"Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference"

Bravo, Francesco ; Escanciano, Juan Carlos ; Van Keilegom, Ingrid

ABSTRACT

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, Wilks' phenomenon breaks down. In this paper we suggest 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, leads to distribution-free inference and it is less sensitive to the first-step estimator than alternative bootstrap methods. Several examples and a simulation study illustrate the generality of the procedure and its good finite sample performance.

CITE THIS VERSION


Le dépôt institutionnel DIAL est destiné au dépôt et à la diffusion de documents scientifiques émanant des membres de l'UCLouvain. Toute utilisation de ce document à des fins lucratives ou commerciales est strictement interdite. L'utilisateur s'engage à respecter les droits d'auteur liés à ce document, principalement le droit à l'intégrité de l'œuvre et le droit à la paternité. La politique complète de copyright est disponible sur la page Copyright policy

DIAL is an institutional repository for the deposit and dissemination of scientific documents from UCLouvain members. Usage of this document for profit or commercial purposes is strictly prohibited. User agrees to respect copyright about this document, mainly text integrity and source mention. Full content of copyright policy is available at Copyright policy

Available at: http://hdl.handle.net/2078.1/165159
DISCUSSION PAPER

2015/16

Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference

BRAVO, F., ESCANCIANO, J.C. AND I. VAN KEILEGOM
Wilks’ Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference

Francesco Bravo∗ Juan Carlos Escanciano†
University of York Indiana University

Ingrid Van Keilegom‡
Université catholique de Louvain

June 18, 2015

Abstract

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, Wilks’ phenomenon breaks down. In this paper we suggest 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, leads to distribution-free inference and it is less sensitive to the first-step estimator than alternative bootstrap methods. Several examples and a simulation study illustrate the generality of the procedure and its good finite sample performance.

Key Words: Empirical likelihood; Semiparametric inference; Stochastic equicontinuity; Wilks’ phenomenon.

AMS 2000 Subject Classification: Primary 62M10; secondary 62G10.

∗Department of Economics, University of York, Heslington, York YO10 5DD, UK. E-mail: francesco.bravo@york.ac.uk. Web Page: https://sites.google.com/a/york.ac.uk/francescobravo/.
†Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/.
‡Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, B1348 Louvain-la-Neuve, Belgium. E-mail: ingrid.vankeilegom@uclouvain.be. Web Page: http://perso.uclouvain.be/ingrid.vankeilegom/.
1 Introduction

Since its introduction as a nonparametric likelihood alternative to likelihood-type bootstrap methods for constructing confidence regions, Owen’s (1988, 1990, 2001) empirical likelihood (EL henceforth) method 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 (DiCiccio, Hall and Romano, 1991), they do not require estimation of scale (internal studentization) and skewness, and finally they are range preserving and transformation respecting. Furthermore, DiCiccio and Romano (1989) 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 $O(n^{-1})$. See Owen (2001) 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 Fan and Zhang (2004) 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 (Wilks, 1938) (i.e. the likelihood ratio statistic is asymptotically distribution-free and converges to a chi-squared distribution) holds for the EL.

In the case of semiparametric models EL has been considered by a number of authors including Wang and Jing (2003) for partially linear models, in Xue and Zhu (2006) for single-index models, Wang and Rao (2001, 2002a, 2002b), Wang, Linton and Härdle (2004) and Wang and Chen (2009) for various missing data problems, and Bertail (2006), Bravo, Chu and Jacho-Chavez (2013) and Bravo (2014) for general semiparametric moment conditions models, among many others. Chen and Van Keilegom (2009) and Xue and Zhu (2012) provide recent surveys on EL inference in the context of semiparametric regression models. For semiparametric models the standard approach uses a two-step (plug-in) procedure in which the first-step estimator replaces the infinite-dimensional parameter, while in the second step the plug in EL ratio is used to obtain inferences for the finite-dimensional parameter. Bertail (2006) has shown that plug-in based on the efficient score leads to the Wilks’ Theorem. However, computing the efficient score may be difficult in many semiparametric models.

More generally, the two-step semiparametric plug-in method typically does not yield asymptotically pivotal test statistics. Indeed, as shown in a general setting by Hjort, McKeague and Van Keilegom (2009), the asymptotic distribution of the resulting plug-in EL ratio is generally a weighted sum of chi-squared random variates with the weights depending (often in a complicated
way) on the distribution of the data. Thus, in these situations the Wilks’ phenomenon does
not hold for the EL ratio, so to obtain asymptotically valid EL inferences three main proposals
have been put forward in the literature. First, the bootstrap, as suggested for example by Wang
and Chen (2009) and Hjort, McKeague and Van Keilegom (2009), among many others. The
proposed bootstrap methods are general in nature, but they require re-estimating the semipara-
metric model in each bootstrap iteration, and thus are computationally expensive, in particular
for constructing confidence regions, since the bootstrap approximation of the critical value has
to be carried out for each candidate of the finite-dimensional parameter. Second, adjusting
the EL by a scale factor such that the adjusted (or rescaled) EL ratio is asymptotically pivotal.
Wang, Linton and Härdle (2004) proposed a specific scale factor; more general adjustments have
been proposed by Xue and Zhu (2006) and Bravo, Chu and Jacho-Chavez (2013). 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. Third, in some specific cases it is possible to correct
the original estimating equation in such a way that the effect of the first-step estimation is re-
moved. This approach has been called in the EL literature “bias-reduced or bias-corrected EL”,
see for example, Zhu and Xue (2006), Zhu, Lin, Cui and Li (2010), and Xue and Xue (2011).
As shown by these authors, this approach has the advantage of not requiring the bootstrap nor
undersmoothing, 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 can be built for other semiparametric models.

This leads us to the main objective of this article, which is to propose a theoretical foundation
for “bias-corrected EL” methods, thereby extending this approach to general semiparametric
models. The result is a general two-step method that can be used to obtain asymptotically
pivotal EL ratios in semiparametric models. 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 Koshevnik and Levit (1976), Pfanzagl (1982), Bickel, Klaassen, Ritov and Wellner
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. Moreover, the method that we propose is asymptotically first-order equivalent to
inference being carried out as if the infinite-dimensional parameter was known. Hence, we should
expect the confidence region to be less sensitive to the first-step estimator than alternative
procedures without the correction. In particular, we should expect less sensitivity to bandwidth or other regularization parameters, with a wider set of permissible bandwidths. In particular, our method generally allows for optimal rates of the first-step estimator. These theoretical results 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 and provides some heuristic explanation as to why the proposed method works, while Section 3 presents the main results. 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. An Appendix gathers all the proofs (Appendix A) and all the tables and figures (Appendix B).

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 $S_Z \subseteq \mathbb{R}^d_z$, and let $\{Z_i\}_{i=1}^n$ be independent copies of $Z$. Assume $Z$ satisfies the estimating equations

$$E[g(Z, \theta_0, \eta_0)] = 0,$$

where $g(\cdot) : S_Z \times \Theta \times \mathcal{E} \to \mathbb{R}^p$ 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 pseudo-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) represents, 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 $p$).

We assume we have at hand a suitable first-step consistent estimator for $\eta_0$, say $\hat{\eta}$. Under this setting, we aim to construct EL based confidence regions for $\theta_0$ 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

$$EL_n(\theta, \eta_0) = \max \left\{ \prod_{i=1}^n np_i : p_i > 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_ig(Z_i, \theta, \eta_0) = 0 \right\},$$
and $\chi^2_{p,\alpha}$ is the $\alpha$-quantile of the chi-squared distribution with $p$ degrees of freedom, $\alpha \in (0,1)$. In practice, $\eta_0$ is unknown and the standard two-step (plug in) approach defines confidence regions of the form $\{\theta \in \Theta : -2 \log EL_n(\theta, \hat{\eta}) < c\}$, for a suitable constant $c$ to be determined. Hjort, McKeague and Van Keilegom (2009) have investigated this method in a general setting, and have shown that if

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(Z_i, \theta_0, \hat{\eta}) \xrightarrow{d} U,$$  

and

$$\frac{1}{n} \sum_{i=1}^{n} g(Z_i, \theta_0, \hat{\eta}) g'(Z_i, \theta_0, \hat{\eta}) \xrightarrow{P} V,$$  

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

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

provided further regularity conditions hold. This convergence result is a generalization of the classical result by Owen (1988, 1990). The asymptotic distribution of the quadratic form $U'V^{-1}U$ is typically not chi-squared, but rather a weighted sum of chi-squared random variables. To explain the discrepancy between parametric and two-step settings, notice that a standard “functional” Taylor argument leads to the expansion

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(Z_i, \theta_0, \hat{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g(Z_i, \theta_0, \eta_0) + \phi(Z_i, \theta_0, h_0)\} + o_P(1),$$  

where $\phi(Z_i, \theta_0, h_0)$ is the so-called pathwise derivative of $\eta \to \mathbb{E}[g(Z_i, \theta_0, \eta)]$, which accounts for the asymptotic impact of the first-step estimate $\hat{\eta}$ on the sample analog of the moment $\mathbb{E}[g(Z_i, \theta_0, \eta)]$, and where $h_0$ may include $\eta_0$ and other nonparametric objects that may appear in the influence function as a result of “functional differentiation”. Hence, if (2.5) and certain moment conditions hold, an application of the standard Central Limit Theorem (CLT) yields $U \xrightarrow{d} N(0, \Sigma)$, where $\sim$ stands for equality in distribution and

$$\Sigma := \mathbb{E}[(g(Z, \theta_0, \eta_0) + \phi(Z, \theta_0, h_0))(g(Z, \theta_0, \eta_0) + \phi(Z, \theta_0, h_0))^\prime],$$

whereas a uniform law of large numbers (ULLN) yields

$$V = \mathbb{E}[g(Z, \theta_0, \eta_0)g'(Z, \theta_0, \eta_0)].$$

These results imply that the limiting distribution in (2.4) is in general a weighted chi-square distribution when $\phi \neq 0$.  

