LIKELIHOOD INFERENCE ON SEMIPARAMETRIC MODELS: AVERAGE DERIVATIVE AND TREATMENT EFFECT

YUKITOSHI MATSUSHITA AND TAISUKE OTSU

Abstract. In the past few decades, much progress has been made in semiparametric modeling and estimation methods for econometric analysis. This paper is concerned with inference (i.e., confidence intervals and hypothesis testing) in semiparametric models. In contrast to the conventional approach based on t-ratios, we advocate likelihood-based inference. In particular, we study two widely applied semiparametric problems, weighted average derivatives and treatment effects, and propose semiparametric empirical likelihood and jackknife empirical likelihood methods. We derive the limiting behavior of these empirical likelihood statistics and investigate their finite sample performance via Monte Carlo simulation. Furthermore, we extend the (delete-1) jackknife empirical likelihood toward the delete-d version with growing d and establish general asymptotic theory. This extension is crucial to deal with non-smooth objects, such as quantiles and quantile average derivatives or treatment effects, due to the well-known inconsistency phenomena of the jackknife under non-smoothness.

1. Introduction

Recent years have witnessed a surge of research using semiparametric and nonparametric modeling techniques to answer empirical economic questions. This is partly because economic theory seldom suggests parametric functional or distribution forms for economic data and partly because of the sharp increase in high-quality and large-scale data sets combined with declining computational cost.

This paper is concerned with inference (i.e., confidence intervals and hypothesis testing) in semiparametric models. The conventional approach to conduct inference on parametric or finite-dimensional objects of interest is based on t-ratios. Typically the confidence interval of a parameter is formed by 'estimate $\pm 2 \cdot$ standard error', where the standard error is computed by taking a sample counterpart of the limiting variance formula of the corresponding semiparametric estimator. A major advantage of this conventional approach is its convenience: it requires only two inputs, the estimate and standard error. However, there are at least two concerns regarding this approach. First, by construction, the confidence interval is always centered around the parameter estimate. This is because we determine the shape of the confidence interval based on the limiting normal approximation. This shape constraint on the confidence interval may not be innocuous in certain situations, such as inference on the variance. Second, it should be emphasized that the conventional confidence interval involves another estimation problem: estimation of the limiting variance. Since the asymptotic variances of semiparametric estimators usually involve nonparametric components, their estimation or computation requires additional nonparametric fit, which demand additional smoothing parameters, such as bandwidths and series lengths.

In this paper, we advocate an alternative inference approach based on semiparametric or nonparametric likelihoods. If the distribution form of the data is known to the researcher, it is possible to invert the likelihood ratio statistic, say $\ell(\theta)$, to construct the confidence set $\{\theta : \ell(\theta) \leq c\}$ for some critical value c based on the chi-squared distribution. This construction obviously circumvents the above critiques on the conventional confidence intervals based on t-ratios. The shape of the likelihood-based confidence set is determined by data emphasis through the likelihood function. Also, the construction does not involve the standard errors. Theoretical properties of the parametric likelihood methods are summarized in Severini (2000). A remarkable feature of the likelihood ratio statistic is it converges in distribution to the chi-squared distribution (called Wilks' phenomenon) so that the critical value c does not contain any unknown objects.

Since Owen's (1988) discovery of empirical likelihood, numerous works extended this likelihood-based inference approach toward semiparametric and nonparametric econometric and statistical problems. For example, Owen (1988) proposed empirical likelihood inference on population means (without specifying the distribution form) and established Wilks' phenomenon. DiCiccio, Hall and Romano (1991) showed that the empirical likelihood ratio statistic admits higher-order refinement, called the Bartlett correction. We refer to Owen (2001) for a review on the method of empirical likelihood.

This paper focuses on two widely applied semiparametric problems in econometrics, weighted average derivatives and treatment effects, and explores empirical likelihood methods for these problems. Average derivatives are widely used to estimate parameters in single index models (e.g., a binary choice model with an unknown link function) and marginal effects of covariates in some nonseparable models (see, Section 2.1 for some references). Treatment effect analysis is one of the most intensively studied topics in econometrics and statistics (see, Section 3.1 for some references). A common feature of these objects is that both are written in the form $\theta = E[g(Z, h)]$ with unknown functions h. Based on this expression, the object θ is often estimated by the sample average $n^{-1} \sum_{i=1}^{n} g(Z_i, \hat{h})$ using a preliminary nonparametric estimator $\hat{h}$. Without h, the problem reduces to inference on the population moment $E[g(Z)]$ with known g, and the empirical likelihood statistic converges to the chi-squared distribution (Wilks' phenomenon). However, if we apply the same method to the plug-in moment function $g(Z_i, \hat{h})$, Wilks' phenomenon may not emerge. Indeed the empirical likelihood ratio generally converges to a weighted chi-squared distribution (Hjort, McKeague and van Keilegom, 2009), where the weights involve unknown nonparametric objects to be estimated. An obvious reason for this is the influence from the estimation error of $\hat{h}$.

This paper employs two modifications of empirical likelihood to recover Wilks' phenomenon for average derivatives and treatment effects. The first approach, called semiparametric empirical likelihood (Bravo, Escanciano and van Keilegom, 2015, and Matsushita and Otsu, 2016), modifies the moment function by adding a correction term to 'undo' the influence of $\hat{h} - h$. Bravo, Escanciano and van Keilegom (2015) developed a general theory of semiparametric empirical likelihood for semiparametric two-step estimators. Matsushita and Otsu (2016) applied this approach to semiparametric three-step estimators investigated in Hahn and Ridder (2013). We

apply the semiparametric empirical likelihood method to the weighted average derivatives and derive Wilks' phenomenon from primitive conditions. Another interesting finding is that semiparametric empirical likelihood inference does not require undersmoothing for the bandwidth parameter. In contrast, conventional (or bootstrap) inference based on the estimator or t-ratio typically requires undersmoothing.

The second approach, called jackknife empirical likelihood (Jing, Yuan and Zhou, 2009, Matsushita and Otsu, 2017), uses so-called jackknife pseudo-values as a moment function to construct the empirical likelihood. In the jackknife method (Quenouille, 1956, and Shao and Tu, 1995, for a review), the jackknife (bias-corrected) estimator and variance estimator are given by the sample average and variance of the pseudo-values, respectively. Therefore, the jackknife pseudo-values may be treated as if they are sample observations (Tukey, 1958). Jing, Yuan and Zhou (2009) employed this idea to construct the empirical likelihood and applied it to one- and two-sample U-statistics. We note that their results are confined to U-statistics with fixed kernels and do not cover statistics with varying kernels because of smoothing parameters. The general theory of jackknife empirical likelihood for semiparametric estimators is developed by Matsushita and Otsu (2017). This paper applies their general results to weighted average derivatives and treatment effects and confirms Wilks' phenomena in these contexts (i.e., convergence of the jackknife empirical likelihood statistics to the chi-squared distribution).

The contributions described so far are applications of the general theory of semiparametric and jackknife empirical likelihood methods to important econometric problems. Another contribution of this paper is to generalize the existing delete-1 jackknife empirical likelihood method to the delete-d version, where d grows with the sample size, and to study its general asymptotic property. It is known that the delete-1 jackknife variance estimate may be inconsistent for non-smooth objects, such as sample quantiles and quantile average derivatives or treatment effects. Shao and Wu (1989) tackled this problem and showed that the delete-d jackknife can recover the consistency of the variance estimate. We establish an analogous result for the delete-d jackknife empirical likelihood and characterize a trade-off between the smoothness of the estimator of interest and the growth rate of d. Intuitively, the less smooth the estimator is, the more we delete.

This paper is organized as follows. In Section 2, we consider weighted average derivatives. After introducing the basic setup in Section 2.1, Sections 2.2 and 2.3 discuss the semiparametric and jackknife empirical likelihood methods, respectively. Section 3 discusses the semiparametric and jackknife empirical likelihood methods for the average treatment effect. Section 4 outlines the general theory of the delete-d jackknife empirical likelihood. In Section 4.1, we mention some applications to quantile average derivatives and treatment effects. In Section 5, we report Monte Carlo simulation results. Finally, Section 6 concludes.

