LIKELIHOOD INFERENCE ON SEMIPARAMETRIC MODELS WITH GENERATED REGRESSORS

YUKITOSHI MATSUSHITA AND TAISUKE OTSU

Abstract. Hahn and Ridder (2013) formulated influence functions of semiparametric three step estimators where generated regressors are computed in the first step. This class of estimators covers several important examples for empirical analysis, such as production function estimators by Olley and Pakes (1996) and propensity score matching estimators for treatment effects by Heckman, Ichimura and Todd (1998). The present paper studies a nonparametric likelihood-based inference method for the parameters in such three step estimation problems. In particular, we apply the general empirical likelihood theory of Bravo, Escanciano and van Keilegom (2018) to modify semiparametric moment functions to account for influences from plug-in estimates into the above important setup, and show that the resulting likelihood ratio statistic becomes asymptotically pivotal without undersmoothing in the first and second step nonparametric estimates.

1. Introduction

There is a class of econometric problems, where the parameter of interest is estimated by three (or more) certain steps. In the first step, generated regressors (say, $\hat{V}_i$) are computed by some parametric or nonparametric estimation. In the second step, a certain nonparametric regression (say, from Y_i on $(X_i, \hat{V}_i)$) is implemented to obtain an estimator $\hat{\gamma}(X_i, \hat{V}_i)$. In the third step, the parameter of interest β is estimated by the sample average or more generally by the method of moments, $n^{-1}\sum_{i=1}^n g(\hat{\gamma}(X_i, \hat{V}_i), \hat{\beta}) = 0$, where g is a vector of moment functions having the same dimension as β. Indeed several important econometric estimators are formulated in this three step manner or interpreted as a special case. Examples include production function estimators by Olley and Pakes (1996), propensity score matching estimators for treatment effects by Heckman, Ichimura and Todd (1998), and various semiparametric estimators that involve generated regressors or control variables.

This three step approach provides an intuitive way to construct a point estimator for the main parameter β. On the other hand, the three step formulation complicates inference methods

The authors would like to thank the Editor, Co-Editor, and anonymous referees for helpful comments. Financial support from the JSPS KAKENHI (26780133, 18K01541) (Matsushita) and the ERC Consolidator Grant (SNP 615882) are gratefully acknowledged (Otsu).

on β that properly take into account the sampling variations contained in $\hat{\beta}$. In particular, it is known that for regression models, the estimation errors in generated regressors should be incorporated to compute the standard errors (Pagan, 1984), and it is not trivial to characterize how the estimation errors of the generated regressors $\hat{V}_i$ contribute to the standard error of $\hat{\beta}$. By applying Newey's (1994) path derivative method, Hahn and Ridder (2013) settled this problem and derived the influence function of $\hat{\beta}$.[1] As shown in Hahn and Ridder (2013), the influence function consists of three components: the main term due to the third step, adjustment for the second step estimation of $\hat{\gamma}$, and adjustment due to the first step estimation of $\hat{V}_i$. The third component is the most challenging one and is further decomposed into two terms associated with the two roles of $\hat{V}_i$'s played in the second step nonparametric regression as a conditioning variable and argument.

In this paper, we consider nonparametric likelihood inference for the parameter β defined in the three step estimation problem by using the method of generalized empirical likelihood (GEL) (Smith, 1997, and Newey and Smith, 2004). Indeed Bravo, Escanciano and van Keilegom (2018, hereafter BEV) developed general empirical likelihood theory for a semiparametric moment function $m(Z, \beta, \hat{h})$ which involves plug-in nonparametric estimates $\hat{h}$. BEV proposed a general approach to modify the moment functions to account for influences from estimation errors in $\hat{h}$ so that the resulting empirical likelihood statistic is asymptotically pivotal and implementation of $\hat{h}$ does not require undersmoothing. The three step estimation problems above may be accommodated into BEV's general setup by setting $\hat{h}(\cdot) = \hat{\gamma}(\cdot, \hat{\varphi}(\cdot))$, where $\hat{\varphi}$ is a nonparametric estimator for the generated regressors. The contribution of this paper is to apply BEV's general empirical likelihood theory to the three step estimation problem. In particular, we show that the resulting GEL statistic becomes asymptotically pivotal and chi-squared distributed. Also, in contrast to inference based on the t-ratio, another desirable feature of our GEL inference is that it does not require undersmoothing for the bandwidths in the first and second step estimation. We emphasize that BEV established their general theory under high-level assumptions and did not consider the three step estimation problem in their examples. Due to the complicated structure of $\hat{h}(\cdot) = \hat{\gamma}(\cdot, \hat{\varphi}(\cdot))$ (especially $\hat{\varphi}$ appearing in the argument of $\hat{\gamma}$) as clarified by Hahn

[1] Mammen, Rothe and Schienle (2016) investigated general theory for semiparametric M-estimators containing generated variables. They provided conditions to guarantee $\sqrt{n}$-consistency and asymptotic normality of the semiparametric estimators and established validity of the bootstrap.

and Ridder (2013), it is not trivial to establish the above results from primitive conditions in the present setup.

For detailed theoretical developments based on primitive conditions, we concentrate on the case where the second step nonparametric estimate $\hat{\gamma}(\cdot)$ is given by the local linear fitting, the nonparametric function $\varphi(\cdot)$ for the generated regressors takes the form of conditional mean, and $\hat{\varphi}(\cdot)$ is given by the kernel regression fitting. Although the detailed analysis is case-by-case, we expect that similar results hold for other nonparametric estimators.

This paper is organized as follows. In Section 2, we present the basic setup and main results. Sections 2.1 and 2.2 consider the cases of parametric and nonparametric first step, respectively. In Section 3, we provide some extensions of our approach to inference on subvectors or functions of β (Section 3.1), the cases of additional variables (Section 3.2), partial means (Section 3.3), and multidimensional $\hat{\gamma}$ (Section 3.4), and other nonparametric likelihood functions (Section 3.5). In Section 4, our method is illustrated using two examples; a simplified version of Olley and Pakes' (1996) estimator (Section 4.1) and propensity score matching estimators (Section 4.2). Section 5 presents some simulation results.

2. Main results

Our notation follows closely that of Hahn and Ridder (2013). Suppose we observe a random sample $\{Y_i, X_i, Z_i\}_{i=1}^n$ of $(Y, X, Z) \in \mathbb{R} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}$. We wish to conduct inference on the k-dimensional vector of parameters β satisfying the moment condition

$$E[g(\mu(X, V), \beta)] = 0, \tag{2.1}$$

where g is a k-dimensional vector of known functions up to $\mu(\cdot, \cdot)$ and β, $\mu(X, V) = E[Y|X, V]$ is the conditional mean, and V is a scalar unobservable regressor expressed as $V = \varphi(X, Z)$ by some unknown function φ. When φ is known up to finite d_α-dimensional parameters α, we denote it by $V = \varphi(X, Z, \alpha)$. We can estimate β in three-steps. First, evaluate the unobservable regressor V_i by its sample counterpart $\hat{V}_i = \varphi(X_i, Z_i, \hat{\alpha})$ based on some parameter estimator $\hat{\alpha}$ of α (called a parametric first step) or $\hat{V}_i = \hat{\varphi}(X_i, Z_i)$ based on a nonparametric estimator (called a nonparametric first step). The sample counterpart $\hat{V}_i$ is often called the generated regressor. Second, estimate the conditional mean function $\mu(X_i, V_i)$ by nonparametric regression of Y_i on

$(X_i, \hat{V}_i)$. We denote the estimated function (evaluated at $(X_i, \hat{V}_i)$) by $\hat{\gamma}(X_i, \hat{V}_i)$.[2] Third, compute the estimator $\hat{\beta}$ for the parameter of interest β by solving $n^{-1}\sum_{i=1}^n g(\hat{\gamma}(X_i, \hat{V}_i), \hat{\beta}) = 0$.

Several estimators in econometrics and statistics are formulated in this three-step manner. Examples include semiparametric estimators with generated regressors, and some average treatment effect estimators. See Section 4 below for some specific examples. Hahn and Ridder (2013) derived the influence function for $\hat{\beta}$ by analyzing carefully the effect of the first step estimation. This paper focuses on (nonparametric) likelihood-based inference on β without undersmoothing the bandwidth to compute $\hat{\gamma}(\cdot, \cdot)$ in the second step (and $\hat{V}_i$ in the nonparametric first step).

2.1. **Case of parametric first step.** We first consider the case where the unobservable regressor $V = \varphi(X, Z, \alpha)$ is generated from a parametric model indexed by α. Let $\hat{\alpha}$ be an estimator of α, which satisfies Assumption P (v) below. In this case, we evaluate the unobservable regressor V_i by the generated regressor $\hat{V}_i = \varphi(X_i, Z_i, \hat{\alpha})$.

To proceed, we fix the nonparametric estimators for the conditional mean function $\mu(x, v) = E[Y|X = x, V = v]$ and its partial derivative $\mu_v(x, v) = \partial\mu(x, v)/\partial v$ with respect to the second argument. To be specific, we hereafter consider the local linear regression from Y_i on $(X_i', \hat{V}_i)$:

$$\min_{\gamma,\gamma_x,\gamma_v} \sum_{i=1}^n K\left(\frac{(X_i - x)'}{h}, \frac{\hat{V}_i - v}{h}\right)\{Y_i - \gamma - (X_i - x)'\gamma_x - (\hat{V}_i - v)\gamma_v\}^2. \tag{2.2}$$

We employ the intercept and slope coefficient of $\hat{V}_i$ as estimators for $\mu(x, v)$ and $\mu_v(x, v)$, respectively. Denote these estimators by $\hat{\gamma}(x, v)$ and $\hat{\gamma}_v(x, v)$, respectively.[3]

Let $g_1(\mu, \beta)$ and $g_2(\mu, \beta)$ be the first and second derivatives of $g(\cdot, \cdot)$ with respect to its first argument evaluated at (μ, β) (i.e., both g_1 and g_2 are k-dimensional), and $\varphi_\alpha(x, z, \alpha)$ be the partial derivative of $\varphi(\cdot, \cdot, \cdot)$ with respect to its third argument evaluated at (x, z, α) (i.e., φ_α is d_α-dimensional). Let ψ be the influence function of $\hat{\alpha}$ (i.e., $\sqrt{n}(\hat{\alpha}-\alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i, Z_i, \alpha)+o_p(1)$ as in Assumption P (v) below). Based on the above notation, we propose the following GEL statistic

$$\ell(\beta) = 2 \sup_{\lambda\in\Lambda_n(\beta)} \sum_{i=1}^n \rho(\lambda'\tilde{g}_i(\beta)) - 2n\rho(0), \tag{2.3}$$

[2]Here we follow the notation of Hahn and Ridder (2013). They reserve the notation $\hat{\mu}(X_i, V_i)$ for (infeasible) nonparametric regression of Y_i on (X_i, V_i).

[3]Here the local linear regression is employed because of its mathematical simplicity and convenience (both $\hat{\gamma}$ and $\hat{\gamma}_v$ are obtained by single least square fitting). Similar results can be derived for other estimators, such as the kernel and local polynomial regression estimators.

where $\rho(\cdot)$ is a concave function on its domain $\mathcal{V}$, an open interval containing zero, and

$$\begin{aligned}\tilde{g}_i(\beta) &= g(\hat{\gamma}(X_i,\hat{V}_i),\beta)+\hat{\Delta}\psi(X_i,Z_i,\hat{\alpha})+g_1(\hat{\gamma}(X_i,\hat{V}_i),\beta)\{Y_i-\hat{\gamma}(X_i,\hat{V}_i)\}, \qquad (2.4)\\ \hat{\Delta} &= \frac{1}{n}\sum_{i=1}^{n}\{Y_i-\hat{\gamma}(X_i,\hat{V}_i)\}g_2(\hat{\gamma}(X_i,\hat{V}_i),\beta)\hat{\gamma}_v(X_i,\hat{V}_i)\varphi_\alpha(X_i,Z_i,\hat{\alpha})',\\ \Lambda_n(\beta) &= \{\lambda:\lambda'\tilde{g}_i(\beta)\in\mathcal{V},\ i=1,\ldots,n\}.\end{aligned}$$

Note that our moment function $\tilde{g}_i(\beta)$ is composed of three terms. The first term in (2.4) is a plug-in version of the original moment function in (2.1), and the others are correction terms to achieve asymptotic pivotalness. The second term is an adjustment due to the first step estimation of $\hat{V}_i$, and the third term is another adjustment due to the second step estimation of $\hat{\gamma}(\cdot,\cdot)$. These correction terms are considered as sample counterparts of the influence functions for the first and second stage estimation derived in Hahn and Ridder (2013) and Newey (1994), respectively.

For the criterion function $\rho(\cdot)$ to define the GEL statistic, popular choices are empirical likelihood ($\rho(v)=\log(1-v)$ and $\mathcal{V}=(-\infty,1)$), exponential tilting ($\rho(v)=-e^v$), and continuous updating GMM (a quadratic $\rho(\cdot)$). See Section 3.5 for a further general class of statistics.

As shown in Newey and Smith (2004), the GEL statistic in (2.3) has the following dual representation

$$\ell(\beta)=2\sup_{\{p_i\}_{i=1}^n}\left\{\sum_{i=1}^{n}h(np_i):\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_i\tilde{g}_i(\beta)=0\right\}, \qquad (2.5)$$

where $h(\cdot)$ is a convex function to measure the discrepancy between the multinomial weights $\{p_i\}$ under the constraint $\sum_{i=1}^n p_i\tilde{g}_i(\beta)=0$ and the unconstrained weights n^{-1}. For example, if $\rho(v)=\log(1-v)$ (empirical likelihood), the dual form is given by $h(np_i)=-\log(np_i)$. Thus, the GEL statistic $\ell(\beta)$ can be interpreted as a conventional likelihood ratio statistic using multinomial weights. For implementation, we employ the form in (2.3) since it involves optimization only for the k-dimensional vector λ.

In the setup of this subsection, we impose the following assumptions.

Assumption P.