\footnote{The precise condition for $U'V^{-1}U$ being a central chi-squared distribution is that $\Sigma V^{-1}\Sigma V^{-1} = \Sigma V^{-1}\Sigma$, in which case the number of degrees of freedom is $v = tr(V^{-1}\Sigma)$; see Rao and Mitra (1971, p.171).}
2.2 A new method: heuristics

Let $m$ denote the augmented moment function (cf. (2.5))

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

and define the modified EL ratio function as

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

Let $\hat{h}$ be a consistent estimate of $h_0$ satisfying some conditions below. The main result of this article shows that under certain regularity conditions

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

forms an asymptotically valid $(1-\alpha)$-confidence region for $\theta_0$.

This statement follows from the general results of Hjort et al. (2009), after showing that

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

and

$$\frac{1}{n} \sum_{i=1}^{n} m\left(Z_i, \theta_0, \hat{h}\right) m'\left(Z_i, \theta_0, \hat{h}\right) \xrightarrow{p} \Sigma, \quad (2.8)$$

where $\Sigma$ is defined in (2.6). Under these conditions, Wilks’ phenomenon is restored, i.e.

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

We provide now some heuristic discussion on the validity of the method, 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 BKRW. This implies that, modulo some regularity conditions, the following invariance property holds

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m\left(Z_i, \theta_0, \hat{h}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Z_i, \theta_0, h_0) + o_P(1). \quad (2.9)$$

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

The next section presents the main result 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.
3 The Main Result

3.1 Notation

We first elaborate further on the model introduced in (2.1). To simplify the notation, we write the moment as \( g(Z, \theta_0, h_0) \), even though \( g \) only depends on \( \eta_0 \). Notice that, though we do not make it explicit in (2.1), the nuisance function \( h_0(\cdot) \) may contain \( \theta_0 \) as an additional argument. In what follows, we suppress \( \theta_0 \) 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 \( \hat{h}(\cdot) \) for \( h_0(\cdot) \) is available with certain convergence properties as specified in Assumption A below. Throughout we use the following notation. Let \( |\cdot| \) denote the Euclidean norm, i.e. \( |A| := (\text{tr}(A'A))^{1/2} \), where \( \text{tr}(A) \) is the trace of the matrix \( A \). For a measurable function \( g \) of \( Z \), define \( \|g\|_\infty := \sup_{z \in S_Z} |g(z)| \) and \( \|g\|_r := (\mathbb{E}[|g(Z)|^r])^{1/r} \), where \( S_Z \) is the support of \( Z \). The function space \( H \), where \( h_0 \) belongs to, is endowed with a pseudo-metric \( \|\cdot\|_H \). Since we assume consistency of \( \hat{h} \) with respect to \( \|\cdot\|_H \), we can redefine \( H \) as \( H_\delta := \{ h \in H : \|h - h_0\|_H \leq \delta \} \), for an arbitrarily small \( \delta > 0 \).

For a measurable function \( f \) we denote \( P_f := \int f dP \), \( P_n f := \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \) and \( G_nf := \sqrt{n}(P_nf - Pf) \).

Henceforth, we will use the concepts of \( \mathbb{P}\)-Glivenko-Cantelli and \( \mathbb{P}\)-Donsker classes; see van der Vaart and Wellner (1996) for definitions.

3.2 Obtaining pathwise derivatives

Consider the semiparametric model \( \mathcal{P}_0 = \{ \mathbb{P}_{\theta_0,h} : h \in \mathcal{H}, \mathbb{P}_{\theta_0,h} \text{ satisfies (2.1), i.e. } \mathbb{E}_h(g(Z, \theta_0, h)) = 0 \} \), where \( \mathbb{E}_h \) denotes the expectation operator with respect to \( \mathbb{P}_{\theta_0,h} \), and with \( \theta_0 \in \Theta \subset \mathbb{R}^p \) known, and \( \mathcal{H} \) a subset of a Banach space. To simplify notation, we sometimes drop \( h_0 \) and denote \( \mathbb{E} \equiv \mathbb{E}_{h_0} \). Define the functional \( \psi : \mathcal{P}_0 \rightarrow \mathbb{R}^p \) by

\[
\psi(\mathbb{P}_{\theta_0,h}) := \mathbb{E}_h[g(Z, \theta_0, h)].
\]  

Consider then a parametric (one-dimensional) submodel \( t \in (0, \varepsilon) \rightarrow \mathbb{P}_{\theta_0,t} \in \mathcal{P}_0 \), satisfying the classical mean-squared differentiability assumption with score \( s_h \), i.e.

\[
\int \left[ d\mathbb{P}_{\theta_0,t}^{1/2} - d\mathbb{P}_{\theta_0}^{1/2} - \frac{1}{2} d\mathbb{P}_{\theta_0}^{1/2} s_h \right]^2 = o(t^2).
\]

The function \( s_h \) has the interpretation of the score function for \( h \) when \( \theta_0 \) is fixed. Denote by \( \mathcal{T}_h(\mathbb{P}_0) \) the linear span of all \( s_h \) thus obtained for different parametric submodels. This is the so-called tangent space of nuisance parameters, see BKRW. Assume the functional \( \psi(\cdot) \) in (3.1)
is differentiable at $\mathbb{P}_{\theta_0, h_0}$ in the sense of van der Vaart (1991). Then, under these conditions, since $\mathbb{E}_{h_t} [g(Z, \theta_0, h_t)] = 0$, by the chain rule

$$0 = \frac{\partial \psi(\mathbb{P}_{\theta_0, h_t})}{\partial t} \bigg|_{t=0} = \frac{\partial \mathbb{E}_{h_t} [g(Z, \theta_0, h_0)]}{\partial t} \bigg|_{t=0} + \frac{\partial \mathbb{E}_{h_0} [g(Z, \theta_0, h_t)]}{\partial t} \bigg|_{t=0} = \mathbb{E} [g(Z, \theta_0, h_0) s_h(Z)] + \mathbb{E} [\phi(Z, \theta_0, h_0) s_h(Z)] = \mathbb{E} [m(Z, \theta_0, h_0) s_h(Z)].$$

That is, the influence function $m(\cdot, \theta_0, h_0)$ is orthogonal to the tangent space of nuisance parameters $T_h(\mathbb{P}_0)$. The function $\phi$ can be obtained by the Riesz representation and the continuity of the derivative of $\gamma(t) := \mathbb{E}_{h_0} [g(Z, \theta_0, h_t)]$ at $t = 0$. That is, by differentiability of the functional $\psi(\mathbb{P}_{\theta_0, h_t})$, the functional $\phi(\mathbb{P}_{\theta_0, h}) = \mathbb{E}_{h_0} [g(Z, \theta_0, h)]$ is also differentiable, and hence there is a continuous linear map $\hat{\phi}_\mathbb{P} : T_h(\mathbb{P}_0) \subset L_2(\mathbb{P}_0) \rightarrow \mathbb{R}^p$ and $\phi \in L_2(\mathbb{P}_0)$ such that $\mathbb{E} [\phi(Z, \theta_0, h_0)] = 0$ and for all $s_h \in T_h(\mathbb{P}_0)$,

$$\frac{\partial \gamma(t)}{\partial t} \bigg|_{t=0} = \hat{\phi}_\mathbb{P} s_h = \mathbb{E} [\phi(Z, \theta_0, h_0) s_h(Z)].$$

Note that $\phi$ is not uniquely determined by the restrictions above. Let $\Pi_h$ denote the orthogonal projection onto the closure of $T_h(\mathbb{P}_0)$. If we take $\phi(Z, \theta_0, h_0) = -\Pi_h g(Z, \theta_0, h_0)$, then (3.2) holds, but there are many other possibilities, corresponding to non-orthogonal projections onto the orthocomplement of $T_h(\mathbb{P}_0)$. The set of permitted $\phi$'s in our method are those that satisfy (3.3) for all $s_h \in T_h(\mathbb{P}_0)$. For a comprehensive discussion on semiparametric efficiency theory and the underlying geometry see, e.g., Pfanzagl (1982) and BKRW. In particular, BKRW contains numerous examples where $\phi$ is computed. There is also an extensive literature in semiparametric inference dealing with the computation of $\phi$ for many semiparametric models of interest (see e.g. Newey (1994)).

Our method takes $\phi$ (hence $m$) as a primitive, computed by the arguments or any of the cited references above, and proceeds to check the regularity conditions given in the next section.

### 3.3 Regularity conditions and main result

We introduce the following regularity conditions.