2. Average derivative

2.1. **Setup.** The setup for this section is introduced as follows. Suppose we observe an independent and identically distributed (iid) sample $\{Y_i, X_i'\}_{i=1}^n$ of (Y, X'), where Y is a scalar dependent variable and X is a k-dimensional vector of continuously distributed explanatory

variables.[1] Let $m(x) = E[Y|X = x]$ be the conditional mean or regression function and $\nabla m(x) = (\partial m(x)/\partial x^{(1)}, \ldots, \partial m(x)/\partial x^{(k)})'$ be its partial derivatives. In this section, we are interested in the weighted average derivative:

$$\theta = E[w(X)\nabla m(X)], \tag{2.1}$$

where w is a known scalar weight function.

The object θ appears in various contexts in empirical studies. As a popular example, consider the single index model $P\{Y = 1|X = x\} = G(x'\beta)$ for the binary dependent variable. If the function G is known (e.g., the probit or logit), then the parameters β can be estimated by the method of maximum likelihood. However, if G is unknown to the researcher (i.e., the model is semiparametric), we cannot implement maximum likelihood estimation. In this case, by noting that $m(x) = G(x'\beta)$, the average derivative in (2.1) can be expressed as

$$\theta = E[w(X)\nabla G(X'\beta)]\beta,$$

where ∇G is the derivative of G. Therefore, θ is proportional to β (note: $E[w(X)\nabla G(X'\beta)]$ is scalar). Since β is identified only up to scale, the above expression can be used as a basis to construct an estimator for the slope parameters β.

As another example, consider the nonseparable model $Y = g(X, U)$, where X and a vector of unobserved variables U are independent. In this case, the average derivative θ may be expressed as

$$\theta = E[w(X)\nabla_1 g(X, U)],$$

where $\nabla_1 g(x, u) = (\partial g(x, u)/\partial x^{(1)}, \ldots, \partial g(x, u)/\partial x^{(k)})'$ is a vector of the partial derivatives with respect to x. Thus, θ is interpreted as the weighted marginal effect of X averaged over X and U.[2]

In order to estimate θ, we introduce an alternative representation of (2.1). Let f be the probability density function of X. Under certain smoothness conditions (see Assumption A (i) below), an application of multivariate integration by parts yields

$$\begin{aligned} \theta &= \int \nabla m(x)\{w(x)f(x)\}dx = -\int m(x)\{f(x)\nabla w(x) + w(x)\nabla f(x)\}dx \\ &= E[Ys(X)], \end{aligned} \tag{2.2}$$

where $s(x) = -\nabla w(x) - w(x)\frac{\nabla f(x)}{f(x)}$. This alternative representation suggests that the average derivative θ can be estimated by the sample average using nonparametric estimates of the function $s(\cdot)$, that is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{s}(X_i), \tag{2.3}$$

[1] If X contains discrete variables such as dummies, the expectations below are understood as conditional expectations for each category of the discrete variables.

[2] Stoker (1989) proposed various tests for functional forms of $m(x)$, such as homogeneity, additivity, and symmetry of derivatives, based on the average first and second derivatives of $m(x)$. Our empirical likelihood approach can be extended to test such hypotheses.

where $\hat{s}(\cdot)$ is a sample counterpart of $s(\cdot)$ given by

$$\hat{s}(x) = -\nabla w(x) - w(x)\frac{\nabla \hat{f}(x)}{\hat{f}(x)}, \tag{2.4}$$

$\hat{f}(x) = \frac{1}{nb^k}\sum_{i=1}^n K\left(\frac{x-X_i}{b}\right)$ is the nonparametric kernel density estimator of $f(x)$ with the (differentiable) kernel function $K(\cdot)$ and the bandwidth b, and $\nabla\hat{f}(x)$ is the vector of its partial derivatives with respect to x.

The average derivative has been studied extensively in the literature of semiparametric econometrics and statistics (e.g., Stoker, 1986, Härdle and Stoker, 1989, Härdle *et al.*, 1992, Newey and Stoker, 1993, and Horowitz and Härdle, 1996), and has been applied in various empirical studies (e.g., Stoker (1989) for cost functions, Härdle, Hildenbrand and Jerison (1991) for demand analysis, Deaton and Ng (1998) for the effect of a tax and subsidy policy change, and Coppejans and Sieg (2005) for nonlinear pricing in labor markets). One popular choice for the weight function is $w(x) = f(x)$ (called the density weighted average derivative), which implies $s(x) = -2\nabla f(x)$ and a simple estimator $\hat{\theta} = -\frac{2}{n}\sum_{i=1}^n Y_i\nabla\hat{f}(X_i)$. This estimator was studied in detail by Powell, Stock and Stoker (1989) and Cattaneo, Crump and Jansson (2010).

2.2. **Semiparametric empirical likelihood.** We first introduce the semiparametric empirical likelihood approach by Bravo, Escanciano and van Keilegom (2015) and Matsushita and Otsu (2016). To motivate this approach, let us begin with a naive application of the conventional empirical likelihood approach. Suppose the derivative $\nabla m(\cdot)$ is known. Then the empirical likelihood function for $\theta = E[w(X)\nabla m(X)]$ is constructed as

$$\ell(\theta) = -2\sup_{\{p_i\}_{i=1}^n}\left\{\sum_{i=1}^n \log(np_i) :\ p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i\{w(X_i)\nabla m(X_i) - \theta\} = 0\right\},$$

which converges to the chi-squared distribution under mild regularity conditions. Based on this result, it seems reasonable to consider a feasible version of $\ell(\theta)$ by replacing $\nabla m(\cdot)$ with a nonparametric estimate $\nabla\hat{m}(\cdot)$. The empirical likelihood function with nonparametric plug-in estimates was studied in Hjort, McKeague and van Keilegom (2009). In particular, they showed that this plug-in version converges to a weighted chi-squared distribution in general, where the weights involve unknown nonparametric objects to be estimated. In other words, the plug-in empirical likelihood is not asymptotically pivotal and computation of the critical values requires an additional estimation step. Obviously the major reason for the lack of pivotalness is the presence of estimation error for $\nabla m(\cdot)$ that will inflate the sampling variation in the moment function, $w(X_i)\nabla\hat{m}(X_i) - \theta$.

The above consideration motivates us to modify the moment function to accommodate the whole sampling variation in the sample moment $n^{-1}\sum_{i=1}^n w(X_i)\nabla\hat{m}(X_i) - \theta$ (or equivalently $\hat{\theta} - \theta$). This idea has been investigated in the literature for different examples (e.g., Bertail, 2006, Zhu and Xue, 2006, and Xue and Xue, 2011). Recently Bravo, Escanciano and van Keilegom (2015) have established a general theory to correct the moment function by utilizing the pathwise derivative with respect to the nonparametric component.

We apply this approach to the weighted average derivatives. For the estimator in (2.3), it is known that

$$\hat{\theta} - \theta = \frac{1}{n}\sum_{i=1}^{n} \psi_i(\theta) + o_p(n^{-1/2}),$$

under certain regularity conditions (e.g., Stoker, 1986, and Newey and Stoker, 1993), where

$$\psi_i(\theta) = w(X_i)\nabla m(X_i) - \theta + s(X_i)\{Y_i - m(X_i)\}. \tag{2.5}$$

Indeed $\psi_i(\theta)$ is the efficient score function for θ because the variance $E[\psi_i(\theta)\psi_i(\theta)']$ equals the semiparametric efficiency bound of θ. Let $\hat{m}(x) = \frac{1}{\hat{f}(x)}\frac{1}{nb^k}\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{b}\right)$ be the nonparametric kernel regression estimator of $m(x)$. The sample counterpart of the efficient score $\psi_i(\theta)$ is given by

$$\hat{\psi}_i(\theta) = w(X_i)\nabla \hat{m}(X_i) - \theta + \hat{s}(X_i)\{Y_i - \hat{m}(X_i)\}. \tag{2.6}$$

By using this counterpart as the moment function for θ, we propose the following empirical likelihood function

$$\ell_S(\theta) = -2 \sup_{\{p_i\}_{i=1}^n} \left\{ \sum_{i=1}^{n} \log(np_i) : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{\psi}_i(\theta) = 0 \right\}. \tag{2.7}$$

Intuitively we add the correction term $\hat{s}(X_i)\{Y_i - \hat{m}(X_i)\}$ in (2.6) to construct the empirical likelihood. The definition in (2.7) involves optimization for n-variables $\{p_1, \ldots, p_n\}$ and is less practical. However, by applying the Lagrange multiplier method, we can obtain its dual form:

$$\ell_S(\theta) = 2 \sup_{\lambda} \sum_{i=1}^{n} \log(1 + \lambda' \hat{\psi}_i(\theta)), \tag{2.8}$$

which involves optimization only for k-variables λ. In practice, we compute $\ell_S(\theta)$ using this dual form. To study the asymptotic property of $\ell_S(\theta)$, we impose the following assumptions.

