat{\eta}(X_i))\left(\frac{\widehat{\mu}_1(X_i)}{\widehat{\eta}(X_i)} + \frac{\widehat{\mu}_0(X_i)}{1-\widehat{\eta}(X_i)}\right)\right)^2}.$$

θ_0	β_0	Wald		Boot		AEL		MEL		MELN	
-2	1	0.091	0.134	0.060	0.112	0.088	0.124	0.059	0.115	0.061	0.117
-2	2	0.089	0.132	0.059	0.110	0.090	0.123	0.058	0.112	0.062	0.117
-2	3	0.093	0.132	0.061	0.110	0.090	0.120	0.058	0.113	0.060	0.115
0	1	0.083	0.121	0.058	0.108	0.085	0.121	0.057	0.110	0.058	0.112
0	2	0.081	0.119	0.057	0.108	0.086	0.122	0.058	0.109	0.058	0.110
0	3	0.082	0.120	0.057	0.109	0.087	0.122	0.057	0.109	0.058	0.111

TABLE 1

Finite sample size (5% left column, 10% right column) of the test for θ_0 in the ATE example for $n = 100$.

Tables 1-2 illustrate that the modified EL ratio based on the pathwise derivative results in a test statistic characterized by good finite sample

not reported.

θ_0	β_0	Wald		Boot		AEL		MEL		MELN	
-2	1	0.078	0.120	0.057	0.108	0.075	0.117	0.055	0.108	0.057	0.109
-2	2	0.077	0.119	0.055	0.107	0.074	0.115	0.054	0.107	0.055	0.110
-2	3	0.075	0.115	0.055	0.106	0.073	0.112	0.073	0.112	0.055	0.106
0	1	0.077	0.116	0.054	0.105	0.076	0.114	0.056	0.105	0.056	0.105
0	2	0.075	0.114	0.055	0.106	0.077	0.110	0.056	0.105	0.055	0.109
0	3	0.076	0.118	0.053	0.103	0.074	0.108	0.055	0.106	0.056	0.110

TABLE 2
Finite sample size (5% left column, 10% right column) of the test for θ_0 in the ATE example for $n = 300$.

properties, typically better than those based on the other competing test statistics. The tables also illustrate that the approximation to the pathwise derivative given in (5.1) yields also a test statistic with good finite sample properties. To further investigate this result we conduct some sensitivity analysis and compute the finite sample size for the five statistics using as bandwidths the values $kb/4$, $k = 1, 2, \ldots, 10$ for $n = 100$. Figure 1 is based on $\theta_0 \in \{-2, 0\}$, $\beta_0 \in \{1, 3\}$ and shows how both modified EL ratio based on the pathwise derivative (MEL) and on its numerical approximation (MELN) are clearly less sensitive to the choice of the bandwidth than the other competing statistics. To further support this result, Figure 2 reports the sensitivity to different bandwidths of the finite sample coverage and average length of the confidence intervals (at the 95% nominal level) for θ_0 and based on the Wald statistic (Wald), its bootstrapped version (Boot) and the same AEL, MEL and MELN statistics described above. The coverage of the confidence interval based on modified test is both more accurate and less sensitive to the bandwidth parameter, while having a shorter length than those based on alternative tests.

We also report power results. Figure 3 shows the size adjusted finite sample power of the tests based on the alternative hypotheses $H_\delta = \theta_0 + \delta$ with $\delta \in \{-1.5, -1.4, ..., -0.1, 0, 0.1, ..., 1.5\}$ for $\theta_0 = -2$ and $\beta_0 = 1$ and $n = 100$; those for the other values of θ_0, β_0 and $n = 300$ are similar and hence are not shown. The figure shows that both MEL and MELN have superior finite sample power compared to all the other competing statistics, which is consistent with our theoretical results in Theorems 3.3 and 3.4.

5.2. *Estimating equations with missing data.* We consider a logit model with missing covariates, similar to the model considered by [74]. The estimating equation is $s(X, W, \theta) = X(Y - \Lambda(X'\theta))$, where $X = (1, X_1, X_2)'$, $\theta_0 = (-1, 1, 2)'$, $\Lambda(\cdot)$ is the cumulative logistic distribution, X_1 and X_2 are, respectively, independent $N(0, 0.25)$ and $U(0, 3)$. In this case the variables