(i): *$\{Y_i,X_i',Z_i'\}_{i=1}^n$ is an iid sample from $(Y,X',Z')\in\mathbb{R}\times\mathbb{X}\times\mathbb{Z}$. $\mathbb{X}$, $\mathbb{Z}$, and $\mathbb{V}$ are compact. The joint density $f(x,v)$ of (X,V) is continuously differentiable to order $s\geq 2$ and bounded away from zero on $\mathbb{X}\times\mathbb{V}$. $\mu(x,v)$ is continuously differentiable to order $s\geq 2$ on $\mathbb{X}\times\mathbb{V}$. For some $p\geq 4$, $E|Y|^p<\infty$ and $E[|Y|^p|X=x,V_*=v]f(x,v)$ is bounded*

over $\mathbb{X} \times \mathbb{V}$. $g(\cdot, \beta)$ *is twice continuously differentiable with respect to the first argument. For some neighborhood* $\mathcal{N}$ *of* α, $\varphi_{\alpha\alpha}(x, z, \alpha)$ *is continuous over* $\mathbb{X} \times \mathbb{Z} \times \mathcal{N}$.

(ii): ρ *is concave and twice continuously differentiable in a neighborhood of zero, and the first and second derivatives (denoted by* ρ_1 *and* ρ_2*, respectively) satisfy* $\rho_1(0) = -1$ *and* $\rho_2(0) = -1$*, respectively.*

(iii): $K(\cdot)$ *integrates to one, is compactly supported and twice differentiable with bounded derivatives, and satisfies* $\int K(u) u_1^{j_1} \cdots u_{d_x+1}^{j_{d_x+1}} du = 0$ *for all vectors of non-negative integers* $(j_1, \ldots, j_{d_x+1})$ *such that* $j_1 + \cdots + j_{d_x+1} < s$.

(iv): *As* $n \to \infty$*, it holds* $n^{1/2} h^{d_x+1} / \log n \to \infty$ *and* $nh^{4s} \to 0$.

(v): $\hat{\alpha}$ *satisfies*

$$\sqrt{n}(\hat{\alpha} - \alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, Z_i, \alpha) + o_p(1), \tag{2.6}$$

with $E|\psi(X, Z, \alpha)|^2 < \infty$ *and* $n^{-1} \sum_{i=1}^{n} \psi(X_i, Z_i, \hat{\alpha}) = o_p(n^{-1/2})$.

(vi): $\hat{\Delta} \xrightarrow{p} \Delta = E[g_2(\mu(X, V), \beta) \mu_v(X, V) \{\mu(X, Z) - \mu(X, V)\} \varphi_\alpha(X, Z, \alpha)']$.

Assumption P (i) collects conditions on the distributions of the observables (Y, X, Z) and unobservable regressor V, and smoothness of the functions g and φ. The compact support assumptions on X, Z, and V may be relaxed by introducing trimming terms to deal with denominator problems for kernel-based estimators. This assumption also requires that the sample is iid (see Remark 7 below for an extension to weakly dependent data). Assumption P (ii) is on the GEL criterion function ρ in (2.3). This assumption is mild enough to cover popular criterions, such as empirical likelihood, exponential tilting, and Cressie-Read's power divergence family. Assumption P (iii) is on the kernel function K in (2.2) to estimate μ and μ_v. This requires that K is an s-th order kernel function. Assumption P (iv) is on the bandwidth h in (2.2). We emphasize that this assumption does not require undersmoothing, i.e., we only require $nh^{4s} \to 0$ instead of $nh^{2s} \to 0$. Thus, for example, the MSE optimal bandwidth is allowed. See Remark 5 below for further discussion. Assumption P (v) is on the first-stage estimator $\hat{\alpha}$. These requirements are typically satisfied for popular estimators, such as the maximum likelihood and generalized method of moments (GMM) estimators, under mild regularity conditions.[4] The function ψ is called the influence function for $\hat{\alpha}$. Assumption P (vi) is a high level assumption

[4]As an example, consider the GMM estimator $\hat{\alpha}$ solving $\{\sum_{i=1}^{n} \partial m_i(\hat{\alpha})/\partial \alpha'\}' W \{\sum_{i=1}^{n} m_i(\hat{\alpha})\} = 0$, where W is a positive definite weight matrix and $m_i(\alpha) = m(X_i, Z_i, \alpha)$. Mild regularity conditions guarantee (2.6) with $\psi(X_i, Z_i, \alpha) = (M'WM)^{-1} M'W m_i(\alpha)$, where $M = E[\partial m_i(\alpha)/\partial \alpha']$. Also, the requirement $\frac{1}{n} \sum_{i=1}^{n} \psi(X_i, Z_i, \hat{\alpha}) = o_p(n^{-1/2})$ can be verified by ensuring $\frac{1}{n} \sum_{i=1}^{n} \frac{\partial m_i(\hat{\alpha})}{\partial \alpha'} \xrightarrow{p} M$ and $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_i(\hat{\alpha}) = O_p(1)$.

on $\hat{\Delta}$ that appears in the correction term of $\tilde{g}_i(\beta)$. This assumption can be verified by applying the law of large numbers for U-statistics.[5]

The main result of this paper, the asymptotic distribution of the GEL statistic, is presented as follows. The proof is given in Appendix A.

Theorem 1. *Consider the setup of this subsection. Under Assumption P, it holds*

$$\ell(\beta) \xrightarrow{d} \chi^2(k).$$

Remark 1. This theorem says that the GEL statistic $\ell(\beta)$ is asymptotically pivotal and converges to the $\chi^2(k)$ distribution. Based on this theorem, the $100(1-\alpha)\%$ asymptotic confidence set is constructed as $CS_\alpha^{GEL} = \{b : \ell(b) \leq \chi^2_{1-\alpha}(k)\}$, where $\chi^2_{1-\alpha}(k)$ is the $(1-\alpha)$-th quantile of the $\chi^2(k)$ distribution. A drawback of CS_α^{GEL} (compared to the conventional one based on the t-ratio) is that it requires a numerical search. If the parameter of interest β is scalar, a grid search can be applied to compute CI_α^{GEL}. For multidimensional β, we can apply the subvector inference as in Section 3.1 below to obtain the confidence set for each element of β.

Remark 2. The correction terms of $\tilde{g}_i(\beta)$ in (2.4) are considered as sample counterparts of the influence functions for the first and second stage estimation derived in Hahn and Ridder (2013) and Newey (1994), respectively. Indeed, our correction terms may be used for the t or Wald test as well. To simplify, suppose $g(\mu(X,V),\beta) = h(\mu(X,V)) - \beta$ for some known function h. Then by Lemma A.6, the asymptotic variance of the estimator $\hat{\beta} = n^{-1}\sum_{i=1}^n h(\hat{\gamma}(X_i,\hat{V}_i))$ can be consistently estimated by $n^{-1}\sum_{i=1}^n \tilde{g}_i(\beta)\tilde{g}_i(\beta)'$.

Remark 3. We can also show that the GEL statistic $\ell(\beta)$ is consistent and converges to the non-central $\chi^2(k)$ distribution with non-centrality $c'G'\Omega^{-1}Gc$ with $G = E\left[\frac{\partial g(\mu(X,V),\beta)}{\partial \beta'}\right]$ under the local alternative hypothesis $H_{1n} : \beta_n = \beta + c/\sqrt{n}$ for some $c \neq 0$ (by modifying Lemma A.4 to show $\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{g}_i(\beta_n) \xrightarrow{d} N(Gc, \Omega)$).

Remark 4. Theorem 1 is considered as a specialization of the empirical likelihood theory of BEV to the three step estimation problem. As in BEV, it is crucial to incorporate the last two terms in (2.4) to achieve the asymptotic pivotalness. Without these terms, the corresponding

[5]Typically, under smoothness assumptions on g, φ, and $\hat{\gamma}$, we can expand $\hat{\Delta}$ around $\hat{\alpha} = \alpha$ (or $\hat{V}_i = V_i$), $\hat{\gamma} = \mu$, and $\hat{\gamma}_v = \mu_v$. Then by the law of large numbers, the main term converges to Δ under finite moment assumptions for Y, $\mu(X,V)$, $g_2(\mu(X,V),\beta)$, $\mu_v(X,V)$, and $\varphi_\alpha(X,Z,\alpha)$. Also the remainder terms are shown to be asymptotically negligible by applying the consistency of $\hat{\alpha}$ and $\hat{\gamma}$ and the law of large numbers for U-statistics to guarantee (stochastic) boundedness of the linear expansion coefficients.

statistic $\ell^{\text{unadjusted}}(\beta) = 2\sup_{\lambda\in\Lambda_n(\beta)}\sum_{i=1}^n \rho(\lambda' g(\hat{\gamma}(X_i,\hat{V}_i),\beta)) - 2n\rho(0)$ converges to the $\chi^2(k)$ distribution multiplied by a constant that depends on some nuisance parameters, and is not asymptotically pivotal.[6] It should be noted that this specialization to the three step estimation problem is not trivial due to the influence of the first step estimation for generated regressors as shown by Hahn and Ridder (2013).

Remark 5. We note that the condition on the bandwidth h to compute $\hat{\gamma}(\cdot,\cdot)$ (Assumption P (iv)) does not require undersmoothing, i.e., we only require $nh^{4s} \to 0$ instead of $nh^{2s} \to 0$. This property is known in the empirical likelihood literature for several setups (e.g., BEV, Zhu and Xue (2006), Zhu, *et al.* (2010), and Xue and Xue (2011)). See also Newey (1994) and Newey, Hsieh and Robins (2004) for analogous discussions in the context of semiparametric M-estimators. Theorem 1 shows that a similar result holds for the three step estimation problem. Intuitively, the first and third terms in $\tilde{g}_i(\beta)$ share the same form as the smoothing bias and these bias terms are automatically cancelled out. We emphasize that in contrast to the GEL confidence set CS_α^{GEL}, the Wald-type confidence set using the asymptotic variance estimator based on Hahn and Ridder's (2013) formula requires undersmoothing for the bandwidth.

Remark 6. If the parameter of interest is explicitly defined as $\beta = h(\mu(X,V))$ for some known function h, then we can apply Theorem 1 by setting $g(\mu(X,V),\beta) = h(\mu(X,V)) - \beta$. If g is linear in μ, then the second term in (2.4) vanishes (by $g_2(\cdot) = 0$), and the moment function simplifies to

$$\tilde{g}_i(\beta) = g(\hat{\gamma}(X_i,\hat{V}_i),\beta) + g_1(\hat{\gamma}(X_i,\hat{V}_i),\beta)\{Y_i - \hat{\gamma}(X_i,\hat{V}_i)\}.$$

Furthermore, based on the argument in Newey (1994, pp. 1357-8), the third term in (2.4) vanishes when $n^{-1}\sum_{i=1}^n g(\hat{\gamma}(X_i,\hat{V}_i),\hat{\beta}) = 0$ is the first-order condition for $\hat{\beta}$ to maximize an objective function (say, $n^{-1}\sum_{i=1}^n q(\hat{\gamma}(X_i,\hat{V}_i),\beta)$) and the limit of $\hat{\gamma}$ indeed maximizes its population counterpart $E[q(\gamma(X,V),\beta)]$ with respect to γ (i.e., γ is concentrated out). See Newey (1994) for some examples.

Remark 7. It is interesting to see whether the iid assumption in Assumption P (i) can be relaxed to allow, for example, weakly dependent data. For the conventional moment condition models

[6] Although $\ell^{\text{unadjusted}}(\beta)$ is not asymptotically pivotal, its adjusted version, obtained by multiplying an adjustment term, has the same local power property as $\ell(\beta)$. However, our simulation results in Section 5 suggest that such an adjusted statistic underperforms in finite samples. Existing papers on semiparametric two step inference (e.g., BEV and Xue and Xue, 2011) also report underperformance of the multiplicative adjustments.

without generated variables, the GEL statistic using block average or smoothed moment functions converges to the chi-squared distribution (Kitamura, 1997, and Smith, 1997). A natural question is whether an analogous result can be established for the present setup. A recent paper by Bravo, Chu and Jacho-Chávez (2017), who studied asymptotic properties of the (smoothed) GMM, GEL, and related estimators for semiparametric moment condition models, allows generated variables under weakly dependent data (without the second step nonparametric estimate $\hat{\gamma}(\cdot)$). Interestingly, they showed that in general, the smoothed GEL estimator becomes asymptotically less efficient than the smoothed GMM estimator in the presence of generated variables. Since our setup is even more complicated, we leave such an generalization for future research.

2.2. **Case of nonparametric first step.** We next consider the case where $V = \varphi(X, Z)$ is written as an unknown function φ and needs to be estimated by some nonparametric method. In particular, we focus on the situation where V is written as the conditional mean (i.e., $V = \varphi(X, Z) = E[U|X, Z]$ for some observable U) and $\varphi(X, Z)$ is estimated by the nonparametric kernel estimator

$$\hat{\varphi}(x, z) = \frac{\sum_{j=1}^{n} \tilde{K}\left(\frac{X_j - x}{b}, \frac{Z_j - z}{b}\right) U_j}{\sum_{j=1}^{n} \tilde{K}\left(\frac{X_j - x}{b}, \frac{Z_j - z}{b}\right)}, \tag{2.7}$$

where $\tilde{K}$ is a kernel function and b is the bandwidth.[7] Let us redefine the generated regressor as $\hat{V}_i = \hat{\varphi}(X_i, Z_i)$.

In this case, we modify the GEL statistic in (2.3) by replacing $\tilde{g}_i(\beta)$ with

$$\tilde{g}_i(\beta) = g(\hat{\gamma}(X_i, \hat{V}_i), \beta) + \hat{\Delta}_{1i}(U_i - \hat{V}_i) + g_1(\hat{\gamma}(X_i, \hat{V}_i), \beta)\{Y_i - \hat{\gamma}(X_i, \hat{V}_i)\}, \tag{2.8}$$

where $\hat{\Delta}_{1i}$ is the nonparametric regression fitted value of $\delta_{1i} = \{Y_i - \hat{\gamma}(X_i, \hat{V}_i)\} g_2(\hat{\gamma}(X_i, \hat{V}_i), \beta)\hat{\gamma}_v(X_i, \hat{V}_i)$ on (X_i, Z_i) satisfying Assumption NP (iii) below.[8]

We impose the following assumptions for the case of nonparametric first step estimators.