Assumption A.

(i): *$\{Y_i, X_i'\}_{i=1}^n$ is an iid sample from $(Y, X') \in \mathbb{R} \times \mathbb{X}$, where $\mathbb{X} \subset \mathbb{R}^k$ is compact. For some $p \geq 2$, $E[|Y|^p | X = x]$ is bounded and $E|Y|^p < \infty$. $E[\psi_i(\theta)\psi_i(\theta)']$ is positive definite, and it holds $\inf_{x:w(x)>0} f(x) > 0$. $w(x)$ is known, bounded, and continuously differentiable. For some $s \geq 2$, $f(x)$ is $(s+1)$ times differentiable, and $f(x)$ and its first $(s+1)$ derivatives are bounded and continuous. $m(x)$ is continuously differentiable, and $f(x)m(x)$ and its first derivative are bounded.*

(ii): *$K(u)$ is even, bounded, and twice differentiable with bounded derivatives, and satisfies $\int K(u)du = 1$, $\int K(u)u_1^{j_1} \cdots u_k^{j_k} du = 0$ for any vector of non-negative integers $(j_1, \ldots, j_k)$ such that $j_1 + \cdots + j_k < s$ and $\int |K(u)|(1 + |u|^s)du < \infty$. Also, $\int |\dot{K}(u)|du < \infty$, $\int |\bar{K}(u)|du < \infty$, where $\dot{K}(u) = \partial K(u)/\partial u$ and $\bar{K}(u) = \sup_{|r| \geq u} |\partial(K(r), \dot{K}(r)')/\partial r|$, respectively.*

These assumptions are mild and standard in the literature (see, e.g., Cattaneo, Crump and Jansson, 2013). The asymptotic distribution of the empirical likelihood statistic $\ell_S(\theta)$ is presented as follows. The proof is given in Appendix A.1.

Proposition 1. *Consider the setup of this section and impose Assumption A. Suppose* $\sqrt{\frac{\log n}{nb^{k+2}}} = o(n^{-1/6})$ *and* $b^s = o(n^{-1/4})$. *Then*

$$\ell_S(\theta) \xrightarrow{d} \chi^2(k).$$

Remark 1. This proposition says that the semiparametric empirical likelihood statistic $\ell_S(\theta)$ is asymptotically pivotal and converges to the $\chi^2(k)$ distribution. Based on this proposition, the $100(1-\alpha)\%$ asymptotic confidence set for θ is constructed as $ELCS_\alpha = \{\theta : \ell_S(\theta) \leq \chi^2_{1-\alpha}(k)\}$, where $\chi^2_{1-\alpha}(k)$ is the $(1-\alpha)$-th quantile of the $\chi^2(k)$ distribution. This property of asymptotic pivotalness is particularly attractive in our setup because the asymptotic variance of the average derivative estimator $\hat{\theta}$ takes a complicated form due to the influence from the nonparametric estimation of the density f and its derivative. Although we can express the asymptotic variance of $\hat{\theta}$ based on the influence function in (2.5), whether we can precisely estimate the asymptotic variance so that the resulting t-ratio is reliable for inference on θ is another problem entirely. In contrast, our empirical likelihood statistic $\ell_S(\theta)$ is internally studentized and circumvents such asymptotic variance estimation.

Remark 2. When we are concerned with the slope parameters β in the binary choice model $P\{Y = 1|X = x\} = G(x'\beta)$, we need to introduce a normalization on θ (e.g., the first element of θ equals 1 or $|\theta| = 1$). For example, if we normalize $\theta = (1, \vartheta')'$, then the empirical likelihood (ratio) statistic for ϑ can be obtained as $L_S(\vartheta) = \ell_S(1, \vartheta) - \min_\vartheta \ell_S(1, \vartheta)$. By applying a similar argument to Smith (1997), we can show that $L_S(\vartheta)$ converges to the $\chi^2(d-1)$ distribution.

Remark 3. If we are interested in the confidence set for some element of θ (say, the j-th element θ_j), our empirical likelihood statistic $L_S(\theta_j)$ can be obtained by replacing $\hat{\psi}_i(\theta)$ in (2.8) with

$$\tilde{\psi}_i(\theta_j) = w(X_i)\nabla_j \hat{m}(X_i) - \theta_j + \hat{s}_j(X_i)\{Y_i - \hat{m}(X_i)\},$$

where $\hat{s}_j(x) = -\nabla_j w(x) - w(x)\frac{\nabla_j \hat{f}(x)}{\hat{f}(x)}$ and "∇_j" means the derivative with respect to the j-th element of x. By an analogous argument, we can show that $L_S(\theta_j) \xrightarrow{d} \chi^2(1)$, and the confidence set for θ_j is given by $\{\theta_j : L_S(\theta_j) \leq \chi^2_{1-\alpha}(1)\}$. We note that in this case, the Lagrange multiplier λ to compute $L_S(\theta_j)$ is scalar, and the computational cost is cheaper than the vector case.

Remark 4. We note that the conditions on the bandwidth b to compute $\hat{f}$ and $\nabla\hat{f}$ do not require undersmoothing, i.e., we only require $nb^{4s} \to 0$ instead of $nb^{2s} \to 0$. Thus, for example, the MSE optimal bandwidth is allowed. This desirable property is known in the empirical likelihood literature for several setups (e.g., Zhu and Xue, 2006, Bravo, Escanciano and van Keilegom, 2015). Proposition 1 shows that a similar result holds for the present setup. Intuitively, the main term (i.e., $w(X_i)\nabla\hat{m}(X_i) - \theta$) and the adjustment term (i.e., $s(X_i)\{Y_i - \hat{m}(X_i)\}$) in (2.6) share the same form for the smoothing bias and these bias terms are automatically cancelled out. We emphasize that in contrast to the empirical likelihood confidence set $ELCS_\alpha$, the Wald-type (or t-ratio-based) confidence set using the asymptotic variance estimator based on the efficient score function in (2.5) requires undersmoothing for the bandwidth.

Remark 5. Another interesting finding is that the condition $\sqrt{\frac{\log n}{nb^{k+2}}} = o(n^{-1/6})$ on the upper bound of the decay rate of the bandwidth is also weaker than the conventional requirement $\sqrt{\frac{\log n}{nb^{k+2}}} = o(n^{-1/4})$. This point is clarified by Rothe and Firpo (2016) in the context of doubly-robust estimators satisfying certain orthogonality conditions. In our setup, the general result of Rothe and Firpo (2016) implies that the rate $o(n^{-1/6})$ is sufficient for the asymptotic normality of $\hat{\theta}$ because the second order variance term has a smaller order. We find that the same result applies to our semiparametric empirical likelihood statistic.

Remark 6. Matsushita and Otsu (2016) extended the semiparametric empirical likelihood approach to the semiparametric three-step estimators considered in Hahn and Ridder (2013). In the present setup, their method can be applied to the case where some elements of X are generated (or estimated) variables. In this case, we need to introduce an additional correction term to the moment function $\psi_i(\theta)$ to recover the asymptotic pivotalness.

2.3. **Jackknife empirical likelihood.** We next consider an alternative inference approach based on the jackknife empirical likelihood. To begin with, we introduce the conventional jackknife method. Let $\hat{\theta}$ be some estimator of θ and $\hat{\theta}^{(-i)}$ be its leave-i-out version, i.e., the estimator computed by the sample without the i-th observation. Then the jackknife bias estimator for $\hat{\theta}$ is given by $(n-1)(\bar{\theta} - \hat{\theta})$ with $\bar{\theta} = n^{-1}\sum_{i=1}^{n} \hat{\theta}^{(-i)}$. By subtracting this bias estimate, the bias-corrected estimator is written as

$$\tilde{\theta} = \hat{\theta} - \bar{\theta} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}_i,$$

where $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)}$. The object $\tilde{\theta}_i$, called the jackknife pseudo-value (Tukey, 1958), may be interpreted as an iid copy of $\hat{\theta}$. By using these pseudo-values, the jackknife estimate for the variance of $\hat{\theta}$ is obtained by $(n-1)^{-1}\sum_{i=1}^{n}(\tilde{\theta}_i - \tilde{\theta})^2$. See Shao and Tu (1995) for a review on the jackknife method.