**Assumption A:**

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

$$\sup_{\|h-h_0\|_H \leq \delta_n} \|G_n m(\cdot, \theta_0, h) - G_n m(\cdot, \theta_0, h_0)\| = o_P(1).$$
(ii) Asymptotic “no bias” condition:
\[ \mathbb{P} m \left( Z, \theta_0, \hat{h} \right) = o_{\mathbb{P}}(n^{-1/2}). \]

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

(iv) Uniform consistency: for all \( \delta_n \downarrow 0 \) and \( \nu(Z, \theta_0, h) := m(Z, \theta_0, h) m'(Z, \theta_0, h), \)
\[ \sup_{\| h-h_0 \|_{\mathcal{H}} \leq \delta_n} | \mathbb{P}_n \nu(\cdot, \theta_0, h) - \mathbb{P}_n \nu(\cdot, \theta_0, h_0) | = o_{\mathbb{P}}(1). \]
Moreover, the matrix \( \Sigma = \mathbb{E} [m(Z, \theta_0, h_0) m'(Z, \theta_0, h_0)] \) is positive definite and finite.

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

Assumption A is a high-level condition that suffices for our method to work. The conditions in A(i-ii) are mild assumptions that impose “smoothness” on the semiparametric model; see, e.g., Chen, Linton and Van Keilegom (2003) for related assumptions. Assumption A(i) is implied by the \( \mathbb{P}−\text{Donsker} \) property of the function class \( \mathcal{F} := \{m(\cdot, \theta_0, h) : h \in \mathcal{H}^\delta \}. \) Related high-level assumptions to the asymptotic “no bias” condition have been considered extensively in the literature, see, for example, BKRW (p. 396), Theorem 6.1(i) in Huang (1996, p. 557), Section 25.8 in van der Vaart (1998), Assumption H2 in Bertail (2006), or Condition M2 in Bickel, Ritov and Stoker (2006); all for the case \( \phi(Z, \theta_0, h_0) = -\Pi_h g(Z, \theta_0, h_0). \) Assumption A(iii) is standard in the literature on semiparametric inference. Assumption A(iv) is implied by the \( \mathbb{P}−\text{Glivenko-Cantelli} \) property of the function class \( \mathcal{M}^2 := \{m(\cdot, \theta_0, h) m'(\cdot, \theta_0, h) : h \in \mathcal{H}^\delta \}. \) Assumption A(v) is required in Hjort, McKeage and Van Keilegom (2009), who discussed sufficient conditions for it to hold.

**Theorem 3.1** If Assumption A holds, then
\[ -2 \log MEL_n(\theta_0, \hat{h}) \xrightarrow{d} \chi_p^2. \]

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:

(i) The map \( h \rightarrow M(h) = \mathbb{E} [m(Z, \theta_0, h)] \) from \( \mathcal{H}^\delta \) to \( \mathbb{R}^p \) is Hadamard differentiable at \( h = h_0 \) with derivative \( V_h(h_0)[h - h_0] \) in all directions \( [h - h_0] \in \mathcal{H}^\delta \); and for all \( h \in \mathcal{H}^\delta \) with \( \delta_n \downarrow 0 \), it holds that

\[
|M(h) - M(h_0) - V_h(h_0)[h - h_0]| \leq c \|h - h_0\|^2_H
\]

for a constant \( c \geq 0 \).

(ii) \( \mathbb{P}(\hat{h} \in \mathcal{H}^\delta) \rightarrow 1 \), and \( \|\hat{h} - h_0\|_H = o_{\mathbb{P}}(n^{-1/4}) \).

(iii) The derivative \( V_h(h_0) \) is zero.

Assumption B(i) requires additional smoothness in the model. This condition holds if \( M(h) \) is twice Hadamard differentiable with bounded second derivative. Assumption B(ii) strengthens the consistency of A(iii). The zero derivative of B(iii) is consistent with (3.2). The proof of the following Lemma is trivial, and hence omitted.

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

### 4 Examples

This section illustrates the general theory above with several examples. Henceforth, we use the following notations and assumptions. We assume that the data \( \{Z_i\}_{i=1}^n \) is a sample of independent and identically distributed (i.i.d) observations. For a generic random vector \( Z \) we denote by \( f_Z \) its (Lebesgue) density with support \( S_Z \). Similarly, \( f_{Y|X} \) denotes the conditional density of \( Y \) given \( X \). We assume \( S_X \) is compact, convex and has non-empty interior. In all the examples below we implicitly assume that the corresponding variance-covariance matrix \( \Sigma \) in (2.6) is finite and positive definite. Finally, 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 \( W \).

#### 4.1 Mean of interval censored data

Suppose we observe \( Z = (Y, W', V)' \), where \( W \) is a \( d_w \)-dimensional vector of covariates and \( Y = 1(U > V) \). The unobservable continuous variable \( U \) satisfies \( \mathbb{E}[|U|] < \infty \). We are interested in inference on \( \theta_0 = \mathbb{E}[U] \). The random variables \( U \) and \( V \) are conditionally independent given \( W \), in short \( U \perp V|W \), and the support of \( U \) is \([0, M]\), \( M \leq \infty \). This is the so-called current status model extensively investigated in the literature; see Groeneboom and Wellner (1992), Huang and Wellner (1997), Jewell and van der Laan (2003), Sun (2006) and Banerjee (2012).
for surveys of this extensive literature. For applications in economics see Lewbel, Linton and McFadden (2010).

Let \( \eta_0(u, w) := \mathbb{P}(U > u| W = w) \) denote the conditional survival function and note that
\[
\theta_0 = -\mathbb{E} \left[ \int_0^M u d\eta_0(u, W) \right] = \mathbb{E} \left[ \int_0^M \eta_0(u, W) du \right].
\]

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

To apply our method we need to compute the pathwise derivative \( \phi \). Newey (1994, p. 1361) has shown in a more general context that this is given by
\[
\phi(Z, \theta_0, h_0) = -(Y - \eta_0(X)) \frac{f_W(W)}{f_X(X)},
\]
where \( h_0 = (\eta_0, f_X, f_W) \) and \( X := (V, W')' \). Hence, we propose using the estimating equation
\[
\mathbb{E} \left[ \theta_0 - \int_0^M \eta_0(u, W) du - (Y - \eta_0(X)) \frac{f_W(W)}{f_X(X)} \right] = 0.
\]
That is, in this example,
\[
m(Z, \theta_0, h_0) = \theta_0 - \int_0^M \eta_0(u, W) du - (Y - \eta_0(X)) \frac{f_W(W)}{f_X(X)},
\]
where \( h_0 = (\eta_0, f_X, f_W) \in \mathcal{H} := C^q_1(\mathcal{S}_X) \times C^q_2(\mathcal{S}_X) \times C^q_3(\mathcal{S}_W), C^q_1(\mathcal{X}) \) is a set of smooth continuous functions on \( \mathcal{X} \) endowed with the sup-norm \( \| \cdot \|_\infty \), as defined in van der Vaart and Wellner (1996, p. 154) with \( q > d_x/2 \), \( d_x = d_w + 1 \), \( C^q_{M,\varepsilon}(\mathcal{S}_X) \) is the subspace of functions \( f \in C^q_M(\mathcal{S}_X) \) such that \( f > \varepsilon \), for some \( \varepsilon > 0 \), and \( \| h_0 \|_\mathcal{H} = \| \eta_0 \|_\infty + \| f_X \|_\infty + \| f_W \|_\infty \).²

The nuisance parameter \( h_0 \) is estimated by a NW estimator:
\[
\hat{\eta}(x) := \frac{n^{-1} \sum_{i=1}^n Y_i K_b(X_i - x)}{f_X(x)}, \quad \hat{f}_X(x) := n^{-1} \sum_{i=1}^n K_b(X_i - x)
\]

²Specifically, if \( \mathcal{X} \) is a convex, bounded subset of \( \mathbb{R}^d \), with non-empty interior, then for any smooth function \( h : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathbb{R} \) and some \( \eta > 0 \), let \( \eta \) be the largest integer smaller than \( \eta \), and
\[
\| h \|_{\infty, \eta} := \max_{|a|_{\leq 2} \in \mathcal{X}} \sup_{x: |x| < \eta} | \frac{\partial^{|a|} h(x)}{\partial x_1 \ldots \partial x_d} | + \max_{|a|_{\neq 2} \in \mathcal{X}} \frac{\partial^{|a|} h(x) - \partial^{|a|} h(y)}{|x - y|^{\eta - |a|}},
\]
where \( |a| = \sum_i a_i \) and \( \partial^{|a|} = \frac{\partial^{|a|}}{\partial x_1 \ldots \partial x_d} \). Further, let \( C^q_M(\mathcal{X}) \) be the set of all continuous functions \( h : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathbb{R} \) with \( \| h \|_{\infty, \eta} \leq M \).
and

\[ \hat{f}_W (w) := n^{-1} \sum_{i=1}^{n} K_b (W_i - w), \]

where, for a generic vector \( x := (x_1, \ldots, x_{d_x})' \), \( K_b (x) := b^{-d_x} \prod_{i=1}^{d_x} k(x_i/b) \) for some univariate kernel function \( k (\cdot) \), \( b \) is a bandwidth parameter, \( X_i := (V_i, W_i')' \) and \( x := (v, w')' \in S_X := \mathcal{S}_V \times \mathcal{S}_W \subset \mathbb{R}^{d_x} \).

We verify our conditions under the following assumption:

**Assumption E1:**

(i) We observe \( Z = (1 (U > V), W', V)' \), where \( U \perp V|W \) and \( S_U = [0, M] \subset \mathcal{S}_V \).