Assumption NP. *In addition to Assumption P (i)-(iv), suppose*

(i): *The joint density $f(x, z)$ of (X, Z) is continuously differentiable to order $s \geq 2$ and bounded away from zero on $\mathbb{X} \times \mathbb{Z}$. The function $\varphi(x, z) = E[U|X = x, Z = z]$ is*

[7] We choose the kernel estimator $\hat{\varphi}(X, Z)$ due to its simplicity of our theoretical developments. Although the proofs become more tedious, we expect that analogous results can be derived for other estimators such as local linear or polynomial estimators.

[8] For example, $\hat{\Delta}_{1i}$ can be obtained as in (2.7) by setting $U_j = \delta_{1j}$ and $(x, z) = (X_i, Z_i)$.

continuously differentiable to order $s \geq 2$ *on* $\mathbb{X} \times \mathbb{Z}$. *For some* $p \geq 4$, $E|U|^p < \infty$ *and* $E[|U|^p|X = x, Z = z]f(x, z)$ *is bounded on* $\mathbb{X} \times \mathbb{Z}$.

(ii): $\tilde{K}(\cdot)$ *satisfies similar conditions as Assumption P (iii). As* $n \to \infty$, *it holds* $n^{1/2}b^{d_x+d_z}/\log n \to \infty$ *and* $nb^{4s} \to 0$.

(iii): $\max_{1\leq i\leq n} |\hat{\Delta}_{1i} - \Delta_i| \xrightarrow{p} 0$, *where* $\Delta_i = E[\{Y_i - \mu(X_i, V_i)\}g_2(\mu(X_i, V_i), \beta)\mu_v(X_i, V_i)|X_i, Z_i]$.

These assumptions play analogous roles as Assumption P (v)-(vi). Assumption NP (i) collects additional conditions on the distribution of the observables and smoothness of the function $\varphi(x, z)$. Assumption NP (ii) is on the kernel function $\tilde{K}$ and bandwidth b to estimate the nonparametric first stage estimator $\hat{\varphi}(x, z)$ in (2.7). Note that similar to the second stage estimation for μ and μ_v (Assumption P (iv)), the first stage estimation for φ also does not require undersmoothing; see further discussion below. Assumption NP (iii) is a high level assumption on $\hat{\Delta}_i$ that appears in the correction term of $\tilde{g}_i(\beta)$. This assumption can be verified by applying certain uniform laws of large numbers.[9]

Similar to the case of a parametric first step, the GEL statistic converges to the $\chi^2(k)$ distribution without undersmoothing.

Theorem 2. *Consider the setup of this subsection. Under Assumption NP, it holds*

$$\ell(\beta) \xrightarrow{d} \chi^2(k).$$

The proof is presented in Appendix B. Similar comments to Theorem 1 apply here. The last two terms of $\tilde{g}_i(\beta)$ in (2.8) recover internal studentization and asymptotic pivotalness. Similar to the bandwidth h for the second step estimator $\hat{\gamma}(\cdot, \cdot)$, Assumption NP (ii) on the bandwidth b for the first step estimator $\hat{\varphi}(\cdot, \cdot)$ does not require undersmoothing, i.e., $nb^{4s} \to 0$ instead of $nb^{2s} \to 0$. This is due to the second term $\hat{\Delta}_{1i}(U_i - \hat{V}_i)$ in (2.8). Without this correction term, there will be a smoothing bias term of order $O(\sqrt{n}h^s)$ from the term $\frac{1}{\sqrt{n}}\sum_{i=1}^n \Delta_i(\hat{V}_i - V_i)$ (see the term M_1 in the proof of Lemma B.4). However, this bias term is cancelled out by the correction

[9]For example, suppose $\hat{\Delta}_{1i}$ is given by the kernel regression

$$\hat{\Delta}_{1i} = \frac{\sum_{j=1}^n \tilde{K}\left(\frac{X_j - X_i}{b}, \frac{Z_j - Z_i}{b}\right)\{Y_i - \hat{\gamma}(X_i, \hat{V}_i)\}g_2(\hat{\gamma}(X_i, \hat{V}_i), \beta)\hat{\gamma}_v(X_i, \hat{V}_i)}{\sum_{j=1}^n \tilde{K}\left(\frac{X_j - X_i}{b}, \frac{Z_j - Z_i}{b}\right)}.$$

In this case, we expand this around $\hat{V}_i = V_i$, $\hat{\gamma} = \mu$, and $\hat{\gamma}_v = \mu_v$. Then the uniform convergence of the kernel estimator (e.g., Hansen, 2008) can be applied to show that the main term converges to Δ_i uniformly over i. The remainder terms are shown to be asymptotically negligible by the consistency of $\hat{\varphi}$, $\hat{\gamma}$, and $\hat{\gamma}_v$ combined with the uniform law of large numbers to bound the linear expansion coefficients.

term $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\Delta}_{1i}(U_i - \hat{V}_i)$ (as in the proof of Lemma B.4), and thus the bandwidth b for the first step estimator $\hat{\varphi}(\cdot,\cdot)$ does not require undersmoothing.

Although our assumptions on the bandwidths h and b are relatively mild, their optimal selection rules are substantial open problems. In the existing literature on two-step semiparametric inference, most papers employ the MSE optimal or cross validation bandwidths for the first stage nonparametric estimation; see, e.g., BEV, Zhu and Xue (2006), Zhu, *et al.* (2010), and Xue and Xue (2011). In our simulation study below, we also choose the bandwidths h and b based on the MSE optimal rate for estimation of μ and φ, respectively, multiplied by several constants to check their robustness. However, it is not obvious whether the optimal bandwidths for nonparametric first stage estimation have desirable properties for inference on the parametric component β of interest. Indeed such literature on bandwidth selection for semiparametric inference is very thin. One promising way is to establish a higher order approximation for the coverage error (or size distortion) by our GEL statistic $\ell(\beta)$, and to choose the bandwidths to minimize the coverage error (see, Nishiyama and Robinson, 2000, and Linton, 2002). Such higher order analysis is complicated even for the two-step inference, and we leave it for future research.[10]

3. Extensions

3.1. **Inference on subvector or function of β.** The results in the previous section focus on inference for the whole vector of parameters β. In this subsection, we extend our nonparametric likelihood approach to inference on subvectors or functions of parameters $\theta = \tau(\beta)$, where $\tau : \mathbb{R}^k \to \mathbb{R}^{k_1}$ for $k_1 \leq k$. To this end, we employ the profile GEL statistic

$$\ell_p(\theta) = \min_{b \in B : \theta = \tau(b)} \ell(b),$$

where B is the parameter space of β. The results in the previous section are extended as follows.

Theorem 3. *Consider the setup of Section 2.1. Suppose (a) Assumption P (ii)–(v) hold true, (b) B is compact, (c) Assumption P (i) holds true for all $\beta \in B$, and $\frac{\partial g(\mu,\beta)}{\partial \beta'}$, $\frac{\partial g_1(\mu,\beta)}{\partial \beta'}$, and $\frac{\partial g_2(\mu,\beta)}{\partial \beta'}$ are continuous at all μ and $\beta \in B$, (d) Assumption P (vi) holds true uniformly over $\beta \in B$, and*

[10] Although formal analysis is beyond the scope of the paper, we conjecture that analogous results can be derived for series estimators (on both μ and φ) by extending our theoretical argument combined with the one in Newey (1994). In particular, the resulting likelihood ratio statistic is expected to be asymptotically pivotal without undersmoothing because of orthogonality of least square projection errors as in Newey (1994, p. 1372). However, our modified moment functions as in (2.4) or (2.8) using series estimators should be employed to obtain asymptotically pivotal likelihood ratio statistics.

(e) τ is continuously differentiable and $\partial\tau/\partial\beta'$ has rank k_1. Then the GEL statistic $\ell_p(\theta)$ using $\tilde{g}_i(\beta)$ in (2.4) satisfies

$$\ell_p(\theta) \xrightarrow{d} \chi^2(k_1).$$

Theorem 4. *Consider the setup of Section 2.2. Suppose (a) Assumption P (ii)–(iv) and NP (i)-(ii) hold true, (b) B is compact, (c) Assumption P (i) holds true for all $\beta \in B$, and $\frac{\partial g(\mu,\beta)}{\partial\beta'}$, $\frac{\partial g_1(\mu,\beta)}{\partial\beta'}$, and $\frac{\partial g_2(\mu,\beta)}{\partial\beta'}$ are continuous at all μ and $\beta \in B$, (d) Assumption NP (iii) holds true uniformly over $\beta \in B$, and (e) τ is continuously differentiable and $\partial\tau/\partial\beta'$ has rank k_1. Then the GEL statistic $\ell_p(\theta)$ using $\tilde{g}_i(\beta)$ in (2.8) satisfies*

$$\ell_p(\theta) \xrightarrow{d} \chi^2(k_1).$$

These results can be used to construct confidence sets for each element of β. Relevant examples include partially linear models with generated regressors discussed in Section 4.1 and estimating equations with missing data and generated covariates (cf. Section 4.2 of BEV). We also note that similar to the results in the previous section, the above theorems do not require undersmoothing for both the first and second stage nonparametric estimation. Finally we expect that analogous arguments can be applied for over-identified moment conditions, where the dimension of g exceeds that of β. In this case, the likelihood ratio statistic $\min_{b\in B:\theta=\tau(b)} \ell(b) - \min_{b\in B} \ell(b)$ will converge to the $\chi^2(k_1)$ distribution.[11]

3.2. **Additional variables in third step.** We now consider an extension to the moment condition

$$E[g(W, \mu(X, V), \beta)] = 0,$$

where $W \in \mathbb{R}^{d_w}$ is a vector of additional variables. The vector W may contain X and Z as subvectors. This extension is useful to accommodate, for example, partially linear models with generated regressors (see, Section 4.1 below).

Our nonparametric likelihood approach can be modified to accommodate additional variables W as follows. Let $g_1(w, \mu, \beta)$ be the partial derivative of $g(\cdot, \cdot, \cdot)$ with respect to its $(d_w + 1)$-th argument evaluated at (w, μ, β). In the case of a parametric first step (i.e., $V = \varphi(X, Z, \alpha)$), the

[11] It is interesting to extend our inference method to the case where the object of interest depends not only on β but also on the first and second stage parameters. For example, in the partially linear model discussed in Section 4.1, one may be interested in the average marginal effect of X, that is $\beta + E\left[\frac{\partial m(\varphi(X,Z))}{\partial V}\frac{\partial\varphi(X,Z)}{\partial X}\right]$. The analysis for such general cases is more complicated and beyond the scope of this paper.

GEL statistic in (2.3) is modified by replacing $\tilde{g}_i(\beta)$ with

$$\begin{aligned}\tilde{g}_i(\beta) &= g(W_i, \hat{\gamma}(X_i, \hat{V}_i), \beta) + \hat{\Delta}_1 \psi(X_i, Z_i, \hat{\alpha}) + \hat{\Delta}_{2i}\{Y_i - \hat{\gamma}(X_i, \hat{V}_i)\}, \\ \hat{\Delta}_1 &= \frac{1}{n}\sum_{i=1}^{n}\{(g_1(W_i, \hat{\gamma}(X_i, \hat{V}_i), \beta) - \hat{\kappa}(X_i, \hat{V}_i))\hat{\gamma}_v(X_i, \hat{V}_i)\varphi_\alpha(X_i, Z_i, \hat{\alpha})' \\ &\qquad +(Y_i - \hat{\gamma}(X_i, \hat{V}_i))\hat{\kappa}_v(X_i, \hat{V}_i)\varphi_\alpha(X_i, Z_i, \hat{\alpha})'\},\end{aligned}$$

$\hat{\kappa}(X_i, \hat{V}_i)$ and $\hat{\kappa}_v(X_i, \hat{V}_i)$ are the intercept and slope coefficient of $\hat{V}_i$ in a local polynomial regression of $g_1(W_i, \hat{\gamma}(X_i, \hat{V}_i), \beta)$ on $(X_i, \hat{V}_i)$, respectively, and $\hat{\Delta}_{2i} = \hat{\kappa}(X_i, \hat{V}_i)$.

In the case of a nonparametric first step (i.e., $V_* = \varphi(X, Z) = E[U|X, Z]$ for some observable U), the statistic in (2.3) is modified by replacing $\tilde{g}_i(\beta)$ with

$$\begin{aligned}\tilde{g}_i(\beta) &= g(W_i, \hat{\gamma}(X_i, \hat{V}_i), \beta) + \hat{\Delta}_{1i}(U_i - \hat{V}_i) + \hat{\Delta}_{2i}\{Y_i - \hat{\gamma}(X_i, \hat{V}_i)\}, \\ \bar{\Delta}_{1i} &= \{g_1(W_i, \hat{\gamma}(X_i, \hat{V}_i), \beta) - \hat{\kappa}(X_i, \hat{V}_i)\}\hat{\gamma}_v(X_i, \hat{V}_i) + \{Y_i - \hat{\gamma}(X_i, \hat{V}_i)\}\hat{\kappa}_v(X_i, \hat{V}_i),\end{aligned}$$

and $\hat{\Delta}_{1i}$ is the nonparametric regression fit of $\bar{\Delta}_{1i}$ on (X_i, Z_i).

For both cases, it can be shown that the GEL statistic $\ell(\beta)$ converges to the $\chi^2(k)$ distribution (without undersmoothing).

3.3. **Partial mean case.** In this subsection, we consider an extension to

$$E[g(W, \mu_1(X, V), \ldots, \mu_L(X, V), \beta)] = 0,$$

where $\mu_l(X, V) = E[Y|X, V, D = d_{(l)}]$ for $l = 1, \ldots, L$ is a vector of conditional means associated with the discrete variable D supported on the values $d_{(1)}, \ldots, d_{(L)}$. This extension is useful to accommodate matching estimators of treatment effects, for example.