These ideas of the jackknife estimate and its variance estimate suggest that the moment functions for empirical likelihood may be constructed by those pseudo-values. Based on the average derivative estimator $\hat{\theta}$ defined in (2.3), we consider the following jackknife pseudo-value

$$\hat{\zeta}_i(\theta) = n(\hat{\theta} - \theta) - (n-1)(\hat{\theta}^{(-i)} - \theta), \tag{2.9}$$

where $\hat{\theta}^{(-i)}$ is the leave-i-out version of $\hat{\theta}$, that is

$$\hat{\theta}^{(-i)} = \frac{1}{n-1}\sum_{j\neq i}^{n} Y_j \hat{s}^{(-i)}(X_j),$$

and $\hat{s}^{(-i)}(x)$ is defined as in (2.4) but using the leave-i-out kernel density estimator $\hat{f}^{(-i)}(x) = \frac{1}{(n-1)b^k}\sum_{j\neq i}^{n} K\left(\frac{x-X_j}{b}\right)$. By utilizing this jackknife pseudo-value as our moment function for θ, we propose the jackknife empirical likelihood function

$$\ell_J(\theta) = -2 \sup_{\{p_i\}_{i=1}^n} \left\{ \sum_{i=1}^{n} \log(np_i) : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{\zeta}_i(\theta) = 0 \right\}. \tag{2.10}$$

By applying the Lagrange multiplier method, the dual form of $\ell_J(\theta)$ is written as

$$\ell_J(\theta) = 2 \sup_{\lambda} \sum_{i=1}^{n} \log(1 + \lambda' \hat{\zeta}_i(\theta)). \tag{2.11}$$

In practice, we compute $\ell_J(\theta)$ by using this dual form. The asymptotic distribution of the empirical likelihood ratio is presented as follows.

Proposition 2. *Consider the setup of this section and impose Assumption A. Suppose* $\sqrt{\frac{\log n}{nb^{k+2}}} = o(n^{-1/4})$ *and* $b^s = o(n^{-1/2})$. *Then*

$$\ell_J(\theta) \xrightarrow{d} \chi^2(k).$$

The proof is similar to that of the delete-d jackknife empirical likelihood in Proposition 4 below. See also Matsushita and Otsu (2017) for details.

Remark 7. Similar to the semiparametric empirical likelihood, the jackknife empirical likelihood statistic is also asymptotically pivotal and converges to the $\chi^2(k)$ distribution. The $100(1-\alpha)\%$ asymptotic confidence set is obtained by $\{\theta : \ell_J(\theta) \leq \chi^2_{1-\alpha}(k)\}$. We can also show that both the semiparametric and jackknife empirical likelihood statistics are asymptotically equivalent and have the same local power function. However, we should note that Proposition 2 is obtained under the assumption of undersmoothing (i.e., $nb^{2s} \to 0$). This is due to the fact that the moment function $\hat{\zeta}_i(\theta)$ for the jackknife empirical likelihood does not result in a cancellation of the bias terms as in the semiparametric empirical likelihood. This is considered as a drawback of the jackknife empirical likelihood. On the other hand, in Matsushita and Otsu (2017), we show that a modification of the jackknife empirical likelihood achieves a desirable robustness property for small bandwidths.

3. Treatment effect

In this section, we consider inference on the average treatment effect. Let $Y_i(1)$ and $Y_i(0)$ denote potential outcomes of unit i with and without exposure to a treatment, respectively. Let $D_i \in \{0, 1\}$ be an indicator variable for the treatment such that $D_i = 1$ if unit i is exposed to the treatment and $D_i = 0$ otherwise. We observe $Z_i = (Y_i, X_i', D_i)'$, where $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ is the observable outcome, and X_i is a k-dimensional vector of covariates. We are interested in the average treatment effect $\tau = E[Y_i(1) - Y_i(0)]$.

Under the so-called unconfoundedness assumption (i.e., $Y(1)$ and $Y(0)$ are independent of D, conditional on X), the average treatment effect can be identified as

$$\tau = E\left[\frac{YD}{\varphi(X)} - \frac{Y(1-D)}{1-\varphi(X)}\right],$$

where $\varphi(x) = P\{D = 1 | X = x\}$ is the propensity score (see, Rosenbaum and Rubin, 1983). Based on this expression, τ may be estimated as

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i D_i}{\hat{\varphi}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{\varphi}(X_i)}\right], \tag{3.1}$$

where $\hat{\varphi}(x) = \frac{1}{\hat{f}(x)} \frac{1}{nb^k} \sum_{i=1}^n D_i K\left(\frac{x-X_i}{b}\right)$ is the kernel estimator of $\varphi(x)$ with $\hat{f}(x) = \frac{1}{nb^k} \sum_{i=1}^n K\left(\frac{x-X_i}{b}\right)$. Hirano, Imbens and Ridder (2003) studied the asymptotic properties of $\hat{\tau}$.

Based on the influence function of $\hat{\tau}$, Bravo, Escanciano and van Keilegom (2015) investigated the semiparametric empirical likelihood statistic as in (2.7) with the moment function

$$\hat{\psi}_i^{ATE}(\theta) = \frac{Y_i D_i}{\hat{\varphi}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{\varphi}(X_i)} - \theta - \{D_i - \hat{\varphi}(X_i)\}\hat{q}(X_i),$$

where

$$\hat{q}(X_i) = \frac{\hat{\mu}_1(X_i)}{\hat{\varphi}(X_i)} - \frac{\hat{\mu}_0(X_i)}{1-\hat{\varphi}(X_i)},$$

and $\hat{\mu}_1(x) = \frac{1}{\hat{f}(x)} \frac{1}{nb^k} \sum_{i=1}^n D_i Y_i K\left(\frac{x-X_i}{b}\right)$ and $\hat{\mu}_0(x) = \frac{1}{\hat{f}(x)} \frac{1}{nb^k} \sum_{i=1}^n (1-D_i) Y_i K\left(\frac{x-X_i}{b}\right)$ are the kernel estimators of $E[Y|X, D=1]$ and $E[Y|X, D=0]$, respectively. Bravo, Escanciano and van Keilegom (2015, Proposition E2) showed that under mild regularity conditions, it holds $\ell_S(\tau) \xrightarrow{d} \chi^2(1)$.

Here we focus on the jackknife empirical likelihood approach. Based on the average treatment effect estimator $\hat{\tau}$ defined in (3.1), we consider the jackknife pseudo-value

$$\hat{\zeta}_i^{ATE}(\theta) = n(\hat{\tau} - \tau) - (n-1)(\hat{\tau}^{(-i)} - \tau),$$

where $\hat{\tau}^{(-i)}$ is the leave-i-out version of $\hat{\tau}$, that is

$$\hat{\tau}^{(-i)} = \frac{1}{n-1} \sum_{j \neq i}^n \left[\frac{Y_j D_j}{\hat{\varphi}^{(-i)}(X_j)} - \frac{Y_j(1-D_j)}{1-\hat{\varphi}^{(-i)}(X_j)} \right],$$

and $\hat{\varphi}^{(-i)}(x)$ is a leave-i-out version of $\hat{\varphi}(x)$. Then the jackknife empirical likelihood function is defined as in (2.10) and its asymptotic property is obtained as follows.

Assumption B.