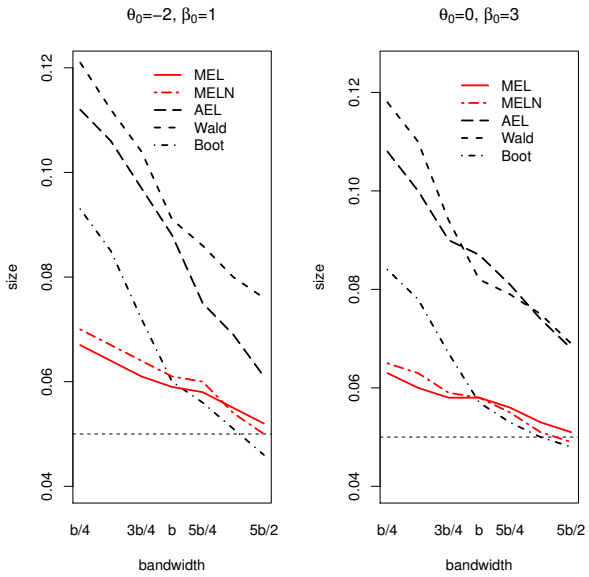

FIG 1. *Finite sample size for MEL (solid curve), MELN (two dashed curve),AEL (long dashed curve), Wald (dashed curve) and Boot (dot dashed curve) in the ATE example for* $n = 100$.

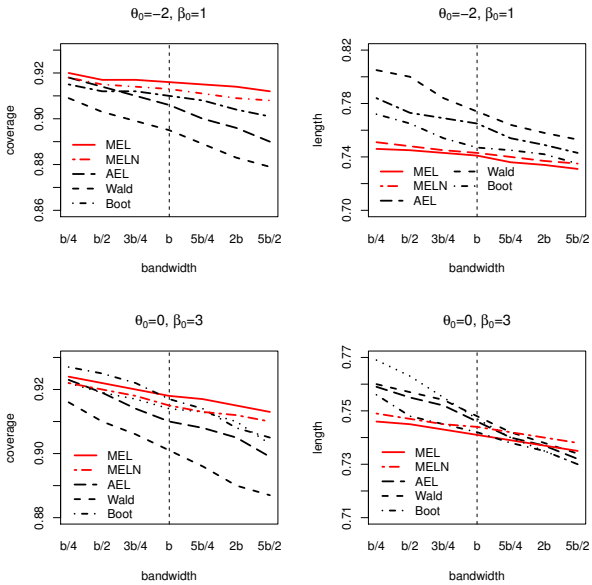

FIG 2. *Finite sample coverage at 95% (left) and average length (right) for MEL (solid curve), MELN (two dashed curve), AEL (long dashed curve), Wald (dashed curve), and Boot (dot dashed curve) in the ATE example for* $n = 100$.

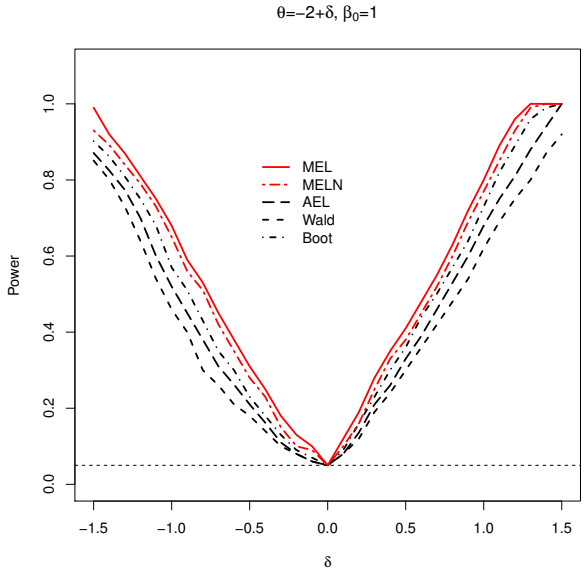

FIG 3. *Finite sample power for MEL (solid curve), MELN (two dashed curve), AEL (long dashed curve), Wald (dashed curve), and Boot (dot dashed curve) in the ATE example for* $n = 100$.

that are always observed are $X = (Y, X_1)'$, while the missing variable is $W = X_2$ with probability of missingness (the propensity score) given by $\text{logit}(\mathbb{P}(X_2 \text{ is missing})) = 0.5 - X_1 - 2Y$ (corresponding to approximately 30% of missing covariates). In the simulations the sample sizes are $n = 100$ and $n = 300$, and $q_0(\cdot)$ and $p_0(\cdot)$ are estimated with a leave-one-out kernel estimator with bandwidths chosen using least squares cross-validation. The statistics we consider are the adjusted EL ratio (AEL), a bootstrap version of it (AELboot), the modified EL ratio based on the pathwise derivative (MEL), a Wald statistic based on (4.2) (Wald), and the modified EL ratio based the analytical approximation (B.7) in Appendix B in the Supplementary Material (MELN) with $V(Z, \theta_0) = (Ds(X, W, \theta_0), D)$, and the analytical derivative

$$\widehat{\delta}_i = \frac{1}{n}\sum_{j=1}^{n}\left[\frac{1 - D_j}{\widehat{p}(X_j)}\left(S_i - \frac{\widehat{q}(X_j)}{\widehat{p}(X_j)}\right) D_i K_b(X_j - X_i)\right], \tag{5.2}$$

where $S_i = s(X_i, W_i, \theta_0)$.