Let $g_{1l}(w, \mu_1, \ldots, \mu_L, \beta)$ be the partial derivative of $g(\cdot, \ldots, \cdot)$ with respect to its $(d_w + l)$-th argument evaluated at $(w, \mu_1, \ldots, \mu_L, \beta)$, $\kappa_l(X_i, V_i) = E[g_{1l}(W_i, \mu_1(X_i, V_i), \ldots, \mu_L(X_i, V_i), \beta)|X_i, V_i]$, and $\pi_l(X_i, V_i) = \Pr\{D_i = d_{(l)}|X_i, V_i\}$. In the case of a parametric first step, the GEL statistic in (2.3) is modified by replacing $\tilde{g}_i(\beta)$ with

$$\begin{aligned}\tilde{g}_i(\beta) &= g(W_i, \hat{\gamma}_1(X_i, \hat{V}_i), \ldots, \hat{\gamma}_L(X_i, \hat{V}_i), \beta) + \hat{\Delta}_1 \psi(X_i, Z_i, \alpha) \\ &\quad + \sum_{l=1}^{L} \mathbb{I}\{D_i = d_{(l)}\}\{Y_i - \hat{\gamma}_l(X_i, \hat{V}_i)\}\frac{\hat{\kappa}_l(X_i, \hat{V}_i)}{\hat{\pi}_l(X_i, \hat{V}_i)},\end{aligned}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, $\hat{\gamma}_l(X_i,\hat{V}_i)$, $\hat{\gamma}_{l,v}(X_i,\hat{V}_i)$, $\hat{\pi}_l(X_i,\hat{V}_i)$, $\hat{\pi}_{l,v}(X_i,\hat{V}_i)$, $\hat{\kappa}_l(X_i,\hat{V}_i)$, and $\hat{\kappa}_{l,v}(X_i,\hat{V}_i)$ are the local polynomial estimators of $\mu_l(X_i,V_i)$, $\partial\mu_l(X_i,V_i)/\partial V_i$, $\pi_l(X_i,V_i)$, $\partial\pi_l(X_i,V_i)/\partial V_i$, $\kappa_l(X_i,V_i)$, and $\partial\kappa_l(X_i,V_i)/\partial V_i$, respectively, and

$$\begin{aligned}
\hat{\Delta}_1 &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l=1}^{L}\begin{pmatrix} g_{1l}(W_i,\hat{\gamma}_1(X_i,\hat{V}_i),\ldots,\hat{\gamma}_L(X_i,\hat{V}_i),\beta) \\ -\frac{\mathbb{I}\{D_i=d_{(l)}\}}{\hat{\pi}_l(X_i,\hat{V}_i)}\hat{\kappa}_l(X_i,\hat{V}_i)\end{pmatrix}\hat{\gamma}_{l,v}(X_i,\hat{V}_i)\right]\varphi_\alpha(X_i,Z_i,\hat{\alpha})' \\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{L}\frac{\mathbb{I}\{D_i=d_{(l)}\}}{\hat{\pi}_l(X_i,\hat{V}_i)}\{Y_i-\hat{\gamma}_l(X_i,\hat{V}_i)\}\hat{\kappa}_l'(X_i,\hat{V}_i)\right)\varphi_\alpha(X_i,Z_i,\hat{\alpha})' \\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{L}\frac{\mathbb{I}\{D_i=d_{(l)}\}}{\hat{\pi}_l(X_i,\hat{V}_i)^2}\{Y_i-\hat{\gamma}_l(X_i,\hat{V}_i)\}\hat{\kappa}_l(X_i,\hat{V}_i)\hat{\pi}_{l,v}(X_i,\hat{V}_i)\right)\varphi_\alpha(X_i,Z_i,\hat{\alpha})'.
\end{aligned}$$

In the case of a nonparametric first step, the statistic in (2.3) is modified by replacing $\tilde{g}_i(\beta)$ with

$$\begin{aligned}
\tilde{g}_i(\beta) &= g(W_i,\hat{\gamma}_1(X_i,\hat{V}_i),\ldots,\hat{\gamma}_L(X_i,\hat{V}_i),\beta)+\hat{\Delta}_{1i}(U_i-\hat{V}_i) \\
&\quad+\sum_{l=1}^{L}\mathbb{I}\{D_i=d_{(l)}\}\{Y_i-\hat{\gamma}_l(X_i,\hat{V}_i)\}\frac{\hat{\kappa}_l(X_i,\hat{V}_i)}{\hat{\pi}_l(X_i,\hat{V}_i)},
\end{aligned}$$

where $\hat{\Delta}_{1i}$ is a nonparametric estimator of

$$\begin{aligned}
\Delta_{1i} &= E\left[\sum_{l=1}^{L}\left(g_{1l}(W_i,\mu_1(X_i,V_i),\ldots,\mu_L(X_i,V_i),\beta)-\frac{\mathbb{I}\{D_i=d_{(l)}\}}{\pi_l(X_i,V_i)}\kappa_l(X_i,V_i)\right)\frac{\partial\mu_l(X_i,V_i)}{\partial V_i}\middle| X_i,Z_i\right] \\
&\quad+E\left[\sum_{l=1}^{L}\frac{\mathbb{I}\{D_i=d_{(l)}\}}{\pi_l(X_i,V_i)}\{Y_i-\mu_l(X_i,V_i)\}\frac{\partial\kappa_l(X_i,V_i)}{\partial V_i}\middle| X_i,Z_i\right] \\
&\quad+E\left[\sum_{l=1}^{L}\frac{\mathbb{I}\{D_i=d_{(l)}\}}{\pi_l(X_i,V_i)^2}\{Y_i-\mu_l(X_i,V_i)\}\kappa_l(X_i,V_i)\frac{\partial\pi_l(X_i,V_i)}{\partial V_i}\middle| X_i,Z_i\right].
\end{aligned}$$

For both cases, it can be shown that the GEL statistic $\ell(\beta)$ converges to the $\chi^2(k)$ distribution (without undersmoothing).

3.4. **Case of multidimensional μ.** Theorem 1 can be generalized to the case of multidimensional μ, where $\mu(X_i,V_i)=(\mu_1(X_i,V_i),\ldots,\mu_L(X_i,V_i))'$ and $\mu_l(X_i,V_i)=E[Y_{l,i}|X_i,V_i]$ for

$l = 1, \ldots, L$. In this case, the GEL statistic in (2.3) is modified by replacing $\tilde{g}_i(\beta)$ with

$$\begin{aligned}\tilde{g}_i(\beta) &= g(\hat{\gamma}(X_i, \hat{V}_i), \beta) + \hat{\Delta}\psi(X_i, Z_i, \hat{\alpha}) + \sum_{l=1}^{L} g_{1l}(\hat{\gamma}(X_i, \hat{V}_i), \beta)\{Y_{l,i} - \hat{\gamma}_l(X_i, \hat{V}_i)\}, \\ \hat{\Delta} &= \sum_{l=1}^{L}\left[\frac{1}{n}\sum_{i=1}^{n}\{Y_{l,i} - \hat{\gamma}_l(X_i, \hat{V}_i)\} g_{2l}(\hat{\gamma}(X_i, \hat{V}_i), \beta)\hat{\gamma}_{l,v}(X_i, \hat{V}_i)\varphi_\alpha(X_i, Z_i, \hat{\alpha})'\right],\end{aligned}$$

where $g_{1l}(\mu, \beta)$ and $g_{2l}(\mu, \beta)$ are the first and second derivatives of $g(\cdot, \cdot)$ with respect to its l-th argument evaluated at (μ, β), respectively, and $\hat{\gamma}_{l,v}(X_i, \hat{V}_i)$ is the slope coefficient of $\hat{V}_i$ in a local polynomial regression of $Y_{l,i}$ on $(X_i, \hat{V}_i)$.

Similarly we can extend Theorem 2 for the nonparametric first step to the case of multidimensional μ. The GEL statistic is modified by replacing $\tilde{g}_i(\beta)$ with

$$\tilde{g}_i(\beta) = g(\hat{\gamma}(X_i, \hat{V}_i), \beta) + \sum_{l=1}^{L} \hat{\Delta}_{1l,i}(U_i - \hat{V}_i) + \sum_{l=1}^{L} g_{1l}(\hat{\gamma}(X_i, \hat{V}_i), \beta)\{Y_{l,i} - \hat{\gamma}_l(X_i, \hat{V}_i)\},$$

where $\hat{\Delta}_{1l,i}$ is the nonparametric regression fitted value of $\delta_{1l,i} = \{Y_{l,i} - \hat{\gamma}_l(X_i, \hat{V}_i)\} g_{2l}(\hat{\gamma}(X_i, \hat{V}_i), \beta)\hat{\gamma}_{l,v}(X_i, \hat{V}_i)$ on (X_i, Z_i).

3.5. **Other nonparametric likelihood functions.** The GEL statistic in (2.3) can be further generalized to allow different criterion functions for the construction of the objective function and implied weights, such as the exponentially tilted empirical likelihood in Schennach (2007) and the generalized power divergence family in Camponovo and Otsu (2014). By using the dual form in (2.5), the general family of statistics can be defined as

$$\bar{\ell}(\beta) = 2\sum_{i=1}^{n} \rho(\bar{\lambda}'\tilde{g}_i(\beta)) - 2n\rho(0),$$

where $\bar{\lambda} = \arg\max_{\lambda \in \Lambda_n(\beta)} \sum_{i=1}^{n} \bar{\rho}(\lambda'\tilde{g}_i(\beta))$ for possibly another criterion $\bar{\rho}$. The GEL statistic in (2.3) corresponds to the case of $\rho = \bar{\rho}$, and the exponentially tilted empirical likelihood corresponds to the case of $\rho(v) = \log(1 - v)$ and $\bar{\rho}(v) = -e^v$.

By adding analogous assumptions on $\bar{\rho}$ (as in Assumption P (ii)), a similar argument as in the proof of Theorem 1 yields that $\bar{\ell}(\beta) \xrightarrow{d} \chi^2(k)$.

4. Examples

4.1. **Partially linear model with generated regressor.** In this subsection, we illustrate our nonparametric likelihood method using a partially linear model with a generated regressor. This

model may be considered as a simplified version of the production function model studied in Olley and Pakes (1996). In particular, we consider inference on the slope parameters β in the partially linear model with an unobservable regressor V:

$$Y = X'\beta + m(V) + \epsilon,$$

where m is an unknown function and $E[\epsilon|X, V] = 0$. The unobservable regressor V is generated by $V = \varphi(X, Z, \alpha)$ (parametric first step) based on observables (X, Z) and φ known up to α, or $V = \varphi(X, Z)$ (nonparametric first step) based on an unknown function φ, which is consistently estimable.[12]

Estimation of β may be interpreted in a three step way. First, we compute the generated regressor $\hat{V}$ as a proxy for V. Second, the functions $\mu_1(v) = E[X|V = v]$ and $\mu_2(v) = E[Y|V = v]$ are estimated by $\hat{\gamma}_1(\hat{V}_i)$ and $\hat{\gamma}_2(\hat{V}_i)$, that is, a nonparametric regression of X on $\hat{V}$ and Y on $\hat{V}$, respectively. Third, the estimator $\hat{\beta}$ can be obtained by solving $n^{-1}\sum_{i=1}^{n}(X_i - \hat{\gamma}_1(\hat{V}_i))\{(Y_i - \hat{\gamma}_2(\hat{V}_i)) - (X_i - \hat{\gamma}_1(\hat{V}_i))'\hat{\beta}\} = 0$. Based on this condition for $\hat{\beta}$, we consider the moment function $g(X, \mu(V), \beta) = (X - \mu_1(V))\{(Y_i - \mu_2(V)) - (X - \mu_1(V))'\beta\}$ to apply our nonparametric likelihood method.

In the case of a parametric first step, using the fact that $m(V) = E[Y - X'\beta|V] = \mu_2(V) - \mu_1(V)'\beta$ and a multidimensional version of Hahn and Ridder (2013, Theorem 4), the influence function of $\hat{\beta}$ is obtained as

$$\begin{aligned}&\{X_i - \mu_1(V_i)\}\{Y_i - X_i'\beta - m(V_i)\}\\ &- \left[E[(Y_i - X_i'\beta - m(V_i))\frac{\partial \mu_1(v)}{\partial v}\varphi_\alpha(X, Z, \alpha)'] + E[(X - \mu_1(V))\frac{\partial m(V)}{\partial V}\varphi_\alpha(X, Z, \alpha)']\right]\sqrt{n}(\hat{\alpha} - \alpha),\end{aligned}$$

The t-ratio is given by estimating the asymptotic variance of this function. We note that by Newey (1994, Proposition 2), there is no contribution from $\hat{\gamma}_1$ and $\hat{\gamma}_2$ in this example.

[12]For example, in Olley and Pakes (1996), $V = \varphi(X, Z)$ corresponds to conditional means $E[y_{t-1}|i_{t-1}, k_{t-1}, a_{t-1}]$ and $E[l_{t-1}|i_{t-1}, k_{t-1}, a_{t-1}]$ where $(y_{t-1}, l_{t-1}, i_{t-1}, k_{t-1})$ are logs of the output, labor inputs, investment, capital inputs at a previous period, respectively, and a_{t-1} is the firm's age. X corresponds to (k_{t-1}, a_{t-1}), and Z corresponds to i_{t-1}. If we parametrize these conditional means and estimate by e.g. least squares, then it will be the case of parametric first step. If we nonparametrically estimate these conditional means by the kernel estimator, it will be the case of nonparametric first step.

By applying the result in Section 3.2, the GEL statistic is defined by (2.3) with

$$\begin{aligned}\tilde{g}_i(\beta) &= \{X_i - \hat{\gamma}_1(\hat{V}_i)\}\{Y_i - X_i'\beta - \hat{m}(\hat{V}_i)\} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \{Y_i - X_i'\beta - \hat{m}(\hat{V}_i)\}\hat{\gamma}_{1,v}(\hat{V}_i)\varphi_\alpha(X_i, Z_i, \hat{\alpha})' \\ +\{X_i - \hat{\gamma}_1(\hat{V}_i)\}\hat{m}_v(\hat{V}_i)\varphi_\alpha(X_i, Z_i, \hat{\alpha})' \end{bmatrix}\psi(X_i, Z_i, \hat{\alpha}),\end{aligned}$$

where $\hat{\gamma}_1(v)$, $\hat{\gamma}_{1,v}(v)$, $\hat{m}(v)$, and $\hat{m}_v(v)$ are the nonparametric estimators of $\mu_1(v)$, $\mu_{1,v}(v) = \frac{\partial \mu_1(v)}{\partial v}$, $\{\mu_2(v) - \mu_1(v)'\beta\}$, and $\{\mu_{2,v}(v) - \mu_{1,v}(v)'\beta\}$, respectively.[13]

In the case of a nonparametric first step, the GEL statistic can be defined by (2.3) with

$$\tilde{g}_i(\beta) = \{X_i - \hat{\gamma}_1(\hat{V}_i)\}\{Y_i - X_i'\beta - \hat{m}(\hat{V}_i)\} + \hat{\Delta}_{1i}(U_i - \hat{V}_i),$$

where $\hat{\Delta}_{1i}$ is the nonparametric regression fit of $[-\{Y_i - X_i'\beta - \hat{m}(\hat{V}_i)\}\hat{\gamma}_{1,v}(\hat{V}_i) - \{X_i - \hat{\gamma}_1(\hat{V}_i)\}\hat{m}_v(\hat{V}_i)]$ on (X_i, Z_i).