- **(i):** *$\{Y_i, D_i, X_i'\}_{i=1}^n$ is an iid sample from $(Y, D, X') \in \mathbb{R} \times \{0,1\} \times \mathcal{X}$, where $Y = DY(1) + (1-D)Y(0)$ and $(Y(1), Y(0)) \perp D|X$. $f(x)$ (the density function of X), $\varphi(x)$, and $\frac{\mu_1(x)}{\varphi(x)} - \frac{\mu_0(x)}{1-\varphi(x)}$ are s times continuously differentiable with bounded derivatives, and $\inf_{x \in \mathcal{X}} f(x) \geq c > 0$. $\hat{\varphi}(\cdot)$, $\hat{\mu}_1(\cdot)$, and $\hat{\mu}_0(\cdot)$ are uniformly consistent over $\mathcal{X}$.*
- **(ii):** *$K(u)$ is even, bounded, and satisfies $\int K(u)du = 1$, $\int K(u) u_1^{j_1} \cdots u_k^{j_k} du = 0$ for any vector of non-negative integers $(j_1, \ldots, j_k)$ such that $j_1 + \cdots + j_k < s$, $\int |K(u)|(1 + |u|^s)du < \infty$. There exist $C, L > 0$ and $v > 1$ such that $|K(u)| \leq C|u|^{-v}$ for all $|u| > L$.*

Proposition 3. *Consider the setup of this section and impose Assumption B. Suppose $\sqrt{\frac{\log n}{nb^k}} = o(n^{-1/4})$ and $b^s = o(n^{-1/2})$. Then*

$$\ell_J(\theta) \xrightarrow{d} \chi^2(k).$$

Similar comments to Propositions 1 and 2 apply. Assumption B is analogous to that of Bravo, Escanciano and van Keilegom (2015, Proposition E2) except for the undersmoothing condition $nb^{2s} \to 0$. Their semiparametric empirical likelihood requires only $nb^{4s} \to 0$. The proof is similar to that of the delete-d jackknife empirical likelihood in Proposition 4 below.

Here we present the jackknife empirical likelihood method based on the estimator $\hat{\tau}$ in (3.1). We expect that a similar result can be obtained for other estimators, such as the the propensity score matching estimator by Heckman, Ichimura and Todd (1998).

4. Delete-d jackknife empirical likelihood: General theory

In this section, we develop a general theory for the delete-d jackknife empirical likelihood. This is a novel extension of the (delete-one) jackknife empirical likelihood by Jing, Yuan and Zhou (2009) to more general setups, and is considered a natural counterpart of the delete-d jackknife method (Shao and Wu, 1989).

We first introduce some notation. Take any estimator $\hat{\theta}$ for a k-vector of parameters θ. Assume that $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is invariant under permutation of the arguments. Let d be an integer less than n, and $\mathcal{S}_{n,d}$ be the collection of subsets of $\{1, \ldots, n\}$ with size $n - d$. For each $s = \{i_1, \ldots, i_{n-d}\} \in \mathcal{S}_{n,d}$, let $\hat{\theta}_s = \hat{\theta}(X_{i_1}, \ldots, X_{i_{n-d}})$ be a leave-d-out counterpart of $\hat{\theta}$, and "$\sum_s$" mean the summation over $s \in \mathcal{S}_{n,d}$. Note that $\mathcal{S}_{n,d}$ has $N = \binom{n}{d}$ elements.

Based on the above notation, the delete-d jackknife variance estimator is defined as

$$V_d = \frac{n-d}{dN} \sum_s (\hat{\theta}_s - \hat{\theta})(\hat{\theta}_s - \hat{\theta})'. \tag{4.1}$$

It is known that the delete-1 jackknife variance estimator V_1 is consistent for sufficiently smooth estimators. On the other hand, if the estimator is not smooth, V_1 may be an inconsistent estimator of the variance of $\hat{\theta}$ (see, Miller, 1974). The most popular example of failure of the delete-1 jackknife is the sample quantile. If θ is scalar, we typically have (see, Shao and Wu, 1989, pp. 1176-1177)

$$nV_1 \xrightarrow{d} \frac{\sigma^2}{4}\xi^2,$$

where σ^2 is the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$ and $\xi \sim \chi^2(2)$. Therefore, v_1 is an inconsistent estimator of σ^2. For this problem, Shao and Wu (1989) showed that the delete-d jackknife variance estimator with diverging d (but slower than n) may recover consistency for σ^2 and characterized a trade-off between smoothness of the estimator and growth rate of d.

In this section, we introduce and study a delete-d version of the jackknife empirical likelihood approach. Define the delete-d jackknife pseudo value as

$$\tilde{\zeta}_s(\theta) = (\hat{\theta} - \theta) + \frac{1}{d}\sqrt{(n-d)(N-d)}\varepsilon_s(\hat{\theta} - \hat{\theta}_s), \tag{4.2}$$

where $\varepsilon_s = +1$ with probability 0.5 and -1 otherwise. The perturbation ε_s is introduced to remove correlations of the second terms in (4.2) across s. Note that when $d = 1$, the delete-d pseudo value $\tilde{\zeta}_s(\theta)$ reduces to the delete-1 version in (2.9) except for the perturbation. Based on these pseudo values, (the dual form of) the delete-d jackknife empirical likelihood is defined as

$$\tilde{\ell}_J(\theta) = \frac{2}{d} \sup_\lambda \sum_s \log(1 + \lambda' \tilde{\zeta}_s(\theta)). \tag{4.3}$$

For the estimator $\hat{\theta}$, we impose the following assumptions.

Assumption D. *Suppose the estimator $\hat{\theta}$ admits the expansion*

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \phi_i + R_n, \tag{4.4}$$

where $\{\phi_i\}$ is an iid sequence with mean zero and finite variance Ω. Also the remainders R_n and $R_{n,s}$ for $\hat{\theta}_s$ satisfy

$$\frac{n(n-d)}{d}|Var(R_n - R_{n,s})| \to 0. \tag{4.5}$$

Finally, $\sqrt{\frac{n(n-d)}{N-d}}\max_s |R_n - R_{n,s}| \xrightarrow{p} 0$.

The assumption for the expansion in (4.4) is mild and typically satisfied for $\sqrt{n}$-consistent estimators. Also, since (4.5) implies $R_n = o_p(n^{-1/2})$ (Shao and Wu, 1989, Lemma 1), the central limit theorem guarantees that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $N(0, \Omega)$. The condition in (4.5) is a key for characterizing the trade-off between the smoothness of the estimator and the growth rate of d. The same condition is employed by Shao and Wu (1989, eq. (3.4)). Intuitively, if the estimator is less smooth, then the remainder component $|E[R_n R_n']|$ tends to be of larger order (or slower decay) and we need d to grow faster so that (4.5) is guaranteed.

Shao and Wu (1989) provided various results and examples to verify the condition in (4.5). For example, if the estimator is sufficiently smooth (e.g., the functional T to define the estimator $\hat{\theta} = T(F_n)$ for the empirical distribution F_n is Fréchet differentiable), then it typically holds that $|Var(R_n - R_{n,s})| = o(n^{-2})$ and the condition in (4.5) is satisfied even if d is bounded. Thus, for sufficiently smooth $\hat{\theta}$, the jackknife variance estimator V_d is consistent even for fixed d. Also if $\hat{\theta}$ is the sample quantile, then by Duttweiler (1973), the remainder decays more slowly and we can obtain $|E[R_n R_n']| = O(n^{-3/2})$. In this case, the condition in (4.5) is satisfied if d diverges faster than $\sqrt{n}$. The last condition in Assumption D is a mild requirement for the remainder to control maximal deviations of the pseudo values.

The asymptotic property of the delete-d jackknife empirical likelihood statistic is presented as follows.

Proposition 4. *Consider the setup of this section. Under Assumption D, it holds*

$$\tilde{\ell}_J(\theta) \xrightarrow{d} \chi^2(k).$$

4.1. **Discussion: Quantile-based methods.** As the sample quantile example suggests, the delete-d jackknife empirical likelihood would be useful to deal with non-smooth objects, especially quantile-based parameters. In this subsection, we mention two examples: quantile average derivative (Chaudhuri, Doksum and Samarov, 1997) and quantile treatment effect (Firpo, 2007). Although formal analyses require a separate paper, we expect that the semiparametric and delete-d jackknife empirical likelihood methods provide valid inference procedures.