The adjusted EL ratio is based on the feasible version of (4.2), namely $\widetilde{s}(Z, \theta) = Ds(X, W, \theta) + (1 - D)\widehat{q}(X, \theta)/\widehat{p}(X)$. In this case the estimated

adjustment is $\widehat{\rho} = tr(\widehat{\Sigma}^{-1}\widehat{Q})/tr(\widehat{V}^{-1}\widehat{Q})$, where

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{\sigma}^2(X_i)}{\widehat{p}(X_i)} + \widehat{q}\left(X_i,\widehat{\theta}\right)\widehat{q}\left(X_i,\widehat{\theta}\right)'\right),$$

$$\widehat{\sigma}^2(x) = \frac{1}{n}\frac{\sum_{i=1}^{n} D_i s(X_i,W_i,\widehat{\theta})s(X_i,W_i,\widehat{\theta})'K_b(X_i-x)}{n^{-1}\sum_{i=1}^{n} D_i K_b(X_i-x)} - \widehat{q}(x,\widehat{\theta})\widehat{q}(x,\widehat{\theta})'$$

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{s}(Z_i,\widehat{\theta})\widetilde{s}(Z_i,\widehat{\theta})', \quad \widehat{Q} = \frac{1}{n}\left(\sum_{i=1}^{n}\widetilde{s}(Z_i,\widehat{\theta})\right)\left(\sum_{i=1}^{n}\widetilde{s}(Z_i,\widehat{\theta})\right)',$$

and $\widehat{q}(x,\theta)$ and $\widehat{p}(x)$ are defined in (4.3). Then, it can be shown that $-2\widehat{\rho}\log EL_n(\theta_0,\widehat{h}) \xrightarrow{d} \chi^2_3$. The bootstrap version of the EL ratio follows the procedure suggested by [65] for imputed (survey) data: (1) for $D_i = 1$ a resample $\{Z_i^*\}_{i=1}^n$ from $\{Z_i\}_{i=1}^n$ and for $D_i = 0$ a resample $\{\widehat{q}^*(X_i^*,\theta)/\widehat{p}^*(X_i^*)\}_{i=1}^n$ from the imputed values $\{\widehat{q}(X_i,\theta)/\widehat{p}(X_i)\}_{i=1}^n$ are drawn to form the bootstrap analogue $\widetilde{s}^*(Z_i^*,\theta)$ of $\widetilde{s}(Z_i,\theta)$; (2) the bootstrap EL ratio statistic $EL_n^*(\theta_0,\widehat{h}^*)$ is computed using the centered version of $\widetilde{s}^*(Z_i^*,\theta_0)$; (3) steps (1)-(2) are repeated B times. The consistency of this bootstrap procedure follows by standard arguments (see for example those used by [74]). Finally, the Wald statistic is

$$W = n\left(\widehat{\theta}-\theta_0\right)'\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\widetilde{s}(Z_i,\widehat{\theta})}{\partial\theta'}\right)\widehat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\widetilde{s}(Z_i,\widehat{\theta})}{\partial\theta'}\right)'\left(\widehat{\theta}-\theta_0\right),$$

where $\widehat{\theta}$ is the maximum empirical likelihood estimator as defined in [74] (for exactly identified estimating equations).

The tables and figures below are based on 1000 replications. Tables 3 and 4 report, respectively, the finite sample size (at the 5% and 10% significance level) of the tests $H_0 : \theta_1 = \theta_{10}$ and $H_0 : \theta_2 = \theta_{20}$ and of the test for the joint hypothesis $H_0 : \theta_1 = \theta_{10}$, $\theta_2 = \theta_{20}$.

	$n = 100$				$n = 300$			
	θ_1		θ_2		θ_1		θ_2	
AEL	0.090	0.123	0.085	0.118	0.075	0.112	0.071	0.111
AELboot	0.059	0.109	0.058	0.107	0.055	0.103	0.056	0.102
MEL	0.057	0.108	0.057	0.106	0.054	0.103	0.054	0.102
MELN	0.058	0.109	0.059	0.108	0.055	0.105	0.055	0.104
Wald	0.104	0.148	0.105	0.135	0.087	0.129	0.080	0.115

TABLE 3

Finite sample size (5% left column, 10% right column) for marginal tests for θ_1 and θ_2 in the missing data example.

	$n = 100$		$n = 300$	
AEL	0.085	0.122	0.079	0.115
AELboot	0.060	0.115	0.057	0.106
MEL	0.056	0.108	0.052	0.103
MELN	0.059	0.110	0.055	0.107
Wald	0.106	0.140	0.092	0.119

TABLE 4

Finite sample size (5% left column, 10% right column) for joint test for (θ_1, θ_2) in the missing data example.