For this example, our main theorems in Section 2 can be applied as follows. We adapt Assumptions P (i) and NP (i) to this example.

Assumption P1. In addition to Assumption P (ii)-(v), suppose

(i): $\{Y_i, X_i', Z_i'\}_{i=1}^n$ is an iid sample from $(Y, X', Z') \in \mathbb{R} \times \mathbb{X} \times \mathbb{Z}$. $\mathbb{X}$, $\mathbb{Z}$, and $\mathbb{V}$ are compact. The density $f(v)$ of V is continuously differentiable to order $s \geq 2$ and bounded away from zero on $\mathbb{V}$. $\mu_1(v)$ and $\mu_2(v)$ are continuously differentiable to order $s \geq 2$ on $\mathbb{V}$. For some $p \geq 4$, $E|X|^p < \infty$, $E|Y|^p < \infty$, and $E[|X|^p|V = v]f(v)$ and $E[|Y|^p|V = v]f(v)$ are bounded over $\mathbb{X} \times \mathbb{V}$.

(ii): For some neighborhood $\mathcal{N}$ of α, $\varphi_{\alpha\alpha}(x, z, \alpha)$ is continuous over $\mathbb{X} \times \mathbb{Z} \times \mathcal{N}$.

Assumption NP1. In addition to Assumptions P (ii)-(iv) and P1 (i), suppose

(i): As $n \to \infty$, it holds $n^{1/2}b^{d_x+d_z}/\log n \to \infty$ and $nb^{4s} \to 0$.

(ii): The joint density $f(x, z)$ of (X, Z) is continuously differentiable to order $s \geq 2$ and bounded away from zero on $\mathbb{X} \times \mathbb{Z}$. The functions $\varphi(x, z) = E[U|X = x, Z = z]$ and $\delta(x, z) = E[\eta|X = x, Z = z]$, where $\eta = \epsilon\mu_{1,v}(V) + \{X - \mu_1(V)\}\{\mu_{2,v}(V) - \mu_{1,v}(V)'\beta\}$, are continuously differentiable to order $s \geq 2$ on $\mathbb{X} \times \mathbb{Z}$. For some $p \geq 4$, $E|U|^p < \infty$,

[13]For example, based on $\hat{\beta}$ obtained above, $\hat{m}(v)$ and $\hat{m}_v(v)$ can be constructed by the local linear regression from the residual $(Y_i - X_i\hat{\beta})$ on the regressor $(\hat{V}_i - v)$, where the intercept and slope estimates correspond to the ones for $\hat{m}(v)$ and $\hat{m}_v(v)$, respectively.

$E|\eta|^p < \infty$, and $E[|U|^p|X = x, Z = z]f(x,z)$ and $E[|\eta|^p|X = x, Z = z]f(x,z)$ are bounded on $\mathbb{X} \times \mathbb{Z}$.

Then analogous arguments yield the limiting distribution of the GEL statistic.

Proposition. Consider the setup of this subsection. Suppose either Assumption P1 (for the parametric first step $V = \varphi(X, Z, \alpha)$) or Assumption NP1 (for the nonparametric first step $V = \varphi(X, Z)$) holds true. Then $\ell(\beta) \xrightarrow{d} \chi^2(k)$.

4.2. **Average treatment effect and counterpart on treated population.** In this subsection, we consider the propensity score matching estimators for the average treatment effect and the one for the treated population. 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 vector of covariates.

First, we consider inference on the average treatment effect $\beta = E[Y_i(1) - Y_i(0)]$. Let $\varphi(x) = \Pr\{D = 1|X = x\}$ be the propensity score and $\hat{\varphi}(x)$ be its nonparametric estimator (i.e., a nonparametric regression of D on X). Also let $\hat{\gamma}_1(\cdot)$ and $\hat{\gamma}_0(\cdot)$ be the nonparametric regression fits from Y on $\hat{\varphi}(X)$ for the treated and untreated samples, respectively. Then the propensity score matching estimator by Heckman, Ichimura and Todd (1998) is defined as $\hat{\beta} = \frac{1}{n}\sum_{i=1}^n \{\hat{\gamma}_1(\hat{\varphi}(X_i)) - \hat{\gamma}_0(\hat{\varphi}(X_i))\}$. This can be interpreted as the method of moments estimator using the moment function $g(X, \mu(V), \beta) = \mu_1(V) - \mu_0(V) - \beta$, where $\mu_1(v) = E[Y|V = v, D = 1]$, $\mu_0(v) = E[Y|V = v, D = 0]$, and $V = \varphi(X)$.

From Hahn and Ridder (2013, Section 4), the influence function of the propensity score matching estimator $\hat{\beta}$ is given by

$$\begin{aligned} & (\mu_1(V_i) - \mu_0(V_i) - \beta) - \left(\frac{m_1(X_i) - \mu_1(V_i)}{\varphi(X_i)} + \frac{m_0(X_i) - \mu_0(V_i)}{1 - \varphi(X_i)}\right)(D_i - \varphi(X_i)) \\ & \quad + \left(\frac{D_i}{\varphi(X_i)}(Y_i - \mu_1(V_i)) - \frac{1 - D_i}{1 - \varphi(X_i)}(Y_i - \mu_0(V_i))\right) \\ = \; & (m_1(X_i) - m_0(X_i) - \beta) + \frac{D_i}{\varphi(X_i)}(Y_i - m_1(X_i)) - \frac{1 - D_i}{1 - \varphi(X_i)}(Y_i - m_0(X_i)), \end{aligned} \tag{4.1}$$

where $m_1(x) = E[Y|X = x, D = 1]$ and $m_0(x) = E[Y|X = x, D = 0]$. By applying the result in Section 3.3, the GEL statistic is defined by (2.3) with

$$\tilde{g}_i(\beta) = (\hat{m}_1(X_i) - \hat{m}_0(X_i) - \beta) + \frac{D_i}{\hat{\varphi}(X_i)}(Y_i - \hat{m}_1(X_i)) - \frac{1 - D_i}{1 - \hat{\varphi}(X_i)}(Y_i - \hat{m}_0(X_i)). \tag{4.2}$$

We note that the influence function in (4.1) is identical for other asymptotically efficient estimators, such as the inverse probability weighted estimator (Hirano, Imbens and Ridder, 2003). Indeed, BEV modified the moment function for the inverse probability weighted estimator and obtained the same function in (4.2).[14] Also it is interesting to note that the correction terms (i.e., the second and third terms in (4.2)) are analogous to additional terms in semiparametric doubly robust estimators (see, Cattaneo, 2010, and Rothe and Firpo, 2016). Rothe and Firpo (2016) showed that in this setup, the semiparametric doubly robust estimator has smaller first order bias and second order variance compared to other estimators. Indeed both this paper and Rothe and Firpo (2016) utilize the same bias cancellation property in $\tilde{g}_i(\beta)$ (see, Remark 5) for valid inference without undersmoothing and point estimation with smaller bias, respectively.

Next, we consider the average treatment effect on the treated population $\beta = E[Y_i(1) - Y_i(0)|D_i = 1]$. To simplify the presentation, we assume $p = \Pr\{D_i = 1\}$ is known as in Hahn and Ridder (2013). In this case, from Hahn and Ridder (2013, Section 4), the influence function of the propensity score matching estimator $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{p}\{\hat{\gamma}_1(\hat{\varphi}(X_i)) - \hat{\gamma}_0(\hat{\varphi}(X_i))\}$ is given by

$$\begin{aligned}
&\frac{D_i}{p}(\mu_1(V_i) - \mu_0(V_i) - \beta) - \frac{m_0(X_i) - \mu_0(V_i)}{p(1 - \varphi(X_i))}(D_i - \varphi(X_i)) \\
&\quad + \left(\frac{D_i}{p}(Y_i - \mu_1(V_i)) - \frac{(1 - D_i)\varphi(X_i)}{p(1 - \varphi(X_i))}(Y_i - \mu_0(V_i))\right) \\
= \;& \frac{D_i}{p}(m_1(X_i) - m_0(X_i) - \beta) + \frac{D_i}{p}(Y_i - m_1(X_i)) - \frac{(1 - D_i)\varphi(X_i)}{p(1 - \varphi(X_i))}(Y_i - m_0(X_i)).
\end{aligned}$$

By applying the result in Section 3.3, the GEL statistic is defined by (2.3) with

$$\tilde{g}_i(\beta) = \frac{D_i}{p}(\hat{m}_1(X_i) - \hat{m}_0(X_i) - \beta) + \frac{D_i}{p}(Y_i - \hat{m}_1(X_i)) - \frac{(1 - D_i)\hat{\varphi}(X_i)}{p(1 - \hat{\varphi}(X_i))}(Y_i - \hat{m}_0(X_i)),$$

where $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$ are nonparametric estimators of $m_1(X_i)$ and $m_0(X_i)$, respectively.

[14] Primitive conditions for $\ell(\beta) \xrightarrow{d} \chi^2(1)$ are provided in Section 4.2 of BEV.

5. Simulation

This section conducts simulation studies to evaluate the finite sample properties of our semiparametric GEL inference method. We consider inference on (i) the average treatment effect, and (ii) a sample selection model whose implied structure is essentially the same as the partial linear model with a generated regressor discussed in Section 4.1.

5.1. **Average treatment effect.** We adopt the simulation design in Ichimura and Linton (2005) and consider inference on the average treatment effect $\beta = E[Y(1)] - E[Y(0)]$. The data generating process is

$$\begin{aligned} X &\sim U[-0.5, 0.5], \qquad T = \mathbb{I}\{X\alpha + \epsilon > 0\}, \\ Y(0) &= 2X + \eta, \qquad Y(1) = Y(0) + \beta, \end{aligned}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, $\alpha = 1$, $\beta = 0$, and (η, ϵ) are mutually independent standard normal random variables. We consider the models where the propensity score $\Pr\{T = 1|X\}$ is nonparametric (Model NP), and parametric $\Pr\{T = 1|X\} = \Phi(X\alpha)$ with the standard normal distribution function $\Phi(\cdot)$ (Model P). For the parametric case, α is estimated by the maximum likelihood. The sample size is $n = 100$, and the results are based on $1,000$ Monte Carlo replications.

We compare the confidence sets for β constructed by (i) Wald-type method (Wald) based on the propensity score matching estimator by Heckman, Ichimura and Todd (1998), (ii) adjusted empirical likelihood (AEL), (iii) semiparametric empirical likelihood (SPEL), (iv) semiparametric exponential tilting (SPET), and (v) semiparametric continuous updating GMM (SPCU). All methods are implemented by the Gaussian kernel. Wald is the conventional approach, SPEL, SPET, and SPCU are our proposals, and AEL is based on the unadjusted moment function (i.e., the first term of $\tilde{g}_i(\beta)$ in (2.4) or (2.8)) followed by a multiplicative correction. More precisely, the confidence set by AEL is

$$\{\beta : \hat{\rho} \cdot \ell^{\text{unadjusted}}(\beta) \leq \chi^2_{1-\alpha}(1)\},$$

where $\ell^{\text{unadjusted}}(\beta)$ is the empirical likelihood ratio $2\sup_\lambda \sum_{i=1}^n \log(1+\lambda g_i(\beta))$ with $g_i(\beta) = \hat{\gamma}_1(\hat{\varphi}(X_i)) - \hat{\gamma}_0(\hat{\varphi}(X_i))$ and $\hat{\rho} = \frac{\sum_{i=1}^n g_i(\hat{\beta})^2}{\sum_{i=1}^n \tilde{g}_i(\hat{\beta})^2}$ is the adjustment term to recover the asymptotic pivotalness.[15]

Table 1 presents empirical coverages of these confidence sets with 0.95 nominal coverage. We consider five different fixed bandwidths: $h_1 = cS_x n^{-1/5}$ for the first step in Model NP, and $h = cS_v n^{-1/5}$ for the second step in both Models NP and P with $c \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$, where S_x and S_v are the sample standard deviations of X and $\hat{V}$, respectively. We observe that Wald and AEL tend to under-cover for large bandwidths, while the proposed GEL methods (SPEL, SPET, and SPCU) are typically less sensitive to the bandwidths for both Models NP and P.