(ii) \( f_X (x), f_W (w)/f_X (x) \) and \( \eta_0 (x) \) are \( r \) times continuously differentiable in \( x \) and \( w \), with uniformly bounded derivatives (including zero derivatives), where \( r \) is as in (iii) below. Moreover, \( h_0 \in \mathcal{H}^b \) and \( \mathbb{P} (\hat{h} \in \mathcal{H}^b ) \to 1 \).

(iii) The kernel function \( k : \mathbb{R} \to \mathbb{R} \) is bounded, symmetric, and satisfies the following conditions: \( \int k (t) \, dt = 1, \int t^l k (t) \, dt = 0 \) for \( l = 1, \ldots, r - 1, \) and \( \int |t^r k (t)| \, dt < \infty \) for some \( r \geq 2 \); and for some \( v > 1, |k(t)| \leq C |t|^{-v} \) 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) \( nb_n^{4r} \to 0 \).

Primitive conditions for \( \mathbb{P} (\hat{h} \in \mathcal{H}^b ) \to 1 \) have been given in Neumeyer and Van Keilegom (2010) and Escanciano, Jacho-Chavez and Lewbel (2014). It is important to note that in E1(iv-b) 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 our high-level assumptions A-B, as the following Proposition shows.

**Proposition E1.** Under Assumption E1, the conclusion of Theorem 3.1 holds 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 Rubin (1974). 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

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

See Rosenbaum and Rubin (1983). This representation suggests the estimator

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

where \(\hat{\eta}\) is a consistent estimator of the propensity score. Hirano, Imbens and Ridder (2003) derived the influence function for \(\hat{\theta}\) and provided sufficient conditions for the asymptotic normality of \(\sqrt{n}(\hat{\theta} - \theta_0)\) when \(\hat{\eta}\) 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 \(\eta_0\) is given by

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

where \(h_0 = (\eta_0, \mu_0, \mu_1) \in \mathcal{H} := C_{1,\varepsilon}^q(S_X) \times C_{M}^q(S_X) \times C_{M}^q(S_X)\), and \(C_{1,\varepsilon}^q(S_X)\) is the subspace of functions \(f \in C_{M}^q(S_X)\) such that \(\varepsilon < f < 1 - \varepsilon\), for some \(\varepsilon, 0 < \varepsilon < 1\), and \(\|h_0\|_\mathcal{H} = \|\eta_0\|_\infty + \|\mu_0\|_\infty + \|\mu_1\|_\infty\). The extra nuisance parameters \(\mu_0\) and \(\mu_1\) 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

$$\iota(x) := \frac{\mu_1(x)}{\eta_0(x)} + \frac{\mu_0(x)}{1-\eta_0(x)}.$$ 

We verify here our conditions for this example when \(\hat{h} = (\hat{\eta}, \hat{\mu}_0, \hat{\mu}_1)\), where

$$\hat{\eta}(x) := \frac{n^{-1}\sum_{i=1}^{n} D_i K_b(X_i - x)}{n^{-1}\sum_{i=1}^{n} K_b(X_i - x)},$$

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

$$\hat{\mu}_0(x) := \frac{n^{-1}\sum_{i=1}^{n} Y_i (1-D_i) K_b(X_i - x)}{n^{-1}\sum_{i=1}^{n} (1-D_i) K_b(X_i - x)}.$$
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) \( f_X (x), \nu (x) \) 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, \( \inf_{x \in S_X} f_X (x) > c > 0 \), \( h_0 \in \mathcal{H}^d \) and \( \mathbb{P}(\hat{h} \in \mathcal{H}^d) \to 1 \).

Proposition E2. Under Assumptions E1(iii-iv) and E2, the conclusion of Theorem 3.1 holds for this example.

4.3 Estimating equations with missing data

Consider inference based on the \( p \) estimating equations

\[
\mathbb{E} [s(Z_1, Z_2, \theta_0)] = 0,
\]

where \( Z_1 \) is a \( d_{Z_1} \)-dimensional random vector that is always observed and \( Z_2 \) is a \( d_{Z_2} \)-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 \( Z = (Z_1', Z_2'D, D)' \). We assume missingness at random, i.e., \( Z_2 \) is independent of \( D \), conditional on \( Z_1 \). Wang and Chen (2009) proposed EL inference based on nonparametric imputation in this general setting. See also Chen, Hong and Tarozzi (2008) 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. Wang and Chen (2009). Here, we apply our method to obtain a version of Wilks’ Theorem in this general setting for missing data.

We modify the approach of Wang and Chen (2009) and consider the estimating equation

\[
g(Z, \theta, \eta_0) = Ds(Z_1, Z_2, \theta) + (1 - D) \frac{q_0(Z_1, \theta)}{p_0(Z_1)},
\]

where \( \eta_0 = (q_0', p_0)' \), \( q_0(Z_1, \theta) := E[Ds(Z_1, Z_2, \theta)|Z_1] \) and \( p_0(Z_1) := E[D|Z_1] \) are infinite-dimensional nuisance parameters. This approach is slightly different from the one in Wang and Chen (2009), who proposed a nonparametric imputation method by sampling from a smoothed nonparametric estimator of the distribution of \( Z_2 \) given \( Z_1 \) 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 Wang and Chen (2009), our method is strictly more efficient than that based on imputing \( Z_2 \) 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. Never-
theless, our main contribution in this example is not the nonparametric imputation of $s$, but
rather obtaining distribution-free semiparametric EL inference.

Wang and Chen (2009, Lemma 1) provided sufficient conditions under which (2.5) holds with
$$ \phi(Z, \theta_0, h_0) = D \left( s(Z_1, Z_2, \theta_0) - \frac{q_0(Z_1, \theta_0)}{p_0(Z_1)} \right) \frac{1 - p_0(Z_1)}{p_0(Z_1)}. $$

Therefore, our method suggests doing inference with the estimating moment
$$ m(Z, \theta_0, h_0) = \frac{D}{p_0(Z_1)} s(Z_1, Z_2, \theta_0) + \left( 1 - \frac{D}{p_0(Z_1)} \right) \frac{q_0(Z_1, \theta_0)}{p_0(Z_1)}. $$

We propose to estimate $h_0 = \eta_0 = (q_0', p_0) \in \mathcal{H} := C_M^q(S_{Z_1}) \times \cdots \times C_M^q(S_{Z_1}) \times C_{1,\delta}(S_{Z_1})$, by
the NW kernel estimators

$$ \hat{q}(z_1, \theta) := \frac{1}{n} \sum_{i=1}^n \frac{D_i s(Z_{1i}, Z_{2i}, \theta) K_b(Z_{1i} - z_1)}{n^{-1} \sum_{j=1}^n K_b(Z_{1j} - z_1)} \quad (4.2) $$

and

$$ \hat{p}(z_1) := \frac{1}{n} \sum_{i=1}^n \frac{D_i K_b(Z_{1i} - z_1)}{n^{-1} \sum_{j=1}^n K_b(Z_{1j} - z_1)}. \quad (4.3) $$

We require the following assumption.

**Assumption E3:**

(i) We observe $Z := (Z_1', Z_2, D, D)'$ with $Z_2 \perp D|Z_1$.

(ii) $f_{Z_1}(z_1), q_0(z_1, \theta)$ and $p_0(z_1)$ are $r$ times continuously differentiable in $z_1$, with uniformly bounded derivatives (including zero derivatives), where $r$ is as in E1(iii). Moreover, $\inf_{z_1 \in S_{Z_1}} f_{Z_1}(z_1) > c > 0$, $h_0 \in \mathcal{H}^\delta$ and $P(\tilde{h} \in \mathcal{H}^\delta) \to 1$.

**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

$$ Q_{Y|X}(\tau|X) = \inf \{ t : \mathbb{P}(T \leq t|X) \geq \tau \} = X'\theta_0, $$

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, \( \Delta = I(T \leq C) \) is the censoring indicator, and \( C \) is the censoring time, which is assumed to be conditionally independent of \( T \) given \( X \). As in Leng and Tong (2013) we take \( X \) one-dimensional, and we consider the estimating equation

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

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 \( \eta_0 \) is estimated by the conditional (local) Kaplan-Meier estimator (Beran, 1981)

\[
\hat{\eta}(t|x) = \prod_{Y_i \leq t, \Delta_i = 0} \left( 1 - \frac{W_i(x, b_n)}{\sum_{j=1}^{n} I(Y_j \geq Y_i) W_j(x, b_n)} \right),
\]

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 weight, \( k \) is a one-dimensional density function, \( k_b(\cdot) = k(\cdot/b)/b \) and \( b \equiv b_n \) is a bandwidth sequence.