Figure 4 shows the sensitivity of the finite sample size of the tests $H_0 : \theta_1 = \theta_{10}$, $H_0 : \theta_2 = \theta_{20}$ and $H_0 : \theta_1 = \theta_{10}$, $\theta_2 = \theta_{20}$ to the bandwidth choice, using the following values: $b/4$, $b/2$, $3b/4$, $5b/4$, $2b$ where b is the cross-validated bandwidth. Figure 5 shows the sensitivity of the finite sample coverage and average length of the confidence intervals (at the 95% nominal level) for the unknown slopes θ_{10} and θ_{20} to the bandwidth choice using the same values as those used for the finite sample size.

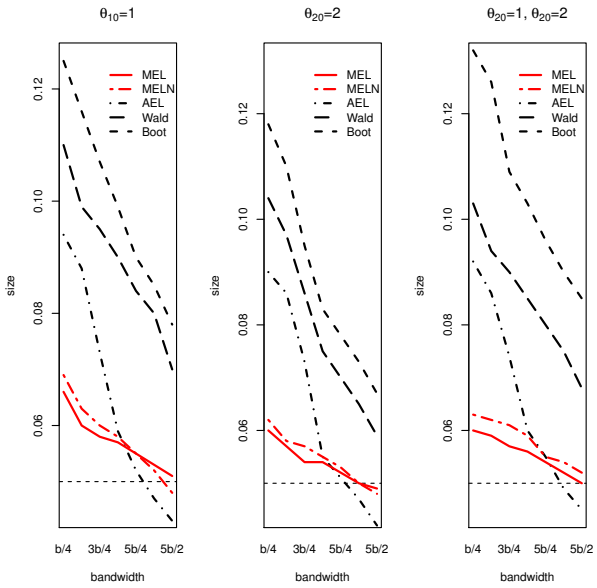

FIG 4. *Finite sample size for MEL (solid curve), MELN (two dashed curve), AEL (long dashed curve), Wald (dashed curve) and AELboot (dot dashed curve) in the missing data example for $n = 100$.*

Figure 6 shows the size adjusted finite sample power of the test based on the alternative hypotheses $H_\delta = \theta_{10} + \delta$ for $\delta \in \{-1, -0.9, ..., -0.1, 0, 0.1, ..., 1\}$ for $\theta_{10} = 1$ and $n = 100$ - those for the other values of θ_{10}, θ_{20} and $n = 300$ are similar and hence are not shown- and the contour plots of the size adjusted

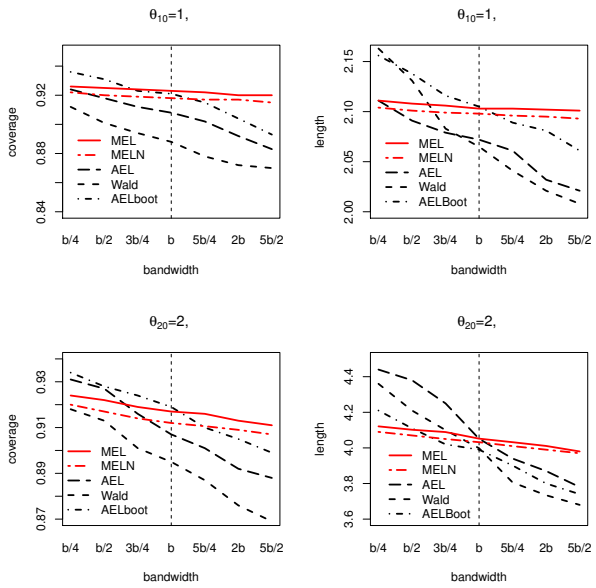

FIG 5. *Finite sample coverage at 95% (left) and average length (right) for MEL (solid curve), MELN (two dashed curve), AEL (long dashed curve), Wald (dashed curve) and AELboot (dot dashed curve) in the missing data example for* $n = 100$.

finite sample power curves for the test of $H_\delta = \theta_1 = \theta_{10}+\delta_1, \theta_2 = \theta_{20}+\delta_2$ over the grid $(\delta_1, \delta_2) \in \{-1, -0.75, ..., 0, ..., 0.75, 1\} \times \{-1, -0.75, ..., 0, ..., 0.75, 1\}$ at the contour level of 0.4. Smaller contour plots indicate higher finite sample power.

Tables 3-4 and Figures 4-6 confirm and strengthen the results of the ATE example, as they indicate that the modified EL proposed in this paper yields test statistics characterized by finite sample properties typically better than those based on other asymptotically equivalent test statistics. As with the ATE example, both modified EL ratios are clearly less sensitive to the bandwidth choice than the other competing statistics and more powerful, confirming the theoretical results of Theorems 3.3 and 3.4.