First, the average derivative for the conditional quantile function is defined as in (2.1) by replacing $m(x)$ with the conditional (τ-th) quantile function $m_\tau(x) = Q_\tau(Y|X = x)$. By using some nonparametric estimator $\hat{m}_\tau$ for m_τ and the integration by parts formula in (2.2), the

parameter $\theta_\tau^D = E[w(X)\nabla m_\tau(X)]$ may be estimated by

$$\hat{\theta}_\tau^D = -\frac{1}{n}\sum_{i=1}^{n} \hat{m}_\tau(X_i)\hat{s}(X_i).$$

Based on Chaudhuri, Doksum and Samarov (1997), the efficient score function for θ_τ^D is written as

$$\psi_i^D(\theta) = w(X_i)\nabla m_\tau(X_i) - \theta + s(X_i)\frac{\tau - \mathbb{I}\{Y_i \le m_\tau(X_i)\}}{f_{Y|X}(m_\tau(X_i)|X_i)}.$$

In this case, the semiparametric empirical likelihood can be constructed as in (2.7) with the sample counterpart of this score function. Also the delete-d jackknife empirical likelihood is defined as in (4.3) by using $\hat{\theta}_\tau^D$.

Next, our approach may also be applied to quantile treatment effects. Let $q_{1,\tau} = \inf_q \Pr\{Y_i(1) \le q\}$ and $q_{0,\tau} = \inf_q \Pr\{Y_i(0) \le q\}$ be the τ-th quantiles of the potential outcomes $Y_i(1)$ and $Y_i(0)$, respectively. The (τ-th) quantile treatment effect is defined as

$$\theta_\tau^{QTE} = q_{1,\tau} - q_{0,\tau}.$$

Based on Firpo (2007), the efficient score function for θ_τ^{QTE} is written as

$$\begin{aligned}\psi_i^{QTE}(\theta) \;=\; & \theta + \frac{D_i}{\varphi(X_i)}\cdot\frac{\tau - \mathbb{I}\{Y_i \le q_{1,\tau}\}}{f_1(q_{1,\tau})} - \frac{D_i - \varphi(X_i)}{\varphi(X_i)}\cdot\frac{\tau - E[\mathbb{I}\{Y_i \le q_{1,\tau}\}|X_i, D_i = 1]}{f_1(q_{1,\tau})} \\ & - \frac{1 - D_i}{1 - \varphi(X_i)}\cdot\frac{\tau - \mathbb{I}\{Y_i \le q_{0,\tau}\}}{f_0(q_{0,\tau})} - \frac{D_i - \varphi(X_i)}{1 - \varphi(X_i)}\cdot\frac{\tau - E[\mathbb{I}\{Y_i \le q_{0,\tau}\}|X_i, D_i = 0]}{f_0(q_{0,\tau})}.\end{aligned}$$

The semiparametric empirical likelihood can be constructed as in (2.7) with the sample counterpart of this score function. Also the delete-d jackknife empirical likelihood is defined as in (4.3) by using the quantile treatment effect estimator $\hat{\theta}_\tau^{QTE}$ by Firpo (2007).

5. Simulation

This section conducts a simulation study to evaluate the finite sample properties of the semiparametric and jackknife empirical likelihood inference methods. We focus on the weighted average derivative and adopt the simulation designs considered in Cattaneo, Crump and Jansson (2013).

In particular, we consider a Tobit model $Y_i = \tilde{Y}_i\mathbb{I}\{\tilde{Y}_i \ge 0\}$ with $\tilde{Y}_i = X_i\beta + \epsilon_i$, $\epsilon_i \sim_{iid} N(0,1)$, and $X_i \sim_{iid} N(0,1)$. We are interested in $\theta = \beta E[w(X)\Phi(X\beta)]$, where $\Phi(\cdot)$ is the standard normal distribution function and the weight function is set as

$$w(x) = \exp\left(-\frac{x^4}{\tau^4(\tau^4 - x^4)}\right)\mathbb{I}\{|x| < \tau\},$$

with the trimming constant $\tau = \Phi^{-1}(0.825)$. We set $\beta = 1$.

We compare three methods to construct confidence intervals for θ: (i) the Wald-type confidence interval (Wald), (ii) the semiparametric empirical likelihood confidence interval (SPEL), and (iii) the jackknife empirical likelihood confidence interval (JEL). We report results implemented by the Gaussian kernel. The sample size is set to $n = 1000$ for each replication.

Table 1 gives the actual coverage rates of all the intervals across 1,000 replications for five different fixed bandwidths constructed as $h_n = cn^{-1/(4+k)}$ with $k = 1$ and $c \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$. The nominal rate is 0.95. Wald intervals tend to under-cover in all cases. JEL intervals tend to over-cover especially when the bandwidth is small. SPEL intervals tend to slightly under-cover, but they are most robust to the choice of bandwidth compared to the other intervals.

c	Wald	SPEL	JEL
0.7	0.915	0.946	0.952
0.8	0.916	0.934	0.935
0.9	0.904	0.943	0.936
1.0	0.902	0.936	0.940
1.1	0.908	0.935	0.952
1.2	0.904	0.931	0.938
1.3	0.894	0.931	0.939

TABLE 1. Coverage probabilities of nominal 95% confidence intervals

6. CONCLUSION

In this paper, we consider semiparametric and jackknife empirical likelihood inference methods for average derivatives and treatment effects, and derive their asymptotic properties. Also, we propose the delete-d jackknife empirical likelihood and establish the general asymptotic theory. The extension to the delete-d version would be useful to deal with non-smooth objects, such as quantile average derivatives and treatment effects. Our simulation results illustrate the usefulness of our inference methods.

Appendix A. Mathematical Appendix

A.1. **Proof of Proposition 1.** Hereafter we suppress "(θ)" and denote $\psi_i = \psi_i(\theta)$ and $\hat{\psi}_i = \hat{\psi}_i(\theta)$. Also define $\Sigma = E[\psi_i \psi_i']$. Suppose

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\psi}_i \quad \xrightarrow{d} \quad N(0, \Sigma), \tag{A.1}$$

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i' \quad \xrightarrow{p} \quad \Sigma, \tag{A.2}$$

$$\max_{1 \le i \le n} |\hat{\psi}_i| \quad = \quad o_p(n^{1/2}). \tag{A.3}$$

Let $\hat{\lambda}$ be the solution of (2.8). By (A.1)-(A.3), the same argument as in the proof of Owen (1990, eq. (2.14)) implies that $\hat{\lambda} = O_p(n^{-1/2})$. The first-order condition for $\hat{\lambda}$ satisfies

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\psi}_i}{1 + \hat{\lambda}' \hat{\psi}_i} = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i' \hat{\lambda} + \frac{1}{n} \sum_{i=1}^{n} \frac{(\hat{\lambda}' \hat{\psi}_i)^2}{1 + \hat{\lambda}' \hat{\psi}_i} \hat{\psi}_i,$$

where the second equality follows from the identity $(1+x)^{-1} = 1 - x + x^2(1+x)^{-1}$. By (A.1)-(A.3) and $\hat{\lambda} = O_p(n^{-1/2})$, we have

$$\hat{\lambda} = \left[\frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i' \right]^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i \right) + o_p(n^{-1/2}).$$

Therefore, an expansion yields

$$\begin{aligned} 2 \sum_{i=1}^{n} \log(1 + \hat{\lambda}' \hat{\psi}_i) &= 2 \sum_{i=1}^{n} \left[\hat{\lambda}' \hat{\psi}_i - \frac{1}{2} (\hat{\lambda}' \hat{\psi}_i)^2 \right] + o_p(1) \\ &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\psi}_i \right)' \left[\frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i' \right]^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\psi}_i \right) + o_p(1). \end{aligned}$$

The conclusion follows by (A.1) and (A.2).

It remains to show (A.1)-(A.3). Below we provide a proof of (A.1). The result in (A.2) can be shown in the same manner. The result in (A.3) follows by a similar argument in Owen (1990, Lemma 3) using the Borel-Cantelli lemma.