It follows from Theorem 3.2 in Du and Akritas (2002) that

\[
\hat{\eta}(t|x) - \eta_0(t|x) = \left. -\frac{\eta_0(t|x)}{f_X(x)} \right|_{t \rightarrow 0} n^{-1} \sum_{i=1}^{n} k_b(X_i - x) \xi(Y_i, \Delta_i, t|x) + R_n(t|x),
\]

where \( \sup_x \sup_{t \leq \tau_x} |R_n(t|x)| = O_P((nb_n)^{-3/4}(\log n)^{3/4}) = o_P(n^{-1/2}) \) provided \( nb_n^3(\log n)^{-3} \rightarrow \infty \), \( \tau_x < \inf \{ t : H(t|x) = 1 \} \) and

\[
\xi(y, \delta, t|x) = -\int_{-\infty}^{h^\vee t} \frac{dH_c(s|x)}{(1 - H(s|x))^2} + \frac{I(y \leq t, \delta = 0)}{1 - H(y|x)},
\]

with \( H(t|x) = \mathbb{P}(Y \leq t|X = x) \) and \( H_c(t|x) = \mathbb{P}(Y \leq t, \Delta = 0|X = x) \). 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 Powell, Stock and Stoker, 1989) it can be easily shown that

\[
n^{-1} \sum_{i=1}^{n} \{ g(Z_i, \theta_0, \hat{\eta}) - g(Z_i, \theta_0, \eta_0) \} = (1 - \tau)n^{-1} \sum_{i=1}^{n} X_i \xi(Y_i, \Delta_i, X_i'\theta_0|X_i) + o_P(n^{-1/2}).
\]

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 \( \theta \) and \( h = (h_1, h_2, h_3)' \),

\[
\phi(Z, \theta, h) = (1 - \tau)X \left[ -\int_{-\infty}^{h^\vee X'\theta} \frac{dh_2(s|X)}{(1 - h_1(s|X))^2} + \frac{I(Y \leq X'\theta, \Delta = 0)}{1 - h_1(Y|X)} \right],
\]
and hence
\[ m(Z, \theta, h) = X \left[ \frac{I(Y - X' \theta \geq 0)}{h_3(X' \theta | X)} - (1 - \tau) \right] + (1 - \tau) \left\{ - \int_{-\infty}^{Y \wedge T' \theta} \frac{dh_2(s | X)}{(1 - h_1(s | X))^2} + \frac{I(Y \leq X' \theta, \Delta = 0)}{1 - h_1(Y | X)} \right\}. \]

The functions \( h_j, j = 1, \ldots, 3 \) are supposed to belong to the space \( G \), defined by
\[ G = \{ g : \mathcal{S}_X \times \mathbb{R} \to [0, 1] : g(x, \cdot) \in \mathcal{B}_M \text{ for all } x \in \mathcal{S}_X, \text{ and } g(\cdot, t) \in C^i_M(\mathcal{S}_X, t), \text{ for all } t \in \mathbb{R} \}, \]
where \( q \geq 1 + \delta \) for some small \( \delta > 0 \), \( B_M = \{ f : \mathbb{R} \to [0, 1] : f \text{ has variation bounded by } M \} \), and \( \mathcal{S}_{X,t} = \{ x \in \mathcal{S}_X : t \leq x' \theta_0 \} \). Define \( \mathcal{H} = \{ (h_1, h_2, h_3)' : h_j \in G, j = 1, \ldots, 3 \} \). We equip \( \mathcal{H} \) with the pseudo-norm \( \|h\|_{\mathcal{H}} = \sum_{j=1}^{3} \sup_{x \in \mathcal{S}_X} \sup_{t \leq x' \theta_0} |h_j(t|x)| \) for \( h = (h_1, h_2, h_3)' \).

Finally, let
\[ \hat{H}(t|x) = \sum_{i=1}^{n} W_i(x, b_n) I(Y_i \leq t), \quad \hat{H}_c(t|x) = \sum_{i=1}^{n} W_i(x, b_n) I(Y_i \leq t, \Delta_i = 0). \]

We are now ready to state the regularity conditions and the main result for this example.

**Assumption E4:**

(i) We observe \( Z = (Y, X', \Delta)' \), where \( Y = T \wedge C, \Delta = I(T \leq C) \), 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 \( t_1, t_2 \in (-\infty, \infty) \),
\[ \left| H(t_1|x) - H(t_2|x) \right| \leq \left| L_1(t_1) - L_1(t_2) \right| \]
\[ \left| \frac{\partial}{\partial x} H(t_1|x) - \frac{\partial}{\partial x} H(t_2|x) \right| \leq \left| L_2(t_1) - L_2(t_2) \right| \]
\[ \left| \frac{\partial}{\partial x} H_c(t_1|x) - \frac{\partial}{\partial x} H_c(t_2|x) \right| \leq \left| L_3(t_1) - L_3(t_2) \right|. \]

(iv) The kernel function \( k \) is a symmetric probability density function with compact support, satisfying \( \int t^r k(t) dt = 0 \) for \( l = 1, \ldots, r - 1 \) and \( \int |t^r k(t)| dt < \infty \) for some \( r \geq 2 \). 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 \( G \).

**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 constructing confidence intervals for the ATE parameter

\[ \theta_0 = \mathbb{E}[Y(1) - Y(0)], \]

using the same design as that used by Ichimura and Linton (2005), where \( Y(0) = 2X + \eta \), \( Y(1) = Y(0) + \theta_0 \), and \( D = I(X\beta_0 + \varepsilon > 0) \) with both \( \eta \) and \( \varepsilon \) independent \( N(0, 1) \), and \( X \) is a \( U[-1/2, 1/2] \) random variable. Notice that \( \beta_0 \) 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 Ichimura and Linton (2005) for details \(^3\). The tables and figures below are based on 1000 replications. The tables report the finite sample coverage and average length at the 95% nominal level of confidence intervals using the popular Hirano, Imbens and Ridder’s (2001) estimator based on the normal approximation (Norm), its bootstrapped version (Boot), the normal approximation of its projected version based on the pathwise derivative (4.1) (Proj), the adjusted EL ratio (AEL), and the modified EL ratio based on the pathwise derivative (4.1) (MEL). The bootstrap estimator is computed as in Li, Racine and Wooldridge (2008) using 500 replications and using the design’s optimal bandwidths, whereas the adjusted EL ratio is based on the following statistic:

\[ -2\hat{\rho} \log EL_n(\theta_0, \hat{h}) \overset{d}{\to} \chi^2_1, \]

with the estimated adjustment

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

Tables 1-2 illustrate that the proposed method results in a test statistic characterized by good finite sample properties, typically better than those based on the other four asymptotically equivalent test statistics. To further investigate this result we conduct some sensitivity analysis

\(^3\)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 not reported.
and compute the finite sample coverage and average length for the five statistics using as bandwidths the values $kb/4$, $k = 1, 2, \ldots, 10$ for $n = 100$. Figures 1-4 are based on $\theta_0 \in \{-2, 0\}$, $\beta_0 \in \{1, 3\}$ and show the trade off bias-variance of kernel estimators, where smaller biases (smaller bandwidths) produce confidence intervals with good coverage but as the bandwidth increases (smaller variances) the length of the confidence interval decreases. Interestingly the proposed method seems to be less sensitive to the choice of the bandwidth.

Figures 1-4 approximately here

5.2 Estimating equations with missing data

We consider a logit model with missing covariates, similar to the model considered by Wang and Chen (2009). The estimating equation is

$$s(Z_1, Z_2, \theta) = X (Y - \Lambda (X'\theta))$$

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 that are always observed are $Z_1 = (Y, X_1)'$, while the missing variable is $Z_2 = X_2$ 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) and a score type statistic based on the inverse probability weighted version (S) of the original estimating equation. Details about implementation of these statistics are given below.

The adjusted EL ratio is based on the imputed estimating equation

$$\tilde{s}(Z, \theta) = Ds(Z_1, Z_2, \theta) + (1 - D) \frac{\hat{q}(Z_1, \theta)}{\hat{p}(Z_1)},$$

which corresponds to the limit case of Wang and Chen (2009)’s estimating equation. In this case the estimated adjustment is

$$\hat{p} = \frac{tr (\hat{\Sigma}^{-1} \hat{Q})}{tr (\hat{V}^{-1} \hat{Q})},$$
\[ \hat{S} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\hat{q}^2(Z_{1i})}{\hat{p}(Z_{1i})} + \hat{q}(Z_{1i}, \hat{\theta}) \hat{q}(Z_{1i}, \hat{\theta})' \right), \]

\[ \hat{\sigma}^2(z_1) = \frac{1}{n} \sum_{i=1}^{n} D_i s(Z_{1i}, Z_{2i}, \hat{\theta}) s(Z_{1i}, Z_{2i}, \hat{\theta})' K_b(Z_{1i} - z_1) \]

\[ \hat{V} = \frac{1}{n} \sum_{i=1}^{n} s(Z_i, \hat{\theta}) s(Z_i, \hat{\theta})', \quad \hat{Q} = \frac{1}{n} \left( \sum_{i=1}^{n} s(Z_i, \hat{\theta}) \left( \sum_{i=1}^{n} s(Z_i, \hat{\theta}) \right)' \right), \]

and \( \hat{q}(z_1, \theta) \) and \( \hat{p}(z_1) \) are defined in (4.2) and (4.3), respectively. Then, it can be shown that

\[ -2\hat{\rho} \log E L_n(\theta_0, \hat{\mu}) \xrightarrow{d} \chi^2. \]