6. Conclusions. In this article we have presented a new way to conduct empirical likelihood two-step inference in semiparametric models. The new method is presented in a general setting, and its major advantage is that, although the estimation procedure is in two steps, Wilks' phenomenon is preserved. This is achieved by using as moment function in the estimating equation the (uniquely defined) influence function of the plug-in sample moment. It is also shown that the limit of this "modified" empirical likelihood is the same as in the case where the nuisance functions would be known.

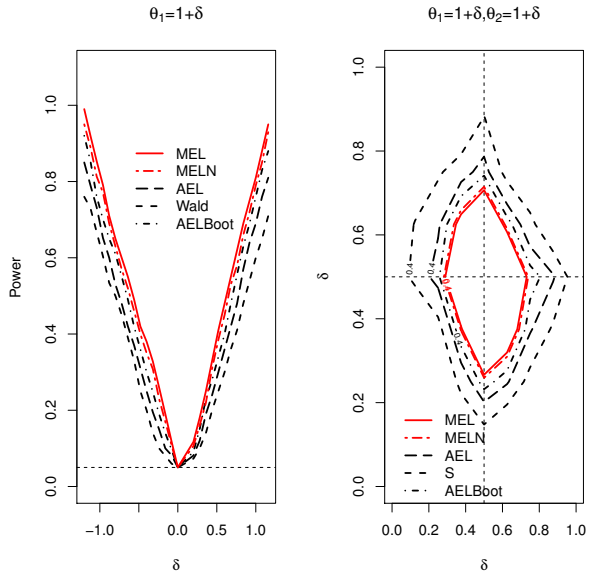

FIG 6. *Finite power (left panel) and finite power contour (right panel) for MEL (solid curve), MELN (two dashed curve), AEL (long dashed curve), Wald (dashed curve) and AELboot (dot dashed curve) in the missing data example for* $n = 100$.

Therefore, it is expected that the way the nuisance parameters are estimated (through e.g. the way a bandwidth parameter is chosen) does not have a major impact on the behavior of the modified empirical likelihood statistic. This might be particularly appealing in situations where first-steps are hard to estimate precisely (such as in high-dimensional settings). Additionally, the proposed modified EL test is efficient (in a Maximin and semiparametric sense). These theoretical results are confirmed by finite sample simulations, which further show that the new method performs favorably compared to competitors.

The ideas of this article can be extended to the problem of estimation of θ_0. An EL estimator based on the modified moments is expected to posses good bias properties, see [53] for linear functionals of densities. [19] have recently investigated the properties of related estimators in a generalized method of moments framework, allowing for machine learning methods as first-steps by virtue of the modified moment functions.

7. Proofs of the Main Results.

Proof of Theorem 3.1. We check the conditions of Theorem 2.1 in [32] (taking in their notation $a_n = 1$ and $m_n = m/\sqrt{n}$). (A0) and (A3) cor-

respond to our Assumption A(v). We check their condition (A1), which corresponds to (2.7). By Assumption A, and the standard Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_0,\widehat{h}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_0,h_0\right) + o_{\mathbb{P}}(1) \xrightarrow{d} N(0,\Sigma).$$

This verifies their assumption (A1) with $U \stackrel{d}{=} N(0,\Sigma)$, where $\stackrel{d}{=}$ stands for equality in distribution. Finally, their assumption (A2) (which corresponds to (2.8)) holds by our Assumption A(iv) and the consistency of $\widehat{h}$. □

Proof of Theorem 3.3. We follow a similar proof strategy as in Theorem 3.1**.** By Assumption A, under the local alternatives H_{1n},

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_0,\widehat{h}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_0,h_0\right) + o_{\mathbb{P}}(1).$$

By Assumption C, and with $\theta_n = \theta_0 + \tau/\sqrt{n}$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_0,h_0\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_n,h_0\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[m\left(Z_i,\theta_0,h_0\right) - m\left(Z_i,\theta_n,h_0\right)\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_n,h_0\right) + \sqrt{n}\mathbb{P}\left[m\left(Z_i,\theta_0,h_0\right) - m\left(Z_i,\theta_n,h_0\right)\right] + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(Z_i,\theta_n,h_0\right) - G_0\tau + o_{\mathbb{P}}(1).
\end{aligned}$$

Define $X_{in} = m\left(Z_i,\theta_n,h_0\right)$. We check the conditions of Lyapounov's Central Limit Theorem. Note $\{X_{in}\}_{i=1}^{n}$ are iid, with zero mean,

$$\lim_{n\to\infty}\mathbb{E}\left[X_{in}X_{in}'\right] = \mathbb{E}\left[m\left(Z,\theta_0,h_0\right)m'\left(Z,\theta_0,h_0\right)\right] < \infty,$$

and

$$\lim_{n\to\infty}\frac{\mathbb{E}\left[\left|X_{in}\right|^{2+\delta}\right]}{n^{\delta/2}} = 0,$$

by Assumption C(iii). This verifies A1 in [32] with $U \stackrel{d}{=} N(-G_0\tau,\Sigma)$. Thus, by Assumption A and Theorem 2.1 in [32], under the local alternatives H_{1n},

$$-2\log MEL_n(\theta_0,\widehat{h}) \xrightarrow{d} U'\Sigma^{-1}U \stackrel{d}{=} Z'Z,$$

where $Z \stackrel{d}{=} N(-\Sigma^{-1/2}G_0\tau, I)$. This allows us to apply existing Maximin theory, see [67]. □