Proof of (A.1). Decompose

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\psi}_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [w(X_i) \nabla \hat{m}(X_i) - \theta + \hat{s}(X_i)\{Y_i - \hat{m}(X_i)\}] \\ &= M_1 + M_2 + M_3, \end{aligned}$$

where

$$\begin{aligned} M_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[w(X_i)\nabla\hat{m}(X_i) - \theta + s(X_i)\{Y_i - \hat{m}(X_i)\}], \\ M_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\hat{s}(X_i) - s(X_i)\}\{Y_i - m(X_i)\}, \\ M_3 &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\hat{s}(X_i) - s(X_i)\}\{\hat{m}(X_i) - m(X_i)\}. \end{aligned}$$

Note that from integration by parts,

$$E[w(X_i)\nabla a(X_i) - s(X_i)a(X_i)] = 0, \tag{A.4}$$

for any vector of differentiable function $a(\cdot)$. For M_1, we denote

$$M_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i(\theta, \hat{h}),$$

where $\hat{h} = (\hat{m}, \nabla\hat{m})$ and $\eta_i(\theta, \hat{h}) = w(X_i)\nabla\hat{m}(X_i) - \theta + s(X_i)\{Y_i - \hat{m}(X_i)\}$. Since $E[\eta_i(\theta, h)] = 0$, we can decompose

$$\begin{aligned} M_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i(\theta, h) + \sqrt{n}E[\eta_i(\theta, \hat{h})] \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\eta_i(\theta, \hat{h}) - E[\eta_i(\theta, \hat{h})]\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\eta_i(\theta, h) - E[\eta_i(\theta, h)]\} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i(\theta, h) + \sqrt{n}E[\eta_i(\theta, \hat{h})] + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i(\theta, h) + o_p(1) \end{aligned}$$

where the second equality follows from the stochastic equicontinuity argument (Chen, Linton and van Keilegom, 2003) and the third equality follows from (A.4) with $a = \hat{m}$. Therefore, the central limit theorem implies $M_1 \xrightarrow{d} N(0, \Sigma)$.

Let $U_i = Y_i - m(X_i)$. For M_2, we further decompose

$$\begin{aligned} M_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(X_i)\left\{\frac{\nabla f(X_i)}{f(X_i)} - \frac{\nabla\hat{f}(X_i)}{\hat{f}(X_i)}\right\}U_i \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{w(X_i)}{f(X_i)}\{\nabla f(X_i) - \nabla\hat{f}(X_i)\}U_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(X_i)\left\{\frac{1}{f(X_i)} - \frac{1}{\hat{f}(X_i)}\right\}\nabla\hat{f}(X_i)U_i \\ &= M_{21} + M_{22}. \end{aligned}$$

For M_{21},

$$\begin{aligned} M_{21} &= -\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w(X_i)}{f(X_i)}U_i\left\{\frac{1}{b^{k+1}}K'\left(\frac{X_i-X_j}{b}\right)-\nabla f(X_i)\right\} \\ &= -\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w(X_i)}{f(X_i)}U_i\left\{\frac{1}{b^{k+1}}K'\left(\frac{X_i-X_j}{b}\right)-E\left[\frac{1}{b^{k+1}}K'\left(\frac{X_i-X_j}{b}\right)\middle| X_i\right]\right\} \\ &\quad -\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w(X_i)}{f(X_i)}U_i\left\{E\left[\frac{1}{b^{k+1}}K'\left(\frac{X_i-X_j}{b}\right)\middle| X_i\right]-\nabla f(X_i)\right\} \\ &= -\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_{ij}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{w(X_i)}{f(X_i)}U_i\left\{E\left[\frac{1}{b^{k+1}}K'\left(\frac{X_i-X_j}{b}\right)\middle| X_i\right]-\nabla f(X_i)\right\}, \end{aligned}$$

where $K'(\cdot)$ is the derivative of $K(\cdot)$ and

$$\xi_{ij}=\frac{w(X_i)}{f(X_i)}U_i\left\{\frac{1}{b^{k+1}}K'\left(\frac{X_i-X_j}{b}\right)-E\left[\frac{1}{b^{k+1}}K'\left(\frac{X_i-X_j}{b}\right)\middle| X_i\right]\right\}.$$

Note that $E[\xi_{ij}|X_j]=0$ because $E[U_i|X_i]=0$. The first term is a second-order degenerate U-statistics. So by using the variance formula (e.g., Serfling, 1980) and Chebyshev's inequality, we can show that $M_{21}=O_p\left(b^s+\frac{1}{\sqrt{nb^{k+2}}}\right)$. By a similar argument, we obtain $M_{22}=O_p\left(b^s+\frac{1}{\sqrt{nb^{k+2}}}\right)$, which implies $M_2=O_p\left(b^s+\frac{1}{\sqrt{nb^{k+2}}}\right)$.

For M_3, we decompose

$$\begin{aligned} M_3 &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{w(X_i)}{f(X_i)}\{\nabla\hat{f}(X_i)-\nabla f(X_i)\}\{\hat{m}(X_i)-m(X_i)\} \\ &\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{w(X_i)\nabla\hat{f}(X_i)}{\hat{f}(X_i)f(X_i)}\{\hat{f}(X_i)-f(X_i)\}\{\hat{m}(X_i)-m(X_i)\}, \end{aligned}$$

and the same argument as in the proof of Rothe and Firpo (2016, Lemma 5) guarantees

$$M_3=O_p\left(b^s+\sqrt{n}b^{2s}+\sqrt{n}\left(\sqrt{\frac{\log n}{nb^{k+2}}}\right)^3\right).$$

Combining these results, (A.1) is obtained.

A.2. **Proof of Proposition 4.** To simplify the presentation, we focus on the case where θ is scalar. Hereafter we denote $\tilde{\zeta}_s=\tilde{\zeta}_s(\theta)$. Suppose

$$\frac{\sqrt{n}}{N}\sum_s\tilde{\zeta}_s \xrightarrow{d} N(0,\Omega), \tag{A.5}$$

$$\frac{dn}{N(N-d)}\sum_s\tilde{\zeta}_s^2 \xrightarrow{p} \Omega, \tag{A.6}$$

By (4.4) and the triangle inequality,

$$\begin{aligned}\frac{d\sqrt{n}}{N-d}\max_s|\tilde{\zeta}_s| &\leq \frac{d}{N-d}\sqrt{n}(\hat{\theta}-\theta)+\sqrt{\frac{n(n-d)}{N-d}}\max_s|R_n-R_{n,s}| \\ &\quad +\sqrt{\frac{n(n-d)}{N-d}}\max_s\left|\frac{1}{n}\sum_{i=1}^n\phi_i-\frac{1}{n-d}\sum_{i\in s}\phi_i\right| \\ &\xrightarrow{p} 0, \end{aligned} \tag{A.7}$$

where the convergence follows from the last condition in Assumption D and $\max_{i=1,\ldots,n}|\phi_i| = o(\sqrt{n})$ by Owen (1990, Lemma 3) using $E|\phi_i|^2<\infty$. Let $\hat{\lambda}$ be the solution of (4.3). By (A.5)-(A.7), the same argument as in the proof of Owen (1990, eq. (2.14)) implies $\hat{\lambda}=O_p\left(\frac{d\sqrt{n}}{N-d}\right)$. By proceeding as in the proof of Proposition 1,

$$\hat{\lambda}=\frac{\sum_s\tilde{\zeta}_s}{\sum_s\tilde{\zeta}_s^2}+o_p\left(\frac{d\sqrt{n}}{N-d}\right).$$

Therefore, an expansion yields

$$\tilde{\ell}_J(\theta)=\frac{2}{d}\sum_s\left[\hat{\lambda}\tilde{\zeta}_s-\frac{1}{2}(\hat{\lambda}\tilde{\zeta}_s)^2\right]+o_p(1)=\frac{\left(\frac{\sqrt{n}}{N}\sum_s\tilde{\zeta}_s\right)^2}{\frac{dn}{N(N-d)}\sum_s\tilde{\zeta}_s^2}+o_p(1).$$

The conclusion follows by (A.5) and (A.6), which are shown below.