The bootstrap version of the EL ratio follows the procedure suggested by Shao and Sitter (1996) 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 \( \{\hat{q}^*(Z_{1i}, \theta) / \hat{p}^*(Z_{1i})\}_{i=1}^{n} \) from the imputed values \( \{\hat{q}(Z_{1i}, \theta) / \hat{p}(Z_{1i})\}_{i=1}^{n} \) are drawn to form the bootstrap analogue \( \hat{q}^*(Z^*_i, \theta) \) of \( \hat{q}(Z_i, \theta) \); (2) the bootstrap EL ratio statistic \( EL_n^*(\theta_0, \hat{\mu}^*) \) is computed using the centered version of \( \hat{q}^*(Z^*_i, \theta_0) \); (3) steps (1)-(2) are repeated \( B \) times. The consistency of this bootstrap procedure follows by standard arguments (see for example those used by Wang and Chen (2009)). Finally, the inverse probability weighted score type statistic \( S \) is

\[ S = \frac{1}{n} \left( \sum_{i=1}^{n} \frac{D_i s(Z_{1i}, Z_{2i}, \hat{\theta})}{\hat{p}(Z_{1i})} \right)' \left( \sum_{i=1}^{n} \frac{D_i s(Z_{1i}, Z_{2i}, \hat{\theta})}{\hat{p}(Z_{1i})} \right)^{-1} \left( \sum_{i=1}^{n} \frac{D_i s(Z_{1i}, Z_{2i}, \hat{\theta})}{\hat{p}(Z_{1i})} \right), \]

where

\[ \hat{\Sigma}_{\hat{p}} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{D_i \hat{V} \text{ar} \left( s(Z_{1i}, Z_{2i}, \hat{\theta}) | Z_{1i} \right)}{\hat{p}(Z_{1i})^2} + \frac{D_i}{\hat{p}(Z_{1i})^2} \hat{q}_p(Z_{1i}, \hat{\theta}) \hat{q}_p(Z_{1i}, \hat{\theta})' \right), \]

\[ \hat{q}_p(z_1, \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i}{R(Z_{1i})} s(Z_{1i}, Z_{2i}, \hat{\theta}) K_b(Z_{1i} - z_1). \]

The tables and figures below are based on 1000 replications. Table 3 reports the finite sample coverage and average length at the 95% nominal level of the confidence intervals for the slope parameters \( \theta_1 \) and \( \theta_2 \). Table 4 reports the finite sample coverage at the 95% nominal level of the confidence region for \( \theta_1 \) and \( \theta_2 \).

Tables 3, 4 approximately here

Figure 5 shows the 95% nominal level confidence regions for \( \theta_1 \) and \( \theta_2 \) for \( n = 100 \); those for \( n = 300 \) are similar and hence are not shown.

Figure 5 approximately here
Tables 3 and 4 and Figure 5 confirm the results of the ATE example and indicate that the proposed method results in a test statistic characterized by finite sample properties typically better than those based on other asymptotically equivalent test statistics. As with the ATE example we conduct some sensitivity analysis and compute the finite sample coverage and average length for the four statistics using as bandwidths the following values: \(b/4\), \(b/2\), \(3b/4\), \(5b/4\), \(2b\) where \(b\) is the cross-validated bandwidth. Figures 6 and 7 show the results for \(n = 100\) and confirm that, while the finite sample properties of all of the proposed statistics depend on the choice of bandwidth, the proposed method is typically less sensitive to that choice.

6 Conclusions

In this article we have presented a new way to conduct empirical likelihood 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 projecting a criterion function on the orthocomplement of the tangent space of nuisance functions. 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. 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 empirical likelihood statistic. This is 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 \(\theta_0\). An EL estimator based on the modified moments is expected to possess good bias properties, as conjectured in Newey, Hsieh and Robins (2004) in the context of semiparametric two-step method of moments estimators. This topic is deferred for future investigation.

Acknowledgements

Juan Carlos Escanciano gratefully acknowledges support by the Spanish Plan Nacional de I+D+I, reference number ECO2012-33053. Ingrid Van Keilegom acknowledges financial support from the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract ‘Projet d’Actions de Recherche Concertées’ (ARC) 11/16-039 of the ‘Communauté française de Belgique’, granted by the ‘Académie universitaire Louvain’.
References


### Appendix A Proofs

**Proof of Theorem 3.1.** We check the conditions of Theorem 2.1 in Hjort, McKeague and Van Keilegom (2009) (taking in their notation \( a_n = 1 \) and \( m_n = m/\sqrt{n} \)). (A0) and (A3) correspond to our Assumption A(v). We check their condition (A1), which corresponds to (2.7).

By Assumption A, and the standard Central Limit Theorem,

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Z_i, \theta_0, \hat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Z_i, \theta_0, h_0) + o_p(1) \overset{d}{\to} N(0, \Sigma).
\]

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

**Proof of Proposition E1.** Assumption A(i) follows by standard empirical process arguments after noticing that the class \( \mathcal{F} = \{g(\cdot, \theta_0, h) + \phi(\cdot, \theta_0, h) : h \in \mathcal{H}\} \) is the sum of two \( \mathbb{P}-\)Donsker classes. A(ii) follows by checking Assumption B. B(i) trivially holds by the linearity in \( \eta_0 \) and because \( f_X \) is bounded away from zero. The rate condition for Assumption B(ii) follows from Assumption E1 and standard results in kernel estimation. The zero derivative condition of B(iii) follows from Newey (1994, p. 1361) and the fact that the derivative with respect to \( f_W \) and \( f_X \) is zero by the conditional restriction

\[
\mathbb{E}[Y - \eta_0(X) | X] = 0 \ a.s.
\]

Assumption A(iv) also holds since the class \( \mathcal{M} := \{m(\cdot, \theta_0, h) : h \in \mathcal{H}^\delta\} \) is \( \mathbb{P}\)-Glivenko-Cantelli with an envelope function \( \theta_0 + C \), where \( C \) is a positive constant. Therefore, the class \( \mathcal{M}^2 := \{m(\cdot, \theta_0, h)^2 : h \in \mathcal{H}^\delta\} \) is \( \mathbb{P}\)-Glivenko-Cantelli.
To verify A(v) first note that by the triangle inequality and for an arbitrarily small $\epsilon > 0$,
\[
P\left(MEL_n(\theta_0, \hat{h}) = 0 \right) = P\left(m(Z_i, \theta_0, \hat{h}) > 0 \text{ or } m(Z_i, \theta_0, \hat{h}) < 0 \text{ for all } Z_i \in \mathcal{Z} \right) \\
\leq \prod_{i=1}^{n} P\left(m(Z_i, \theta_0, h_0) > \epsilon \right) \\
+ \prod_{i=1}^{n} P\left(m(Z_i, \theta_0, h_0) < -\epsilon \right) \\
+ P\left(\sup_{Z_i \in \mathcal{Z}} \left|m(Z_i, \theta_0, \hat{h}) - m(Z_i, \theta_0, h_0)\right| > \epsilon \right) \rightarrow 0
\]
by the facts that $E(m(Z, \theta_0, h_0)) = 0$, $E(m(Z, \theta_0, h_0)^2) < \infty$ and
\[
\max_i \left| m\left(Z_i, \theta_0, \hat{h}\right) - m\left(Z_i, \theta_0, h_0\right) \right| \\
\leq \max_i \left| \int_0^{M} (\hat{\eta} - \eta_0)(u, W_i)du \right| + \max_i \left| (\hat{\eta} - \eta_0)(X_i) \frac{\hat{f}_W(W_i)}{f_X(X_i)} \right| \\
+ \max_i |Y_i - \eta_0(X_i)| \max_i \left| \frac{\hat{f}_W(W_i)}{f_X(X_i)} - \frac{f_W(W_i)}{f_X(X_i)} \right| \\
= o_P(1),
\]
by a standard result on the uniform consistency of kernel estimators (see for example Masry, 1996). To show the second part of A(v) note that by the triangle inequality and Chebychev’s inequality,
\[
P\left(\max_i \left| m\left(Z_i, \theta_0, \hat{h}\right) \right| > 2\sqrt{n}\epsilon \right) \\
\leq P\left(\max_i \left| m\left(Z_i, \theta_0, h_0\right) \right| > \sqrt{n}\epsilon \right) \\
+ \sum_i P\left(\int_0^{M} (\hat{\eta} - \eta_0)(u, W_i)du - Y_i \left(\frac{\hat{f}_W(W_i)}{f_X(X_i)} - \frac{f_W(W_i)}{f_X(X_i)}\right) \right. \\
+ (\hat{\eta} - \eta_0)(X_i) \frac{\hat{f}_W(W_i)}{f_X(X_i)} + \eta_0(X_i) \left(\frac{\hat{f}_W(W_i)}{f_X(X_i)} - \frac{f_W(W_i)}{f_X(X_i)}\right) \left. \right| > \sqrt{n}\epsilon \right) \\
\leq P\left(\max_i \left| m\left(Z_i, \theta_0, h_0\right) \right| > \sqrt{n}\epsilon \right) + \frac{4}{\epsilon^2} \left[ E\left(\int_0^{M} (\hat{\eta} - \eta_0)(u, W)du\right)^2 \right. \\
+ E\left(\frac{(Y - \eta_0(X))}{f_X(X)} \left(\frac{\hat{f}_W(W)}{f_X(X)} - \frac{f_W(W)}{f_X(X)}\right) \right)^2 \\
+ E\left(\frac{(\hat{\eta} - \eta_0)(X) \frac{\hat{f}_W(W)}{f_X(X)}}{1}\right)^2 \right] \\
= o(1),
\]
where the last equality follows from the Cauchy-Schwarz inequality combined with the $L_2$-consistency of kernel estimators (Andrews, 1995, Stone, 1982).