Proof of Theorem 3.4. From the proof of Theorem 3.3**,** we obtain

$$-2\log MEL_n(\theta_0, \widehat{h}) = T_n' T_n + o_P(1),$$

where

$$T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} -\Sigma^{-1/2} m\left(Z_i, \theta_0, h_0\right).$$

By Corollary 3 in [20] the optimality will follow if we prove $T_n = \xi_n(h_0) + o_P(1)$, where $\xi_n(h_0) := (nB^*)^{-1/2} \sum_{i=1}^{n} S_\theta^*(Z_i, h_0)$, $S_\theta^*(Z_i, h_0)$ is the so-called efficient score and $B^* := Var(S_\theta^*)$ the efficient information, see [20] for details. By Lemma 1 in [2], see page 940 (A.33),

$$S_\theta^* = -G_0'\Sigma^{-1}m.$$

Hence, $B^* = G_0'\Sigma^{-1}G_0$ and $B^{*-1/2}S_\theta^* = -\left(G_0'\Sigma^{-1/2}\right)^{-1} G_0'\Sigma^{-1}m = -\Sigma^{-1/2}m$. Thus, $T_n = \xi_n(h_0) + o_P(1)$ and the optimality follows. □

Acknowledgements. We are grateful to the Co-Editor, the Associate Editor and referees for suggestions that have improved our paper.

SUPPLEMENT

Supplement to the paper: "Two-step semiparametric empirical likelihood inference". The supplement contains four appendices: Appendix A gathers all the proofs for the examples, Appendix B proves the validity of a general numerical algorithm for estimating the pathwise derivative, Appendix C extends the main result of the paper to the case of overidentified models, and Appendix D shows an auxiliary result regarding Donsker and Glivenko-Cantelli classes.

REFERENCES

[1] Ackerberg, D., X. Chen, and J. Hahn, (2012). A Practical Asymptotic Variance Estimator for Two-step Semiparametric Estimators. *Rev. Econ. and Stat.* 94, 481-498.

[2] Ackerberg, D., X. Chen, J. Hahn and Z. Liao (2014). Asymptotic Efficiency of Semiparametric Two-step GMM. *Rev. of Econ. Stud.* 81, 919-943.

[3] Akritas, M.G. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scand. J. Statist.*, **28**, 549-568.

[4] Andrews, D.W.K. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric. Theory*, **11**, 560-596.

[5] Banerjee, M. (2012). Current Status Data in the 21st Century: Some Interesting Developments. *Interval-Censored Time-to-Event Data: Methods and Applications*, 1-31, Chapman and Hall/CRC Biostatistics Series.

[6] Beran, R. (1981). *Nonparametric regression with randomly censored survival data.* Technical Report, Univ. California, Berkeley.

[7] Bertail P. (2006). Empirical likelihood in some semi-parametric models. *Bernoulli*, **12**, 299-331.

[8] Bickel, P.J., Klaassen, C.A., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore.

[9] Bickel,P.J., Y. Ritov, and T. M. Stoker (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Ann. Stat.,* **34**, 721-741.

[10] Bravo, F. (2004). Empirical likelihood based inference with applications to some econometric models. *Econometric. Theory,* **20**, 231-264.

[11] Bravo, F. (2018). *Second order asymptotics for nonparametric conditional moment restrictions*. Working paper.

[12] Bravo, F., Chu, B. and Jacho-Chavez, D.T. (2017). Semiparametric estimation of moment conditions models with weakly dependent data. *J. Nonpar. Statist.*, **29**, 108-136.

[13] Bravo, F., Escanciano, J.C. and Van Keilegom, I. (2018). Supplement to "Two-step semiparametric empirical likelihood inference".

[14] Chen, X.H., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.*, **36**, 808-843.

[15] Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591-1608.

[16] Chen, S.X. and Cui, H. (2006). On the Bartlett correctability of empirical likelihood in the presence of nuisance parameters. *Biometrika*, **93**, 215-220.

[17] Chen, S.X. and Cui, H. (2007). On the second order properties of empirical likelihood with moment restrictions. *J. Econometrics*, **141**, 492-516.

[18] Chen, S.X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *TEST*, **18**, 415-447.

[19] Chernozhukov, V., Escanciano, J.C., Newey, W.K., Ichimura, H., J. Robins (2017). Locally Robust Semiparametric Estimation. Working paper.