We also investigate the power properties of the tests for $H_0 : \beta = 0$ under the alternative hypotheses $H_1 : \beta = \Delta$ for $\Delta = -0.4, -0.2, 0.2, 0.4$. Table 2 reports the calibrated powers of all the tests across 1,000 replications (i.e., the rejection frequencies of the tests where the critical values are given by the Monte Carlo 95th percentiles of these test statistics under H_0) with $c = 1$ for the bandwidths. We find that the proposed GEL methods outperform the conventional Wald and AEL methods.

c	Wald	AEL	SPEL	SPET	SPCU	Wald	AEL	SPEL	SPET	SPCU
			Model NP					Model P		
0.5	0.933	0.931	0.942	0.940	0.941	0.928	0.931	0.932	0.930	0.933
1.0	0.941	0.880	0.945	0.944	0.945	0.940	0.897	0.941	0.942	0.944
1.5	0.943	0.801	0.946	0.946	0.948	0.936	0.852	0.944	0.943	0.945
2.0	0.916	0.761	0.946	0.945	0.947	0.933	0.823	0.945	0.944	0.945
2.5	0.875	0.740	0.946	0.942	0.944	0.930	0.775	0.944	0.945	0.944

TABLE 1. Empirical coverages of nominal 95% confidence intervals ($n = 100$)

Δ	Wald	AEL	SPEL	SPET	SPCU	Wald	AEL	SPEL	SPET	SPCU
			Model NP					Model P		
-0.4	0.717	0.342	0.754	0.761	0.759	0.771	0.551	0.784	0.786	0.786
-0.2	0.247	0.179	0.253	0.252	0.254	0.306	0.204	0.315	0.313	0.311
0.2	0.206	0.141	0.302	0.299	0.298	0.242	0.167	0.248	0.246	0.245
0.4	0.662	0.409	0.782	0.784	0.789	0.779	0.596	0.789	0.789	0.788

TABLE 2. Calibrated powers of tests under $H_1 : \beta = \Delta$ (5% size, $n = 100$)

[15] In Model P, $\tilde{g}_i(\beta) = g_i(\beta) + \hat{\Delta}\psi(X_i, \hat{\alpha})$, where

$$\hat{\Delta} = -\frac{1}{n}\sum_{i=1}^n \left(\frac{\hat{m}_1(X_i) - \hat{\gamma}_1(\varphi(X_i,\hat{\alpha}))}{\varphi(X_i,\hat{\alpha})} + \frac{\hat{m}_0(X_i) - \hat{\gamma}_0(\varphi(X_i,\hat{\alpha}))}{1-\varphi(X_i,\hat{\alpha})} \right) \frac{\partial \varphi(X_i,\hat{\alpha})}{\partial \alpha}.$$

5.2. **Sample selection model.** We consider the following sample selection model:

$$\begin{aligned} Y_i &= \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + \epsilon_i, \quad \text{where } Y_i \text{ is only observed if } D_i = 1, \\ D_i &= \mathbb{I}\{\alpha_0 + W_i\alpha_1 + X_{1i}\alpha_2 + X_{2i}\alpha_3 + \eta_i > 0\}, \end{aligned}$$

for $i = 1, \ldots, n$, where $(\beta_0, \beta_1, \beta_2, \alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-1, 1, 1, -0.1, 0.1, -0.1, 0.1)$, $W_i \sim U[0, 10]$, $X_{ji} = 0.2W_i + \sqrt{1 - 0.2^2}X^*_{ji}$, $X^*_{ji} \sim U[0, 10]$ for $j = 1, 2$, and $(\epsilon_i, \eta_i) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right)$.[16] The sample size is $n = 200$, and the results are based on 1,000 Monte Carlo replications.

First, we consider the model where both the joint distribution of the error terms and the functional form of the selection equation are of unknown forms (Model NP). An implication of this model is that (see, e.g., Ahn and Powell, 1993)

$$E[Y|W, X_1, X_2, D = 1] = X_1\beta_1 + X_2\beta_2 + m(V), \qquad \text{where } V = E[D|W, X_1, X_2].$$

Second, we consider the model where an additional single-index restriction $f(W, X_1, X_2) = \alpha_0 + W\alpha_1 + X_1\alpha_2 + X_2\alpha_3$ is imposed (Model P). This model implies (see, e.g., Powell, 2001, and Newey, 2009)

$$E[Y|W, X_1, X_2, D = 1] = X_1\beta_1 + X_2\beta_2 + m(V), \qquad \text{where } V = W\alpha_1 + X_1\alpha_2 + X_2\alpha_3.$$

We employ Ichimura's (1993) estimator to estimate $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ in the first step.[17]

We compare four methods (Wald, SPEL, SPET, and SPCU) to construct confidence sets for β_2. All methods are implemented by the Gaussian kernel. Table 3 presents empirical coverages of these confidence sets with 95% nominal coverage. We consider five different bandwidths: $h_w = cS_w n^{-1/7}$ and $h_{x_j} = cS_{x_j} n^{-1/7}$ $(j = 1, 2)$ for the first step in Model NP, and $h = cS_v n^{-1/5}$ for the second step in both Models NP and P, and $c \in \{0.5, 0.7, 1.0, 1.3, 1.5\}$, where S_w, S_{x_1}, S_{x_2} and S_v are the sample standard deviations of W, X_1, X_2 and $\hat{V}$, respectively. For Model NP, Wald tends to over-cover for large bandwidths, while the proposed GEL methods are less sensitive to the bandwidths. For Model P, all the methods exhibit similar coverage properties.[18]

[16] In a preliminary simulation study, we also consider heteroskedastic error terms with $\epsilon^*_i = (1 + 0.02x_i^2)\epsilon_i$ and $\eta^*_i = (1 + 0.02x_i^2)\eta_i$. Since the results are similar, we only present the results for the homoskedastic case.

[17] The bandwidth is chosen as $h = 1.06S_v n^{-1/5}$, where S_v is the sample standard deviation of $W\hat{\alpha}_1 + X_1\hat{\alpha}_2 + X_2\hat{\alpha}_3$.

[18] Both Wald and AEL methods do not require undersmoothing and allow the MSE optimal bandwidth for the second step in this model because the moment condition for the parametric component of the partially linear model has the double robustness property (see, Rothe and Firpo, 2016).

We also investigate the power properties of the tests for $H_0 : \beta_2 = 1$ under the alternative hypotheses $H_1 : \beta_2 = 1 + \Delta$ for $\Delta = -0.2, -0.1, 0.1, 0.2$. Table 4 reports the calibrated powers of all the tests across 1,000 replications with $c = 1$ for the bandwidth. The proposed GEL methods have slightly better power for Model NP, while all the methods exhibit similar power for Model P.

c	Wald	SPEL	SPET	SPCU	Wald	SPEL	SPET	SPCU
		Model NP				Model P		
0.5	0.954	0.964	0.964	0.966	0.958	0.948	0.950	0.955
0.7	0.967	0.955	0.953	0.955	0.958	0.955	0.955	0.963
1.0	0.976	0.949	0.951	0.956	0.966	0.965	0.963	0.966
1.3	0.975	0.966	0.965	0.970	0.972	0.970	0.969	0.975
1.5	0.971	0.952	0.950	0.953	0.980	0.980	0.979	0.981

TABLE 3. Empirical coverages of nominal 95% confidence intervals ($n = 200$)

Δ	Wald	SPEL	SPET	SPCU	Wald	SPEL	SPET	SPCU
		Model NP				Model P		
-0.2	0.992	0.996	0.997	0.997	0.844	0.844	0.845	0.847
-0.1	0.771	0.790	0.793	0.794	0.492	0.498	0.494	0.488
0.1	0.630	0.642	0.650	0.630	0.506	0.513	0.511	0.516
0.2	0.990	0.990	0.991	0.991	0.858	0.858	0.855	0.860

TABLE 4. Calibrated powers of tests under $H_1 : \beta = 1 + \Delta$ (5% size, $c = 1$, $n = 200$)

6. Conclusion

In this paper we propose a nonparametric likelihood inference method for parameters defined in three step estimation problems considered in Hahn and Ridder (2013). In particular, we show that the generalized empirical likelihood statistic based on moment functions modified to account for influences from three step estimation is asymptotically pivotal without undersmoothing in the first and second step nonparametric estimates. Our method is illustrated by a partially linear model with a generated regressor and propensity score matching estimators. Finally, as mentioned in the remarks and footnotes, there are several directions of future research, such as an extension of the proposed method to weakly dependent data, formal analysis for plug-in estimators using series estimation methods, higher-order analysis to develop an optimal bandwidth selection method, and inference on more general objects which may depend on the first and second stage parameters.

Appendix A. Appendix for Theorem 1

Hereafter, we use the following notation. By suppressing dependence on $(X_j - x)/h$, define

$$\begin{aligned}
\xi_j(v) &= [1, (X_j - x)/h, (V_j - v)/h]', \qquad \hat{\xi}_j(v) = [1, (X_j - x)/h, (\hat{V}_j - v)/h]', \\
\Phi(V_j, v) &= e_1' \left[\frac{1}{nh^{d_x+1}} \sum_{j=1}^n \xi_j(v)\xi_j(v)' K\left(\frac{X_j - x}{h}, \frac{V_j - v}{h} \right) \right]^{-1} \xi_j(v) K\left(\frac{X_j - x}{h}, \frac{V_j - v}{h} \right), \\
\Phi(\hat{V}_j, v) &= e_1' \left[\frac{1}{nh^{d_x+1}} \sum_{j=1}^n \hat{\xi}_j(v)\hat{\xi}_j(v)' K\left(\frac{X_j - x}{h}, \frac{\hat{V}_j - v}{h} \right) \right]^{-1} \hat{\xi}_j(v) K\left(\frac{X_j - x}{h}, \frac{\hat{V}_j - v}{h} \right).
\end{aligned}$$

where $e_1 = (1, 0, \ldots, 0)'$. Then we denote

$$\begin{aligned}
\hat{\mu}(X_i, V_i) &= \frac{1}{nh^{d_x+1}} \sum_{j=1}^n \Phi(V_j, V_i) Y_j, \\
\hat{\gamma}(X_i, V_i) &= \frac{1}{nh^{d_x+1}} \sum_{j=1}^n \Phi(\hat{V}_j, V_i) Y_j, \\
\hat{\gamma}(X_i, \hat{V}_i) &= \frac{1}{nh^{d_x+1}} \sum_{j=1}^n \Phi(\hat{V}_j, \hat{V}_i) Y_j.
\end{aligned}$$

Recall that $\hat{\mu}(X_i, V_i)$ is (infeasible) nonparametric regression of Y_i on (X_i, V_i) as in Hahn and Ridder (2013). Also, let $\Phi_v(\cdot, \cdot)$ be the derivative with respect to its second argument, $\varphi_{\alpha,i} = \varphi_\alpha(X_i, Z_i, \alpha)$, and $\Omega = E[\xi\xi']$, where

$$\xi = g(\mu(X, V), \beta) + \Delta\psi(X, Z, \alpha) + g_1(\mu(X, V), \beta)\{Y - \mu(X, V)\}.$$

A.1. **Lemmas.**

Lemma A.1. *Under Assumption P,*

$$\begin{aligned}
\max_{1 \le i \le n} |\hat{\mu}(X_i, V_i) - \mu(X_i, V_i)| &= o_p(n^{-1/4}), \\
\max_{1 \le i \le n} |\hat{\gamma}(X_i, V_i) - \mu(X_i, V_i)| &= o_p(n^{-1/4}), \\
\max_{1 \le i \le n} |\hat{\gamma}(X_i, \hat{V}_i) - \mu(X_i, V_i)| &= o_p(n^{-1/4}).
\end{aligned}$$

Proof. By Assumption P (i), both X_i and V_{*i} are compactly supported, and their joint density is bounded away from zero. Thus, an application of Hansen (2008, Theorem 10) yields the first statement. The second and third statements follow by expansions around $\hat{\alpha} = \alpha$ combined with $\sqrt{n}(\hat{\alpha} - \alpha) = O_p(1)$ (by Assumption P (v)) and the first statement.

Lemma A.2. *Under Assumption P,*

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g(\hat{\gamma}(X_i,\hat{V}_i),\beta)-g(\hat{\gamma}(X_i,V_i),\beta)\}\\ =\;& E[g_1(\mu(X_i,V_i),\beta)\mu_v(X_i,V_i)\varphi'_{\alpha,i}]\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1).\end{aligned}$$

Proof. Observe that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g(\hat{\gamma}(X_i,\hat{V}_i),\beta)-g(\hat{\gamma}(X_i,V_i),\beta)\}\\ =\;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_1(\mu(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^{n}\{\Phi(\hat{V}_j,\hat{V}_i)-\Phi(\hat{V}_j,V_i)\}Y_j+o_p(1)\\ =\;& \frac{1}{n}\sum_{i=1}^{n}g_1(\mu(X_i,V_i),\beta)\left\{\frac{1}{nh^{d_x+1}}\sum_{j=1}^{n}\Phi_v(V_j,V_i)Y_j\right\}\varphi'_{\alpha,i}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1)\\ =\;& \frac{1}{n}\sum_{i=1}^{n}g_1(\mu(X_i,V_i),\beta)\mu_v(X_i,V_i)\varphi'_{\alpha,i}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1)\\ =\;& E[g_1(\mu(X_i,V_i),\beta)\mu_v(X_i,V_i)\varphi'_{\alpha,i}]\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1),\end{aligned}$$

where the first equality follows from expansions around $\hat{\gamma}(X_i,\hat{V}_i)=\hat{\gamma}(X_i,V_i)$ and $\hat{\gamma}(X_i,V_i)=\mu(X_i,V_i)$, Lemma A.1, and boundedness of h_2 (by Assumption P (i)), the second equality follows from an expansion around $\hat{\alpha}=\alpha$ and $\sqrt{n}(\hat{\alpha}-\alpha)=O_p(1)$ (by Assumption P (v)) combined with boundedness of $g_1(\mu(x,v),\beta)$ over $\mathbb{X}\times\mathbb{V}$ and $\varphi_\alpha(x,z,\alpha)$ and $\varphi_{\alpha\alpha}(x,z,\alpha)$ over $\mathbb{X}\times\mathbb{Z}\times\mathcal{N}$ (Assumption P (i)), the third equality follows from the uniform convergence of the derivative of the local linear estimator, and the last equality follows from the law of large numbers.