Proof of Proposition E2. We consider the classes of functions

$$
G = \left\{ z = (y, d, x') \to l(z, \eta) = \frac{yd}{\eta(x)} - \frac{y(1 - d)}{1 - \eta(x)} : \eta \in \mathcal{C}_{1, \varepsilon}^q(S_X) \right\}
$$

and $H = \mathcal{C}_{1, \varepsilon}^q(S_X) \times \mathcal{C}_M^q(S_X) \times \mathcal{C}_M^q(S_X)$, with $\|h\|_H = \|\eta\|_\infty + \|\mu_0\|_\infty + \|\mu_1\|_\infty$, for $h = (\eta, \mu_0, \mu_1) \in H$. We show first that $G$ is $\mathbb{P}$–Donsker. Notice that, for $\varepsilon < \eta < 1 - \varepsilon$,

$$
\left| \frac{\partial l(z, \eta)}{\partial \eta} \right| = \left| \frac{-yd}{\eta^2} - \frac{y(1 - d)}{(1 - \eta)^2} \right| \leq \frac{2|y|}{\varepsilon^2} \equiv C(y).
$$

Then, for any $\eta$ and $\eta_1$,

$$
|l(z, \eta) - l(z, \eta_1)| \leq C(y)\|\eta - \eta_1\|_\infty.
$$

We use brackets of the form $[l(z, \eta_j) - \delta C(y), l(z, \eta_j) + \delta C(y)]$, with $\eta_j$ the center of balls of radius $\delta$ covering $\mathcal{C}_{1, \varepsilon}^q(S_X)$. These brackets have $\| \cdot \|_2$–size of $2\delta \|C(\cdot)\|_2$. For any $\eta \in \mathcal{C}_{1, \varepsilon}^q(S_X)$ there exists $j \in \{1, ..., N_\delta \equiv N(\delta, \mathcal{C}_{1, \varepsilon}^q(S_X), \| \cdot \|_\infty)\}$ such that $\|\eta - \eta_j\|_\infty < \delta$. Moreover, by the previous display,

$$
|l(z, \eta) - l(z, \eta_j)| \leq C(y)\delta.
$$

Therefore,

$$
N([2\delta \|C(\cdot)\|_2, G, \| \cdot \|_2) \leq N(\delta, \mathcal{C}_{1, \varepsilon}^q(S_X), \| \cdot \|_\infty),
$$

and $G$ is $\mathbb{P}$–Donsker provided $q > d_x/2$, see van der Vaart and Wellner (1996, p.154). Similarly, the class

$$
\Phi := \left\{ z \to \phi(z, \theta_0, h) = (d - \eta(x)) \left( \frac{\mu_1(x)}{\eta(x)} + \frac{\mu_0(x)}{1 - \eta(x)} \right) : h \in H \right\}
$$

is $\mathbb{P}$–Donsker. Then, Assumption A(i) holds.

To check for A(ii) we verify Assumption B. The moment is twice Hadamard differentiable with a bounded second derivative and a zero first derivative. The rate condition for Assumption B(ii) follows from Assumption E1 and standard results in kernel estimation. The zero derivative condition follows from Newey (1994, p. 1361) and the fact that the derivative with respect to $\mu_1$ and $\mu_0$ of $h \to \mathbb{E}[\phi(Z, \theta_0, h)]$ at $h_0$ is zero by the conditional restriction

$$
\mathbb{E}[D - \eta_0(X)|X] = 0 \ a.s.
$$

To see that the second derivative is bounded, we use that the propensity score is bounded away from zero and one to conclude that

$$
\left| \frac{\partial^2 l(z, \theta_0, \eta)}{\partial \eta^2} \right| = \left| \frac{-2yd}{\eta^3} + \frac{2y(1 - d)}{(1 - \eta)^3} \right| \leq \frac{4|y|}{\varepsilon^3} \equiv C_2(y),
$$

27
and similarly for $\partial^2 \phi(z, \theta_0, h)/\partial h^2$. Finally, the verification of Assumption A(iv) follows from the class

$$\mathcal{F} := \{m(\cdot, \theta_0, h) : h \in \mathcal{H}\},$$

being $\mathbb{P}$–Donsker, and hence, $\mathbb{P}$–Glivenko-Cantelli, and the fact that $\mathcal{F}$ has a square integrable envelope function by E2(i). To check A(v) the same arguments as those used in the proof of Proposition E1 show first that $\mathbb{P}(\text{MEL}_n(\theta_0, \hat{h}) = 0) \to 0$ since $\max_i |\hat{\eta}_i - \eta_0| = o_p(1)$ and $\max_i |\hat{\iota}_i - \iota_0| = o_p(1)$ where $\hat{\eta}_i = \hat{\eta}(X_i)$, $\hat{\iota}_i = \hat{\mu}_1(X_i)/\hat{\eta}_i + \hat{\mu}_0(X_i)/(1 - \hat{\eta}_i)$. To verify the second claim note that, for any $\delta > 0$,

$$\mathbb{P} \left( \max_i |m(Z_i, \theta_0, h_0)\mathbb{E} \left[ \frac{Y(nDZ_1)}{\eta \theta_0} \right]^2 + \mathbb{E} \left[ \frac{Y(1 - D)}{(1 - \eta)(1 - \hat{\eta})} \right]^2 \right),$$

and the conclusion follows as in in proof of the second part of Proposition E1. □

**Proof of Proposition E3.** By same arguments used in Propoposition E2, it is shown that the class of functions $\mathcal{F} := \{m(\cdot, \theta_0, h) : h \in \mathcal{H}\}$ is $\mathbb{P}$–Donsker. Then, Assumption A(i) holds. To check A(ii) we verify Assumption B. The moment is twice pathwise differentiable at $h = h_0$ with uniformly bounded second derivatives, as $p_0 \in C_1^{q, \varepsilon}(S_{Z_1})$. The rate condition for Assumption B(ii) follows from Assumption E1 and standard results in kernel estimation. The zero derivative condition follows from Newey (1994, p. 1361). That is, the pathwise derivative with respect to $q_0$ is clearly zero, since

$$\mathbb{E} \left[ \left( 1 - \frac{D}{p_0(Z_1)} \right) \bigg| Z_1 \right] = 0 \text{ a.s.}$$

The derivative with respect to $p_0$ is more involved, but it is equally zero by the last display and the condition

$$\mathbb{E} \left[ \frac{D}{p_0^2(Z_1)} \left( s(Z_1, Z_2, \theta_0) - \frac{q_0(Z_1, \theta_0)}{p_0(Z_1)} \right) \bigg| Z_1 \right] = 0 \text{ a.s.},$$

which holds by the conditional independence assumption. Finally, Assumption A(iv) follows because the class $\mathcal{F}$ is $\mathbb{P}$–Donsker. To check A(v) the first part follows as in the proof of Proposition E1 since $\max_i |\hat{p}(Z_{1i}) - p_0(Z_{1i})| = o_p(1)$ and $\max_i |\hat{q}(Z_{1i}, \theta_0) - q_0(Z_{1i}, \theta_0)| = o_p(1)$. 28
For the second part, note that for any \( \delta > 0 \),
\[
\mathbb{P} \left( \max_i \left| m \left( Z_i, \theta_0, \hat{h} \right) \right| > 2\sqrt{n} \delta \right) \leq \mathbb{P} \left( \max_i \left| m \left( Z_i, \theta_0, h_0 \right) \right| > \sqrt{n} \delta \right) \\
+ \frac{3}{\delta^2} \mathbb{E} \left[ \mathbb{E}^2 \left[ \frac{D_s \left( Z_1, Z_2, \theta_0 \right)}{p_0 \left( Z_1 \right) p_0 \left( Z_1 \right)} \frac{p_0 \left( Z_1 \right) - \hat{p} \left( Z_1 \right)}{p_0 \left( Z_1 \right) - \hat{p} \left( Z_1 \right)} \right]^2 \right] \\
+ \frac{3}{\delta^2} \mathbb{E} \left[ \left( 1 - \frac{D}{\hat{p} \left( Z_1 \right)} \right) \frac{\hat{q} \left( Z_1, \theta_0 \right) - q_0 \left( Z_1, \theta_0 \right)}{\hat{p} \left( Z_1 \right)} \right]^2 \right] \\
+ \frac{3}{\delta^2} \mathbb{E} \left[ \left( 1 - \frac{D}{\hat{p} \left( Z_1 \right)} \right) \frac{q_0 \left( Z_1, \theta_0 \right)}{\hat{p} \left( Z_1 \right) p_0 \left( Z_1 \right)} \left( \hat{p} \left( Z_1 \right) - p_0 \left( Z_1 \right) \right) \right]^2 \to 0,
\]
as in the proof of the second part of Proposition E1.

\( \square \)

**Proof of Proposition E4.** We start by verifying assumption A(i). Lemma 6.1 in Lopez (2009) shows that the class \( \mathcal{G} \) satisfies for any \( \epsilon > 0 \),
\[
\log N_{[\,]} (\epsilon, \mathcal{G}, \| \cdot \|_\mathcal{H} ) = O(\epsilon^{-2/(1+\delta)}),
\]
and hence we also have that \( \log N_{[\,]} (\epsilon, \mathcal{H}, \| \cdot \|_\mathcal{H} ) = O(\epsilon^{-2/(1+\delta)}) \). Define \( \mathcal{M} = \{ z \to m(z, \theta_0, h) : h \in \mathcal{H} \} \). Since the map \( h \to m(z, \theta_0, h) \) is uniformly Lipschitz continuous (in the sense that \( |m(z, \theta_0, h) - m(z, \theta_0, \tilde{h})| \leq c(z)\|h - \tilde{h}\|_\mathcal{H} \), with \( \mathbb{E}(c^2(Z)) < \infty \)), it easily follows (as in the proof of Theorem 3 in Chen, Linton and Van Keilegom, 2003) that
\[
\log N_{[\,]} (\epsilon, \mathcal{M}, \| \cdot \|_2 ) = O(\epsilon^{-2/(1+\delta)}),
\]
and hence \( \mathcal{M} \) is \( \mathbb{P} \)-Donsker, which implies A(i).