[20] Choi, S., Hall, W.J. and Schick, A. (1996). Asymptotically uniformly most powerful tests in parametric and semiparametric models. *Ann. Statist.*, **24**, 841-861.

[21] DiCiccio, T. and Romano, (1989). On adjustments based on the signed root of the empirical likelihood ratio statistics. *Biometrika*, **76**, 447-456.

[22] DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.*, **19**, 1053-1061.

[23] Du, Y. and Akritas, M.G. (2002). I.i.d representations of the conditional Kaplan-Meier process for arbitrary distributions. *Math. Methods Statist.*, **11**, 152-182.

[24] Escanciano, J.C., D. T. Jacho-Chavez and A. Lewbel (2014). Uniform convergence of weighted sums of non- and semi-parametric residuals for estimation and testing. *J. Econometrics*, **178**, 426-443.

[25] Fan, J. and Zhang, J. (2004). Sieve empirical likelihood tests for nonparametric functions. *Ann. Statist.*, **32**, 1858-1907.

[26] Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.*, **20,** 1306-1328.

[27] Groeneboom, P. and Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser, Basel.

[28] Hampel, F. (1968). *Contributions to the Theory of Robust Estimation*, PhD thesis, University of California, Berkeley.

[29] Hampel, F. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383-393.

[30] He, S., Liang, W., Shen, J. and Yang G. (2016). Empirical likelihood for right censored lifetime data. *J. Amer. Stat. Assoc.,* **111**, 646-655.

[31] Hirano, K., Imbens, G.W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica,* **71**, 1161-1189.

[32] Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.*, **37**, 1079-1111.

[33] Hong, H. Mahajan, A. and D. Nekipelov (2015): "Extremum estimation and numerical derivatives," *J. Econometrics* 188, 250-263.

[34] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.*, **24**, 540-568.

[35] Huang, J. and J. A. Wellner (1997). *Interval Censored Survival Data: A Review of Recent Progress*. Eds. D. Lin and T. Fleming. Springer-Verlag, New York.

[36] Ichimura, H. and Linton, O. (2005). Asymptotic expansions for some semiparametric program evaluation estimators. *Identification and Inference for Econometric Models*, edited by D. Andrews and J. Stock, Cambridge University Press, NY.

[37] Ichimura, H. and W. Newey (2017). The Influence Function of Semiparametric Estimators. CEMMAP working paper.

[38] Inglot, T. and Ledwina, T. (2006). Data-driven score tests for homoscedastic linear regression model : asymptotic results. *Probab. Math. Statist.*, **26**, 41-61.

[39] Jewell, N. P. and Van der Laan, M. (2003). Current status data: Review, recent developments and open problems. *Handbook of Statistics,* 5, 291-306.

[40] Koshevnik, Y. A., and Levit, B. Y. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Applic.*, **21**, 738-753.

[41] Leng, C. and Tong, X. (2013). A quantile regression estimator for censored data. *Bernoulli*, **19**, 344-361.

[42] Lepage, Y. (1973). A maximin test for means. *Stat. Neerlandica*, **27**, 1.

[43] Lewbel, A., O.B. Linton and D. McFadden (2011). Estimating features of a distribution from binomial data. *J. Econometrics.*, **162**, 170-188.

[44] Li. G., Lin, L. and Zhu, L. (2012). Empirical likelihood for a varying coefficient partially linear model with diverging number of parameters. *J. Mult. Anal.,* **105**, 85-111.

[45] Li, Q., Racine, J. and Wooldridge, J. (2008). Estimating average treatment effects with continuous and discrete covariates: the case of Swan-Ganz catherization. *Amer. Econ. Review: Papers Proc.*, **98**, 357-362.

[46] Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Comm. Statist.: Theory Meth.*, **40**, 2639-2660.