Proof of (A.5). By (4.4),

$$\begin{aligned}\frac{\sqrt{n}}{N}\sum_s\tilde{\zeta}_s &= \frac{\sqrt{n}}{N}\sum_s\left\{(\hat{\theta}-\theta)+\frac{1}{d}\sqrt{(n-d)(N-d)}\varepsilon_s(\hat{\theta}-\hat{\theta}_s)\right\} \\ &= \sqrt{n}(\hat{\theta}-\theta)+\frac{\sqrt{n(n-d)(N-d)}}{dN}\sum_s\varepsilon_s(R_n-R_{n,s}) \\ &\quad +\frac{\sqrt{n(n-d)(N-d)}}{dN}\sum_s\varepsilon_s\left(\frac{1}{n}\sum_{i=1}^n\phi_i-\frac{1}{n-d}\sum_{i\in s}\phi_i\right) \\ &\equiv T_1+T_2+T_3.\end{aligned}$$

By the assumption in (4.4) and the central limit theorem, we have $T_1\xrightarrow{d}N(0,\Omega)$. For T_2, observe that $E[T_2]=0$ and

$$Var(T_2)=\left(\frac{n(n-d)}{d}Var(R_n-R_{n,s})\right)\left(\frac{N-d}{dN}Var(\varepsilon_s)\right)\to 0,$$

by (4.5). Thus, the Markov inequality implies $T_2\xrightarrow{p}0$. For T_3, observe that $E[T_3]=0$ and

$$\begin{aligned}Var(T_3) &= \frac{n(n-d)(N-d)}{d^2N}Var\left(\frac{1}{n}\sum_{i=1}^n\phi_i-\frac{1}{n-d}\sum_{i=d+1}^n\phi_i\right)Var(\varepsilon_s) \\ &= \frac{n(n-d)(N-d)}{d^2N}\left\{\frac{d}{n^2}+\frac{d^2}{n^2(n-d)}\right\}Var(\phi_i)Var(\varepsilon_s)\to 0.\end{aligned}$$

Thus, we obtain $T_3\xrightarrow{p}0$. Combining these results, (A.5) is obtained.

Proof of (A.6). Decompose

$$\begin{aligned}\frac{dn}{N(N-d)}\sum_s \tilde{\zeta}_s^2 &= \frac{dn}{N(N-d)}\sum_s \left\{(\hat{\theta}-\theta)+\frac{1}{d}\sqrt{(n-d)(N-d)}\varepsilon_s(\hat{\theta}-\hat{\theta}_s)\right\}^2 \\ &= \frac{dn}{N-d}(\hat{\theta}-\theta)^2+\frac{n(n-d)}{d}\frac{1}{N}\sum_s \varepsilon_s^2(\hat{\theta}-\hat{\theta}_s)^2 \\ &\quad +2\sqrt{n}(\hat{\theta}-\theta)\sqrt{\frac{n(n-d)}{N-d}}\frac{1}{N}\sum_s \varepsilon_s(\hat{\theta}-\hat{\theta}_s) \\ &\equiv A_1+A_2+2A_3.\end{aligned}$$

For A_1, since the assumption in (4.4) guarantees $\sqrt{n}(\hat{\theta}-\theta)=O_p(1)$, we have

$$A_1=\frac{d}{N-d}\{\sqrt{n}(\hat{\theta}-\theta)\}^2 \xrightarrow{p} 0.$$

For A_2, since $\varepsilon_s^2=1$ by construction, we have

$$A_2=\frac{n(n-d)}{d}\frac{1}{N}\sum_s(\hat{\theta}-\hat{\theta}_s)^2=nV_d,$$

where V_d is the delete-d jackknife variance estimator in (4.1) considered by Shao and Wu (1989). Thus, Shao and Wu (1989, Theorem 1) directly imply

$$A_2 \xrightarrow{p} \Omega.$$

For A_3, a similar argument to the proof of (A.5) yields $A_3 \xrightarrow{p} 0$. Combining these results, the result in (A.6) follows.

References

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

[2] Bravo, F., Escanciano, J. C. and I. van Keilegom (2015) Wilks' phenomenon in two-step semiparametric empirical likelihood inference, Working paper.

[3] Cattaneo, M. D., Crump, R. K. and M. Jansson (2010) Robust data-driven inference for density-weighted average derivatives, *Journal of the American Statistical Association*, 105, 1070-1083.

[4] Cattaneo, M. D., Crump, R. K. and M. Jansson (2013) Generalized jackknife estimators of weighted average derivatives, *Journal of the American Statistical Association*, 108, 1243-1256.

[5] Chaudhuri, P., Doksum, K. and A. Samarov (1997) On average derivative quantile regression, *Annals of Statistics*, 25, 715-744.

[6] Coppejans, M. and H. Sieg (2005) Kernel estimation of average derivatives and differences, *Journal of Business and Economic Statistics*, 23, 211-225.

[7] Deaton, A. and S. Ng (1998) Parametric and nonparametric approaches to price and tax reform, *Journal of the American Statistical Association*, 93, 900-909.

[8] DiCiccio, T. J., Hall, P. and J. Romano (1991) Empirical likelihood is Bartlett-correctable, *Annals of Statistics*, 19, 1053-1061.

[9] Firpo, S. (2007) Efficient semiparametric estimation of quantile treatment effects, *Econometrica*, 75, 259-276.

[10] Hahn, J. and G. Ridder (2013) Asymptotic variance of semiparametric estimators with generated regressors, *Econometrica*, 81, 315-340.

[11] Härdle, W., Hart, J., Marron, J. and A. Tsybakov (1992) Bandwidth choice for average derivative estimation, *Journal of the American Statistical Association*, 87, 218-226.

[12] Härdle, W., Hildenbrand, W. and M. Jerison (1991) Empirical evidence on the law of demand, *Econometrica*, 59, 1525-1549.

[13] Härdle, W. and T. M. Stoker (1989) Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association*, 84, 986-995.

[14] Heckman, J. J., Ichimura, H. and P. Todd (1998) Matching as an econometric evaluation estimator, *Review of Economic Studies*, 65, 261-294.

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

[16] Hjort, N. L., McKeague, I. W. and I. van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079-1111.

[17] Horowitz, J. L. and W. Härdle (1996) Direct semiparametric estimation of single-index models with discrete covariates, *Journal of the American Statistical Association*, 91, 1632-1640.

[18] Jing, B. Y., Yuan, J. and W. Zhou (2009) Jackknife empirical likelihood, *Journal of the American Statistical Association*, 104, 1224-1232.

[19] Matsushita, Y. and T. Otsu (2016) Likelihood inference on semiparametric models with generated regressors, Working paper.

[20] Matsushita, Y. and T. Otsu (2017) ...

[21] Newey, W. K. and T. M. Stoker (1993) Efficiency of weighted average derivative estimators and index models, *Econometrica*, 61, 1199-1223.

[22] Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, 75, 237-249.

[23] Owen, A. B. (2001) *Empirical Likelihood*, Chapman & Hall/CRC.

[24] Powell, J. L., Stock, J. H. and T. M. Stoker (1989) Semiparametric estimation of index coefficients, *Econometrica*, 57, 1403-1430.

[25] Quenouille, M. H. (1956) Notes on bias in estimation, *Biometrika*, 43, 353-360.

[26] Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70, 41-55.

[27] Rothe, C. and S. Firpo (2016) Properties of doubly robust estimators when nuisance functions are estimated nonparametrically, Working paper.

[28] Severini, T. A. (2000) *Likelihood Methods in Statistics*, Oxford University Press.

[29] Shao, J. and D. Tu (1995) *The Jackknife and Bootstrap*, Springer.

[30] Shao, J. and C. F. J. Wu (1989) A general theory for jackknife variance estimation, *Annals of Statistics*, 17, 1176-1197.

[31] Smith, R. J. (1997) Alternative semi-parametric likelihood approaches to generalised method of moments estimation, *Economic Journal*, 107, 503-519.

[32] Stoker, T. M. (1986) Consistent estimation of scaled coefficients, *Econometrica*, 54, 1461-1481.

[33] Stoker, T. M. (1989) Tests of additive derivative constraints, *Review of Economic Studies*, 56, 535-552.

[34] Tukey, J. W. (1958) Bias and confidence in not-quite large samples, *Annals of Mathematical Statistics*, 29, 614.

[35] Xue, L. and D. Xue (2011) Empirical likelihood for semiparametric regression model with missing response data, *Journal of Multivariate Analysis*, 102, 723-740.

[36] Zhu, L. and L. Xue (2006) Empirical likelihood confidence regions in a partially linear single- index model, *Journal of the Royal Statistical Society*, B, 68, 549-570.

Department of Industrial Engineering and Economics, Graduate School of Engineering, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo 152-8550, Japan.

E-mail address: matsushita.y.ab@m.titech.ac.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: t.otsu@lse.ac.uk