Lemma A.3. *Under Assumption P,*

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g(\hat{\gamma}(X_i,V_i),\beta)-g(\hat{\mu}(X_i,V_i),\beta)\}\\ =\;& -E[g_1(\mu(X_i,V_i),\beta)\mu_v(X_i,V_i)\varphi'_{\alpha,i}]\sqrt{n}(\hat{\alpha}-\alpha)+\Delta\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1).\end{aligned}$$

Proof. Let $\mu_{xv,i}=\left(\mu(X_i,V_i),\frac{\partial\mu(X_i,V_i)}{\partial x}h,\frac{\partial\mu(X_i,V_i)}{\partial v}h\right)'$. Decompose

$$Y_j=\mu'_{xv,i}\hat{\xi}_j(V_i)-\{\mu'_{xv,i}\hat{\xi}_j(V_i)-\mu'_{xv,i}\xi_j(V_i)\}+\{\mu(X_j,V_j)-\mu'_{xv,i}\xi_j(V_i)\}+\{\mu(X_j,Z_j)-\mu(X_j,V_j)\}+\epsilon_j,$$

where the error term $\epsilon_j = Y_j - \mu(X_j, Z_j)$ satisfies $E[\epsilon_j | X_j, Z_j] = 0$. By this expression, we can write as

$$\hat{\gamma}(X_i, V_i) - \hat{\mu}(X_i, V_i) = m_i^A + m_i^B + m_i^C + m_i^D + m_i^E,$$

where

$$\begin{aligned}
m_i^A &= \frac{1}{nh^{d_x+1}} \sum_{j=1}^{n} \Phi(\hat{V}_j, V_i) \mu'_{xv,i} \hat{\xi}_j(V_i) - \frac{1}{nh^{d_x+1}} \sum_{j=1}^{n} \Phi(V_j, V_i) \mu'_{xv,i} \xi_j(V_i), \\
m_i^B &= -\frac{1}{nh^{d_x+1}} \sum_{j=1}^{n} \Phi(\hat{V}_j, V_i) \{\mu'_{xv,i} \hat{\xi}_j(V_i) - \mu'_{xv,i} \xi_j(V_i)\}, \\
m_i^C &= \frac{1}{nh^{d_x+1}} \sum_{j=1}^{n} \{\Phi(\hat{V}_j, V_i) - \Phi(V_j, V_i)\} \{\mu(X_j, V_j) - \mu'_{xv,i} \xi_j(V_i)\}, \\
m_i^D &= \frac{1}{nh^{d_x+1}} \sum_{j=1}^{n} \{\Phi(\hat{V}_j, V_i) - \Phi(V_j, V_i)\} \{\mu(X_j, Z_j) - \mu(X_j, V_j)\}, \\
m_i^E &= \frac{1}{nh^{d_x+1}} \sum_{j=1}^{n} \{\Phi(\hat{V}_j, V_i) - \Phi(V_j, V_i)\} \epsilon_j.
\end{aligned}$$

Note that $m_i^A = 0$ by construction. Thus, an expansion of $g(\hat{\gamma}(X_i, V_i), \beta)$ around $\hat{\gamma}(X_i, V_i) = \hat{\mu}(X_i, V_i)$ and Lemma A.1 yield

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g(\hat{\gamma}(X_i, V_i), \beta) - g(\hat{\mu}(X_i, V_i), \beta)\} \\
= \; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\hat{\mu}(X_i, V_i), \beta) \{\hat{\gamma}(X_i, V_i) - \hat{\mu}(X_i, V_i)\} + o_p(1) \\
= \; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\hat{\mu}(X_i, V_i), \beta) \{m_i^B + m_i^C + m_i^D + m_i^E\} + o_p(1) \\
\equiv \; & M^B + M^C + M^D + M^E + o_p(1).
\end{aligned}$$

For M^B, we have

$$\begin{aligned}
M^B &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(\hat{\mu}(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^{n}\Phi(\hat{V}_j,V_i)(\mu'_{xv,i}\hat{\xi}_j(V_i)-\mu'_{xv,i}\xi_j(V_i)) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(\hat{\mu}(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^{n}\Phi(\hat{V}_j,V_i)\mu_v(X_i,V_i)(\hat{V}_j-V_j) \\
&= -\frac{1}{n}\sum_{i=1}^{n} g_1(\mu(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^{n}\Phi(V_j,V_i)\mu_v(X_i,V_i)\varphi'_{\alpha,j}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1) \\
&= -\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{nh^{d_x+1}}\sum_{i=1}^{n}\Phi(V_i,V_j)g_1(\mu(X_i,V_i),\beta_*)\mu_v(X_i,V_i)\right\}\varphi'_{\alpha,j}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1) \\
&= -\frac{1}{n}\sum_{j=1}^{n} g_1(\mu(X_j,V_j),\beta)\mu_v(X_j,V_j)\varphi'_{\alpha,j}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1) \\
&= -E[g_1(\mu(X_i,V_i),\beta_*)\mu_v(X_i,V_i)\varphi'_{\alpha,i}]\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1),
\end{aligned}$$

where the first equality is the definition of M^B, the second equality follows from the definitions of $\hat{\xi}_j(V_i)$ and $\xi_j(V_i)$, the third equality follows from expansions around $\hat{\mu}(X_i,V_i)=\mu(X_i,V_i)$ and $\hat{\alpha}=\alpha$ combined with Lemma A.1, $\sqrt{n}(\hat{\alpha}-\alpha)=O_p(1)$, and Assumption P (i), the fourth equality follows by exchanging the order of summations and the fact that $\sum_{i=1}^n\Phi(V_j,V_i)a_i=\sum_{i=1}^n\Phi(V_i,V_j)a_i$ for any a_i (because it is the intercept of the weighted OLS), the fifth equality follows from the uniform convergence of the local linear estimator, and the last equality follows from the law of large numbers.

For M^C, we have

$$\begin{aligned}
M^C &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(\hat{\mu}(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^{n}\{\Phi(V_i,\hat{V}_j)-\Phi(V_i,V_j)\}\{\mu(X_j,V_j)-\mu'_{xv,i}\xi_j(V_i)\} \\
&= \frac{1}{n}\sum_{i=1}^{n} g_1(\mu(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^{n}\Phi_v(V_i,V_j)\{\mu(X_j,V_j)-\mu'_{xv,i}\xi_j(V_i)\}\varphi'_{\alpha,j}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1) \\
&= o_p(1),
\end{aligned}$$

where the first equality is the definition of M^C and the fact that $\sum_{i=1}^n\Phi(V_j,V_i)a_i=\sum_{i=1}^n\Phi(V_i,V_j)a_i$ for any a_i, the second equality follows from expansions around $\hat{\mu}(X_i,V_i)=\mu(X_i,V_i)$ and $\hat{\alpha}=\alpha$ combined with Lemma A.1, $\sqrt{n}(\hat{\alpha}-\alpha)=O_p(1)$, and Assumption P (i), and the third equality follows by exchanging the order of summations and the uniform convergence of the derivative of the local linear estimator.

For M^D, we have

$$\begin{aligned} M^D &= \frac{1}{\sqrt{n}}\sum_{i=1}^n g_1(\hat{\mu}(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^n\{\Phi(V_i,\hat{V}_j)-\Phi(V_i,V_j)\}\{\mu(X_j,Z_j)-\mu(X_j,V_j)\} \\ &= \frac{1}{n}\sum_{i=1}^n g_1(\mu(X_i,V_i),\beta)\frac{1}{nh^{d_x+1}}\sum_{j=1}^n\Phi_v(V_i,V_j)\{\mu(X_j,Z_j)-\mu(X_j,V_j)\}\varphi'_{\alpha,j}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1) \\ &= \frac{1}{n}\sum_{j=1}^n g_2(\mu(X_j,V_j),\beta)\mu_v(X_j,V_j)\{\mu(X_j,Z_j)-\mu(X_j,V_j)\}\varphi'_{\alpha,j}\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1) \\ &= \Delta'\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1), \end{aligned}$$

where the first equality is the definition of M^D and the fact that $\sum_{i=1}^n\Phi(V_j,V_i)a_i=\sum_{i=1}^n\Phi(V_i,V_j)a_i$ for any a_i, the second equality follows from expansions around $\hat{\mu}(X_i,V_i)=\mu(X_i,V_i)$ and $\hat{\alpha}=\alpha$ combined with Lemma A.1, $\sqrt{n}(\hat{\alpha}-\alpha)=O_p(1)$, and Assumption P (i), the third equality follows by exchanging the order of summations and the uniform convergence of the derivative of the local linear estimator, and the last equality follows from the law of large numbers.

For M^E, a similar argument to M^C using $E[\epsilon|X,Z]=0$ yields $M_E=o_p(1)$. Therefore, combining the results for all terms, the conclusion follows.

Lemma A.4. *Under Assumption P,* $\frac{1}{\sqrt{n}}\sum_{i=1}^n\tilde{g}_i(\beta)\xrightarrow{d}N(0,\Omega)$.

Proof. By Lemmas A.2 and A.3,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\{g(\hat{\gamma}(X_i,\hat{V}_i),\beta)-g(\hat{\mu}(X_i,V_i),\beta)\}=\Delta\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1).$$

By this and an expansion of $g(\hat{\mu}(X_i,V_i),\beta)$ around $\hat{\mu}(X_i,V_i)=\mu(X_i,V_i)$, we can decompose

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\tilde{g}_i(\beta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n g(\mu(X_i,V_i),\beta)+M_1+M_2+o_p(1),$$

where

$$\begin{aligned} M_1 &= \Delta\sqrt{n}(\hat{\alpha}-\alpha)+\hat{\Delta}\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(X_i,Z_i,\hat{\alpha}), \\ M_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left[g_1(\mu(X_i,V_i),\beta)\{\hat{\mu}(X_i,V_i)-\mu(X_i,V_i)\}+g_1(\hat{\gamma}(X_i,\hat{V}_i),\beta)\{Y_i-\hat{\gamma}(X_i,\hat{V}_i)\}\right]. \end{aligned}$$

Thus, suppose we have

$$M_1 = \Delta \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, Z_i, \alpha) + o_p(1), \tag{A.1}$$

$$M_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\mu(X_i, V_i), \beta)\{Y_i - \mu(X_i, V_i)\} + o_p(1), \tag{A.2}$$

Then the central limit theorem implies the conclusion.

Since the relation (A.1) follows from Assumption P (v)-(vi), it remains to show (A.2). Decompose

$$M_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\mu(X_i, V_i), \beta)\{Y_i - \mu(X_i, V_i)\} + M_{21} + M_{22} + M_{23},$$

where

$$\begin{aligned}
M_{21} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\mu(X_i, V_i), \beta)\{\hat{\mu}(X_i, V_i) - \mu(X_i, V_i)\}, \\
M_{22} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g_1(\hat{\gamma}(X_i, \hat{V}_i), \beta) - g_1(\mu(X_i, V_i), \beta)\}\{Y_i - \mu(X_i, V_i)\}, \\
M_{23} &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\hat{\gamma}(X_i, \hat{V}_i), \beta)\{\hat{\gamma}(X_i, \hat{V}_i) - \mu(X_i, V_i)\}.
\end{aligned}$$

For M_{22}, the same argument to the proof of Lemma A.2 and Lemma A.3 implies

$$M_{22} = \frac{1}{n} \sum_{i=1} \Delta\{Y_i - \mu(X_i, V_i)\}\sqrt{n}(\hat{\alpha} - \alpha) + o_p(1) = o_p(1).$$

For M_{23}, we further decompose

$$\begin{aligned}
M_{23} &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\mu(X_i, V_i), \beta)\{\hat{\gamma}(X_i, \hat{V}_i) - \mu(X_i, V_i)\} \\
&\quad -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g_1(\hat{\gamma}(X_i, \hat{V}_i), \beta) - g_1(\mu(X_i, V_i), \beta)\}\{\hat{\gamma}(X_i, \hat{V}_i) - \mu(X_i, V_i)\} \\
&= M_{231} + M_{232}.
\end{aligned}$$

From the same argument to the proof of Lemma A.2 and Lemma A.3 (by setting $g(\cdot)$ as the identity map), we have

$$\begin{aligned}
M_{231} &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\mu(X_i, V_i), \beta) \left[\{\hat{\gamma}(X_i, \hat{V}_i) - \hat{\mu}(X_i, V_i)\} + \{\hat{\mu}(X_i, V_i) - \mu(X_i, V_i)\}\right] \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(\mu(X_i, V_i), \beta)\{\hat{\mu}(X_i, V_i) - \mu(X_i, V_i)\} + o_p(1).
\end{aligned}$$

For M_{232}, we have

$$\begin{aligned} M_{232} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\{g_1(\hat{\gamma}(X_i,\hat{V}_i),\beta)-g_1(\hat{\gamma}(X_i,V_i),\beta)\}+\{g_1(\hat{\gamma}(X_i,V_i,\beta))-g_1(\hat{\mu}(X_i,V_i),\beta)\} \\ &\qquad +\{g_1(\hat{\mu}(X_i,V_i),\beta)-g_1(\mu(X_i,V_i),\beta)\}] \\ &\quad \times\Big[\{\hat{\gamma}(X_i,\hat{V}_i)-\hat{\gamma}(X_i,V_i)\}+\{\hat{\gamma}(X_i,V_i)-\hat{\mu}(X_i,V_i)\}+\{\hat{\mu}(X_i,V_i)-\mu(X_i,V_i)\}\Big] \\ &= o_p(1). \end{aligned}$$

The last equality follows the same argument as above combined with the standard argument for degenerated U-statistics.

Finally, note that M_{21} and the main term of M_{231} are cancelled out. Therefore, the conclusion follows.

Lemma A.5. *Under Assumption P,* $\max_{1\le i\le n}|\tilde{g}_i(\beta)| = o_p(n^{1/p})$.

Proof. The proof is similar to that of Newey and Smith (2004, Lemma A1).

Lemma A.6. *Under Assumption P,* $n^{-1}\sum_{i=1}^{n}\tilde{g}_i(\beta)\tilde{g}_i(\beta)' \xrightarrow{p} \Omega$.

Proof. The proof follows by a similar argument to the proof of Lemma A.4.

A.2. **Proof of Theorem 1.** First, by Lemmas A.4, A.5 and A.6, the same arguments as in the proof of Newey and Smith (2004, Lemma A2) imply that $\hat{\lambda} = O_p(n^{-1/2})$.

Next, we obtain an asymptotic approximation for $\hat{\lambda}$. The first-order condition for $\hat{\lambda}$ satisfies

$$0 = \frac{1}{n}\sum_{i=1}^{n}\rho_1(\hat{\lambda}'\tilde{g}_i(\beta))\tilde{g}_i(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\beta) + \frac{1}{n}\sum_{i=1}^{n}\rho_2(\bar{\lambda}'\tilde{g}_i(\beta))\tilde{g}_i(\beta)\tilde{g}_i(\beta)'\hat{\lambda},$$

where the second equality follows from an expansion around $\hat{\lambda} = 0$, and $\bar{\lambda}$ is a point on the line joining $\hat{\lambda}$ and 0. By applying Lemmas A.4, A.5 and A.6, and $\hat{\lambda} = O_p(n^{-1/2})$, we have $\max_{1\le i\le n}|\bar{\lambda}'\tilde{g}_i(\beta)| = o_p(1)$ and

$$\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\beta)\tilde{g}_i(\beta)'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\beta) + o_p(n^{-1/2}). \tag{A.3}$$

Finally, a Taylor expansion yields

$$\begin{aligned} & 2\sum_{i=1}^{n}\rho(\hat{\lambda}'\tilde{g}_i(\beta)) - 2n\rho(0) \\ = \;& 2\sum_{i=1}^{n}\left[\hat{\lambda}'\tilde{g}_i(\beta) - \frac{1}{2}\rho_1(\tilde{\lambda}'\tilde{g}_i(\beta))\hat{\lambda}'\tilde{g}_i(\beta)\tilde{g}_i(\beta)'\hat{\lambda}\right] + o_p(1) \\ = \;& \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{g}_i(\beta)\right)'\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\beta)\tilde{g}_i(\beta)'\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{g}_i(\beta)\right) + o_p(1), \end{aligned} \tag{A.4}$$

where $\tilde{\lambda}$ is a point on the line joining $\hat{\lambda}$ and 0, and the second equality follows from (A.3) and $\max_{1\le i\le n}|\tilde{\lambda}'\tilde{g}_i(\beta)| = o_p(1)$. The conclusion follows by Lemmas A.4 and A.6.