[47] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.*, **17**, 571-599.
[48] Matsushita, Y. and Otsu, T. (2016). Likelihood inference on semiparametric models with generated regressors. LSE working paper.
[49] Matsushita, Y. and Otsu, T. (2018). Likelihood inference on semiparametric models: average derivative and treatment effect. *Jap. Econ. Rev.*, **69**, 2.
[50] Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Mult.. Anal.*, **101**, 1067-1078.
[51] Newey, W.K. (1994a). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, **10**, 233-253.
[52] Newey, W.K. (1994b). The asymptotic variance of semiparametric estimators. *Econometrica*, **62**, 1349-1382.
[53] Newey, W.K., F. Hsieh, and J. Robins (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica,* **72**, 947-962.
[54] Newey, W.K. and Smith R.J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, **72**, 219-255.
[55] Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.
[56] Owen, A.B. (1990). Empirical likelihood ratio confidence Regions. *Ann. Statist.*, **18**, 90-120.
[57] Owen, A.B. (2001). *Empirical Likelihood*. Chapman and Hall, London.
[58] Pakes, A. and Pollard D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**, 1027-1057.
[59] Pfanzagl, J. (1982). *Contributions to a general statistical theory*. Lecture Notes in Statistics. 13. New York: Springer-Verlag.
[60] Powell, J.L., Stock, J.H. and Stoker, T.M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403-1430.
[61] Qin, J. and Lawless, J. (1994). Empirical Likelihood and General Estimating Equations. *Ann. Statist*, 22, 300-325.
[62] Rao, C.R. and S. K. Mitra (1971). textitGeneralized inverse of matrices and its applications. John Wiley and Sons, New York.
[63] Rosembaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.
[64] Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.*, **66**, 688-701.
[65] Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. *J. Amer. Statist. Assoc.*, **91**, 1278-1288.
[66] Stone, C.J. (1982). Optimal global rate of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040-1053.
[67] Strasser, H. (1985). *Mathematical Theory of Statistics*. De Gruyter, Berlin.
[68] Sun, J. (2006): *The Statistical Analysis of Interval-censored Failure Time Data*. Springer-Verlag, New York.
[69] Tang, X., Li, J. and Lian,H. (2013). Empirical likelihood for partially linear proportional hazards models with growing dimensions. *J. Mult. Anal.,* **121**, 22-32.
[70] Tang, C.Y. and Qin Y. (2012). An efficient empirical likelihood approach for estimating equations with missing data. *Biometrika*, **99**, 1001-1007.
[71] Van der Vaart, A.W. (1991). On differentiable functionals. *Ann. Stat.,* **19**, 178-204.
[72] Van der Vaart, A.W. (1998). *Asymptotics Statistics*. Cambridge University Press, Cambridge.

[73] Van der Vaart, A.W. and Wellner, J.A (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New-York.
[74] Wang, D. and Chen, S.X. (2009). Empirical likelihood for estimating equation with missing values. *Ann. Statist.*, **37**, 490-517.
[75] Wang, Q.-H. and Jing, B.-Y. (2003). Empirical likelihood for partial linear models. *Ann. Inst. Statist. Math.*, **55**, 585-595.
[76] Wang, Q., Linton, O. and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.*, **99**, 334-345.
[77] Wang, Q.H. and Rao, J.N.K. (2001). Empirical likelihood for linear regression models under imputation for missing responses. *Canad. J. Statist.*, **29**, 597-608.
[78] Wang, Q.H. and Rao, J.N.K. (2002a). Empirical likelihood-based inference in linear models with missing data. *Scand. J. Statist.*, **29**, 563-576.
[79] Wang, Q.H. and Rao, J.N.K. (2002b). Empirical likelihood-based inference under imputation for missing response data. *Ann. Stat.,* **30**, 896-924.
[80] Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **9**, 60-62.
[81] Wu, T.T., Li, G. and Tang, C. (2015). Empirical likelihood for censored linear regression and variable selection. *Scand. J. Stat.,* **42**, 798-812.
[82] Yang, G., Cui, X. and Hou, S. (2017). Empirical likelihood confidence regions in the single-index model with growing dimensions. *Commun. Stat.,* **46**, 7562-7579.
[83] Xue, L. (2009). Empirical likelihood confidence intervals for response mean with data missing at random. *Scand. J. Stat.,* **36**, 671-685.
[84] Xue, L. and Wang, Q. (2012). Empirical likelihood for single-index varying-coefficient models. *Bernoulli*, **18**, 836-856.
[85] Xue, L. and Xue, D. (2011). Empirical likelihood for semiparametric regression model with missing response data. *J. Multiv. Anal.,* **102**, 723-740.
[86] Xue, L.-G. and Zhu, L. (2006). Empirical likelihood for single-index models. *J. Multiv. Anal.*, **97**, 1295-1312.
[87] Xue, L.G. and Zhu, L. (2007). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika,* **94**, 921-937.
[88] Xue, L.G. and Zhu, L. (2012). Empirical likelihood in some nonparametric and semiparametric models. *Statist. Interf.*, **5**, 367-378.
[89] Zheng, M., Zhao, Z. and Yu, W. (2012). Empirical likelihood methods based on influence functions. *Statist. Interf.,* **5**, 355-366.
[90] Zhu, L., L. Lu, X. Cui and G. Li (2010). Bias-corrected empirical likelihood in a multi-link semiparametric model. *J. Mult. Anal.,* **101**, 850-868.
[91] Zhu, L. and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. Roy. Statist. Soc. - Ser. B*, **68**, 549-570.

Department of Economics
University of York
Heslington, York YO10 5DD
UK
E-mail: francesco.bravo@york.ac.uk

Department of Economics
Universidad Carlos III de Madrid
Calle Madrid 126, Getafe, 28902
Spain
E-mail: jescanci@eco.uc3m.es

ORSTAT
KU Leuven
Naamsestraat 69, 3000 Leuven
Belgium
E-mail: ingrid.vankeilegom@kuleuven.be