For Assumption A(ii), we use a direct approach. Using the fact that \( \mathbb{E}(m(Z, \theta_0, h_0)) = 0 \), straightforward calculations show that
\[
\mathbb{E}(m(Z, \theta_0, \hat{h})) \\
= \mathbb{E} \left[ X \left\{ - \frac{1 - \tau}{\eta_0(X' \theta_0 | X)} \left( \hat{\eta}(X' \theta_0 | X) - \eta_0(X' \theta_0 | X) \right) \right. \right. \\
\left. \left. - (1 - F(X' \theta_0 | X)) \left( \int_{-\infty}^{X' \theta_0} \frac{\hat{H}(s | X) - H(s | X)}{(1 - H(s | X))^2} dH_c(s | X) \right. \right. \right. \\
\left. \left. \left. + \frac{\hat{H}_c(X' \theta_0 | X) - H_c(X' \theta_0 | X)}{1 - H(X' \theta_0 | X)} \right) \right. \right. \right. \\
+ \int_{-\infty}^{X' \theta_0} \frac{\hat{H}_c(s | X) - H_c(s | X)}{(1 - H(s | X))^2} dH(s | X) \right) \right] \right].
\]
Using the i.i.d. representation of \( \hat{\eta}(X' \theta_0 | X) - \eta_0(X' \theta_0 | X) \) given in (4.4), the latter can be written
as

\[ n^{-1} \sum_{i=1}^{n} \mathbb{E} \left[ X \frac{1 - \tau}{f_X(x)} k_b(X_i - x) \xi(Y_i, \Delta_i, X' \theta_0 | X) \right] \]

\[ - (1 - \tau) \sum_{i=1}^{n} \mathbb{E} \left[ X W_i(X, b_n) \left( \int_{-\infty}^{X' \theta_0} \frac{I(Y_i \leq s) - H(s | X)}{(1 - H(s | X))^2} dH_c(s | X) \right) \right] \]

\[ + \frac{I(Y_i \leq X' \theta_0, \Delta_i = 1) - H_c(X' \theta_0 | X)}{1 - H(X' \theta_0 | X)} \left( \int_{-\infty}^{X' \theta_0} \frac{I(Y_i \leq s, \Delta_i = 1) - H_c(s | X)}{(1 - H(s | X))^2} dH(s | X) \right) \]

\[ + o_p(n^{-1/2}) \]

\[ = n^{-1} \sum_{i=1}^{n} \mathbb{E} \left[ X (1 - \tau) k_b(X_i - X) \left\{ \frac{1}{f_X(x)} - \frac{1}{\hat{f}_X(x)} \right\} \xi(Y_i, \Delta_i, X' \theta_0 | X) \right] + o_p(n^{-1/2}) \]

\[ = o_p(n^{-1/2}), \]

with \( \hat{f}_X(x) = n^{-1} \sum_{i=1}^{n} k_b(X_i - x) \), since \( \sup_x |\hat{f}_X(x) - f_X(x)| = O_p((nb_n)^{-1/2}(\log n)^{1/2} + b_n^r) \) and \( \sup_x |n^{-1} \sum_{i=1}^{n} k_b(X_i - x)\xi(Y_i, \Delta_i, x' \theta_0 | x)| = O_p((nb_n)^{-1/2}(\log n)^{1/2} + b_n^r) \), which are \( o_p(n^{-1/4}) \) provided \( nb_n^r \to 0 \) and \( nb_n^2(\log n)^{-2} \to \infty \). Hence, A(ii) follows.

We next consider A(iii). It follows from Proposition 1 and 2 in Akritas and Van Keilegom (2001) and Lemma 6.2 in Lopez (2009) that \( \mathbb{P}(\hat{h} \in \mathcal{H}^\delta) \to 1 \). Moreover, \( \|\hat{h} - h_0\|_{\mathcal{H}} = o_p(1) \), because of the uniform consistency of the Nadaraya-Watson estimator and of the Beran (1981) estimator. Finally, A(iv) follows using similar arguments as for A(i), and A(v) can be shown in the same way as for the previous examples.

\[ \square \]

### Appendix B  Tables and figures

<table>
<thead>
<tr>
<th>( \theta_0 )</th>
<th>( \beta_0 )</th>
<th>Norm</th>
<th>Boot</th>
<th>AEL</th>
<th>MEL</th>
<th>Proj</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1</td>
<td>0.895</td>
<td>0.774</td>
<td>0.907</td>
<td>0.747</td>
<td>0.906</td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
<td>0.901</td>
<td>0.772</td>
<td>0.915</td>
<td>0.758</td>
<td>0.905</td>
</tr>
<tr>
<td>-2</td>
<td>3</td>
<td>0.903</td>
<td>0.771</td>
<td>0.914</td>
<td>0.762</td>
<td>0.908</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.902</td>
<td>0.753</td>
<td>0.915</td>
<td>0.748</td>
<td>0.908</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.904</td>
<td>0.748</td>
<td>0.913</td>
<td>0.752</td>
<td>0.910</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0.901</td>
<td>0.745</td>
<td>0.917</td>
<td>0.742</td>
<td>0.910</td>
</tr>
</tbody>
</table>

Table 1: Finite sample coverage (left column) and average length (right column) of 95 % confidence intervals for \( \theta_0 \) in the ATE example, and for \( n = 100 \).
Figure 1: Finite sample coverage (left) and average length (right) for MEL (solid curve), Normal (dashed curve), AEL (long dashed curve), Boot (dot dashed curve) and Proj (two dashed curve), in the ATE example for $n = 100$.

Figure 2: Finite sample coverage (left) and average length (right) for MEL (solid curve), Normal (dashed curve), AEL (long dashed curve), Boot (dot dashed curve) and Proj (two dashed curve), in the ATE example for $n = 100$. 

Figure 3: Finite sample coverage (left) and average length (right) for MEL (solid curve), Normal (dashed curve), AEL (long dashed curve), Boot (dot dashed curve) and Proj (two dashed curve), in the ATE example for $n = 100$.

Figure 4: Finite sample coverage (left) and average length (right) for MEL (solid curve), Normal (dashed curve), AEL (long dashed curve), Boot (dot dashed curve) and Proj (two dashed curve), in the ATE example for $n = 100$. 
Figure 5: Confidence regions for MEL (solid curve), AEL (dashed curve), AELboot (dot dashed curve) and S (long dashed curve), in the missing data example for $n = 100$.

Figure 6: Finite sample coverage (left) and average length (right) for MEL (solid curve), AEL (dashed curve), AEL (long dashed curve), AELboot (dot dashed curve) and S (long dashed curve), in the missing data example for $n = 100$. 
<table>
<thead>
<tr>
<th>$\theta_0$</th>
<th>$\beta_0$</th>
<th>Norm</th>
<th>Boot</th>
<th>AEL</th>
<th>MEL</th>
<th>Proj</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1</td>
<td>0.923</td>
<td>0.704</td>
<td>0.945</td>
<td>0.678</td>
<td>0.946</td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
<td>0.928</td>
<td>0.694</td>
<td>0.942</td>
<td>0.687</td>
<td>0.941</td>
</tr>
<tr>
<td>-2</td>
<td>3</td>
<td>0.931</td>
<td>0.693</td>
<td>0.941</td>
<td>0.685</td>
<td>0.940</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.929</td>
<td>0.685</td>
<td>0.940</td>
<td>0.680</td>
<td>0.937</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.930</td>
<td>0.678</td>
<td>0.941</td>
<td>0.682</td>
<td>0.939</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0.928</td>
<td>0.677</td>
<td>0.943</td>
<td>0.675</td>
<td>0.938</td>
</tr>
</tbody>
</table>

Table 2: Finite sample coverage (left column) and average length (right column) of 95 % confidence intervals for $\theta_0$ in the ATE example, and for $n = 300$.

<table>
<thead>
<tr>
<th>$n = 100$</th>
<th>$n = 300$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
</tr>
<tr>
<td>AEL</td>
<td>0.911</td>
</tr>
<tr>
<td>AELboot</td>
<td>0.921</td>
</tr>
<tr>
<td>MEL</td>
<td>0.923</td>
</tr>
<tr>
<td>S</td>
<td>0.888</td>
</tr>
</tbody>
</table>

Table 3: Finite sample coverage (left column) and average length (right column) of 95 % confidence intervals for $\theta_1$ and $\theta_2$ in the missing data example.

<table>
<thead>
<tr>
<th>$n = 100$</th>
<th>$n = 300$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEL</td>
<td>0.902</td>
</tr>
<tr>
<td>AELboot</td>
<td>0.934</td>
</tr>
<tr>
<td>MEL</td>
<td>0.936</td>
</tr>
<tr>
<td>S</td>
<td>0.895</td>
</tr>
</tbody>
</table>

Table 4: Finite sample coverage of 95 % confidence regions for $(\theta_1, \theta_2)$ in the missing data example.
Figure 7: Finite sample coverage (left) and average length (right) for MEL (solid curve), AEL (dashed curve), AEL (long dashed curve), AELboot (dot dashed curve) and S (long dashed curve), in the missing data example for $n = 100$. 