Appendix B. Appendix for Theorem 2

B.1. **Lemmas.**

Lemma B.1. *Under Assumption NP,*

$$\begin{aligned} \max_{1\le i\le n}|\hat{V}_i - V_i| &= o_p(n^{-1/4}), \\ \max_{1\le i\le n}|\hat{\mu}(X_i,V_i) - \mu(X_i,V_i)| &= o_p(n^{-1/4}), \\ \max_{1\le i\le n}|\hat{\gamma}(X_i,V_i) - \mu(X_i,V_i)| &= o_p(n^{-1/4}), \\ \max_{1\le i\le n}|\hat{\gamma}(X_i,\hat{V}_i) - \mu(X_i,V_i)| &= o_p(n^{-1/4}). \end{aligned}$$

Proof. The first statement follows from Assumption NP (i)-(ii) and the same argument as in Lemma A.1. The second statement is the same as in Lemma A.1. The third and fourth statements follow by expansions around $\hat{V}_i = V_{*i}$ combined with the first and second statements.

Lemma B.2. *Under Assumption NP,*

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g(\hat{\gamma}(X_i,\hat{V}_i),\beta) - g(\hat{\gamma}(X_i,V_i),\beta)\} \\ = \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(\mu(X_i,V_i),\beta)\mu_v(X_i,V_i)(\hat{V}_i - V_i) + o_p(1). \end{aligned}$$

Proof. This follows from Lemma B.1 and the same argument as in Lemma A.2.

Lemma B.3. *Under Assumption NP,*

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g(\hat{\gamma}(X_i,V_i),\beta)-g(\hat{\mu}(X_i,V_i),\beta)\}\\
=\;& -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(\mu(X_i,V_i),\beta)\mu_v(X_i,V_i)(\hat{V}_i-V_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i(\hat{V}_i-V_i)+o_p(1).
\end{aligned}$$

Proof. By the same argument as in Lemma A.3, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g(\hat{\gamma}(X_i,V_i),\beta)-g(\hat{\mu}(X_i,V_i),\beta)\}\\
=\;& -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(\mu(X_i,V_i),\beta)\mu_v(X_i,V_i)(\hat{V}_i-V_i)\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_2(\mu(X_j,V_j),\beta)\mu_v(X_j,V_j)\{\mu(X_j,Z_j)-\mu(X_j,V_j)\}(\hat{V}_i-V_i)+o_p(1).
\end{aligned}$$

Applying the standard argument using degenerated U-statistics to the last term yields the conclusion.

Lemma B.4. *Under Assumption NP,* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{g}_i(\beta)\xrightarrow{d}N(0,E[\zeta_i\zeta_i'])$*, where*

$$\zeta_i = g(\mu(X_i,V_i),\beta)+\Delta_i(U_i-V_i)+g_1(\mu(X_i,V_i),\beta)\{Y_i-\mu(X_i,V_i)\}.$$

Proof. By Lemmas B.2 and B.3,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{h(\hat{\gamma}(X_i,\hat{V}_i))-h(\hat{\mu}(X_i,V_i))\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i(\hat{V}_i-V_i)+o_p(1).$$

By this and an expansion of $h(\hat{\mu}(X_i,V_i))$ around $\hat{\mu}(X_i,V_i)=\mu(X_i,V_i)$, we can decompose

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{g}(\beta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g(\mu(X_i,V_i),\beta)+M_1+M_2+o_p(1),$$

where

$$\begin{aligned}
M_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i(\hat{V}_i-V_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\Delta}_{1i}(U_i-\hat{V}_i),\\
M_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g_1(\mu(X_i,V_i),\beta)(\hat{\gamma}(X_i,V_i)-\mu(X_i,V_i))+g_1(\hat{\gamma}(X_i,\hat{V}_i),\beta)\{Y_i-\hat{\gamma}(X_i,\hat{V}_i)\}\right\}.
\end{aligned}$$

Suppose we have

$$M_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i(U_i - V_i) + o_p(1), \tag{B.1}$$

$$M_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(\mu(X_i, V_i), \beta)\{Y_i - \mu(X_i, V_i)\} + o_p(1), \tag{B.2}$$

Then the central limit theorem implies the conclusion. For (B.1), by using the relation that $\hat{V}_i - V_i = (U_i - V_i) - (U_i - \hat{V}_i)$, we have

$$\begin{aligned} M_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i(U_i - V_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\Delta}_{1i} - \Delta_i)(U_i - \hat{V}_i) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i(U_i - V_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\Delta}_{1i} - \Delta_i)(U_i - V_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\Delta}_{1i} - \Delta_i)(\hat{V}_i - V_i) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i(U_i - V_i) + o_p(1). \end{aligned}$$

The last equality follows from the standard argument using degenerated U-statistics. Finally, (B.2) follows from the same argument as in Lemma A.4.

B.2. **Proof of Theorem 2.** We can show Theorem 2 by arguments that are similar to those which were used in the proof of Theorem 1, using Lemmas B.2-B.4. Therefore, we omit the details.

Appendix C. Proofs of Theorems 3 and 4

Since the proofs are similar, we only present the proof of Theorem 3.

Let $\tilde{\beta} = \arg\min_{b:\theta=\tau(b)} \ell(b)$. By proceeding as in Newey and Smith (2004, Theorems 3.1 and 3.2) and Qin and Lawless (1995, eq. (3.6)), it can be shown that (under $\theta = \tau(\beta)$) $\tilde{\beta} \xrightarrow{p} \beta$ and

$$\sqrt{n}(\tilde{\beta} - \beta) = -PG'\Omega^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{g}_i(\beta) + o_p(1), \tag{C.1}$$

where $P = V - VH'(HVH')^{-1}HV$, $V = (G'\Omega^{-1}G)^{-1}$, $H = \frac{d\tau(\beta)}{d\beta'}$, and $G = E\left[\frac{\partial g(\mu(X,V),\beta)}{\partial\beta'}\right]$.

By applying a similar argument to establish (A.4), we obtain

$$\ell_p(\theta) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{g}_i(\tilde{\beta})\right)'\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\tilde{\beta})\tilde{g}_i(\tilde{\beta})'\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{g}_i(\tilde{\beta})\right) + o_p(1).$$

By Lemma A.6 combined with consistency of $\tilde{\beta}$, we have $\frac{1}{n}\sum_{i=1}^n \tilde{g}_i(\tilde{\beta})\tilde{g}_i(\tilde{\beta})' \xrightarrow{p} \Omega$. Also an expansion around $\tilde{\beta} = \beta$ and (C.1) imply

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{g}_i(\tilde{\beta}) = (I - GPG'\Omega^{-1})\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{g}_i(\beta) + o_p(1).$$

Combining these results,

$$\ell_p(\theta) = \left(\Omega^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{g}_i(\beta)\right)' A(A'A)^{-1}A' \left(\Omega^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{g}_i(\beta)\right) + o_p(1),$$

where $A = \Omega^{1/2}(G')^{-1}H'$. Since $\Omega^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{g}_i(\beta) \xrightarrow{d} N(0, I)$ by Lemma A.4 and $A(A'A)^{-1}A'$ is an idempotent matrix with rank k_1, the conclusion follows.

Appendix D. Proof of Proposition

Proof for the case of parametric first step $V = \varphi(X, Z, \alpha)$. It is enough to verify $\hat{\Delta} \xrightarrow{p} \Delta$, where

$$\begin{aligned}
\hat{\Delta} &= \frac{1}{n}\sum_{i=1}^n \left[\{Y_i - X_i'\beta - \hat{m}(\hat{V}_i)\}\hat{\gamma}_{1,v}(\hat{V}_i)\varphi_\alpha(X_i, Z_i, \hat{\alpha})' + \{X_i - \hat{\gamma}_1(\hat{V}_i)\}\hat{m}_v(\hat{V}_i)\varphi_\alpha(X_i, Z_i, \hat{\alpha})'\right], \\
\Delta &= E\left[\{Y - X'\beta - m(V))\}\mu_{1,v}(V)\varphi_\alpha(X, Z, \alpha)' + \{X - \mu_1(V)\}m_v(V)\varphi_\alpha(X, Z, \alpha)'\right].
\end{aligned}$$

with $\mu_{1,v}(V) = \frac{\partial \mu_1(V)}{\partial v}$ and $m_v(V) = \frac{\partial m(V)}{\partial v}$. This follows from the similar argument as in Lemma A.1 (e.g., $\max_{1\le i\le n}|\hat{m}(\hat{V}_i) - m(V_i)| = o_p(1)$).

Proof for the case of nonparametric first step $V = \varphi(X, Z)$. It is enough to verify $\max_{1\le i\le n}|\hat{\Delta}_{1i} - \Delta_i| \xrightarrow{p} 0$. Indeed we have

$$\max_{1\le i\le n}\left|\left[\{Y_i - X_i'\beta - \hat{m}(\hat{V}_i)\}\hat{\gamma}_{1,v}(\hat{V}_i) + \{X_i - \hat{\gamma}_1(\hat{V}_i)\}\hat{m}_v(\hat{V}_i)\right] - [\epsilon_i\mu_{1,v}(V_i) + \{X_i - \mu_1(V_i)\}m_v(V_i)]\right| = o_p(1),$$

from the similar argument as in Lemma B.1 (e.g., $\max_{1\le i\le n}|\hat{m}(\hat{V}_i) - m(V_i)| = o_p(1)$). Then the conclusion follows from the fact that

$$\max_{1\le i\le n}\left|\begin{array}{c}\text{nonparametric regression fit of } \epsilon_i\mu_{1,v}(V_i) + \{X_i - \mu_1(V_i)\}m_v(V_i) \text{ on } (X_i, Z_i) \\ -E[\epsilon_i\mu_{1,v}(V_i) + \{X_i - \mu_1(V_i)\}m_v(V_i)|X_i, Z_i]\end{array}\right| = o_p(1).$$

References

[1] Ahn, H. and J. L. Powell (1993) Semiparametric estimation of censored selection models with a nonparametric selection mechanism, *Journal of Econometrics*, 58, 3-29.

[2] Bravo, F., Chu, B. and D. T. Jacho-Chávez (2017) Semiparametric estimation of moment condition models with weakly dependent data, *Journal of Nonparametric Statistics*, 29, 108-136.

[3] Bravo, F., Escanciano, J. C. and I. van Keilegom (2018) Two-step semiparametric empirical likelihood inference, forthcoming in *Annals of Statistics*.

[4] Camponovo, L. and T. Otsu (2014) On Bartlett correctability of empirical likelihood in generalized power divergence family, *Statistics and Probability Letters*, 86, 38-43.

[5] Cattaneo, M. (2010) Efficient semiparametric estimation of multi-valued treatment effects under ignorability, *Journal of Econometrics*, 155, 138–154.

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

[7] Hansen, B. E. (2008) Uniform convergence rates for kernel estimation with dependent data, *Econometric Theory*, 24, 726-748.

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

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

[10] Ichimura, H. (1993) Estimation of single index models, *Journal of Econometrics*, 58, 71–120.

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

[12] Kitamura, Y. (1997) Empirical likelihood methods with weakly dependent process, *Annals of Statistics*, 25, 2084-2102.

[13] Linton, O. (2002) Edgeworth approximations for semiparametric instrumental variable estimators and test statistics, *Journal of Econometrics*, 106, 325-368.

[14] Mammen, E., Rothe, C. and M. Schienle (2016) Semiparametric estimation with generated covariates, *Econometric Theory*, 32, 1140-1177.

[15] Newey, W. K. (1994) The asymptotic variance of semiparametric estimators, *Econometrica*, 62, 1349-1382.

[16] Newey, W. K. (2009) Two-step series estimation of sample selection models, *Econometrics Journal*, 12, S217-S229.

[17] Newey, W. K., Hsieh, F. and J. M. Robins (2004) Twicing kernels and small bias property of semiparametric estimators, *Econometrica*, 72, 947-962.

[18] Newey, W. K. and R. J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica*, 72, 219-255.

[19] Nishiyama, Y. and P. M. Robinson (2000) Edgeworth expansions for semiparametric averaged derivatives, *Econometrica*, 68, 931-979.

[20] Olley, G. S. and A. Pakes (1996) The dynamics of productivity in the telecommunications equipment industry, *Econometrica*, 64, 1263-1297.

[21] Pagan, A. (1984) Econometric issues in the analysis of regressions with generated regressors, *International Economic Review*, 25, 221-247.

[22] Powell, J. L. (2001) Semiparametric estimation of censored selection models, in Hsiao, C., Morimune, K. and J. Powell (eds.), *Nonlinear Statistical Modeling*, 165-196, Cambridge University Press.

[23] Qin, J. and J. Lawless (1995) Estimating equations, empirical likelihood and constraints on parameters, *Canadian Journal of Statistics*, 23, 145-159.

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

[25] Schennach, S. M. (2007) Point estimation with exponentially tilted empirical likelihood, *Annals of Statistics*, 35, 634-672.

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

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

[28] Zhu, L., Lin, L., Cui, X. and G. Li (2010) Bias-corrected empirical likelihood in a multi-link semiparametric model, *Journal of Multivariate Analysis*, 101, 850-868. d

[29] 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.

Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan.

Email address: matsushita.y@r.hit-u.ac.jp

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

Email address: t.otsu@lse.ac.uk