Manuscript submitted to *The Econometrics Journal*, pp. 1–20.

Testing conditional moment restriction models using empirical likelihood

Yves G. Berger[†]

[†]*Economic, Social and Political Sciences, University of Southampton, Southampton, SO17 1BJ, UK.*
E-mail: y.g.berger@soton.ac.uk

Summary An empirical likelihood test is proposed for parameters of models defined by conditional moment restrictions, such as models with non-linear endogenous covariates, with or without heteroscedastic errors or non-separable transformation models. The number of empirical likelihood constraints is given by the size of the parameter, unlike alternative semi-parametric approaches. We show that the empirical likelihood ratio test is asymptotically pivotal, without explicit studentisation. A simulation study shows that the observed size is close to the nominal level, unlike alternative empirical likelihood approaches. It also offers a major advantages over two-stage least-squares, because the relationship between the endogenous and instrumental variables does not need to be known. An empirical likelihood model specification test is also proposed.

Keywords: *Endogenous covariate, Fourier transform, heteroscedasticity, model-specification, two-stage least-squares.*

1. INTRODUCTION

The theory of empirical likelihood is well established for unconditional moment restrictions. See Owen (2001) and Chen and Van Keilegom (2009) for reviews. In statistical modelling and econometrics, conditional moment restrictions models are often considered (Amemiya, 1977; Chamberlain, 1987; Hansen and Singleton, 1982; Newey, 1993; Ai and Chen, 2003; Domínguez and Lobato, 2004; Smith, 2007; Chen and Pouzo, 2009; Lavergne and Patilea, 2013). These models are defined by the conditional expectation of an estimation function being zero, when evaluated at the target parameter. For example, (non-)linear regression models with endogenous covariates with instrumental variables, models with heteroscedastic errors, transformation models, non-linear (in the parameter) simultaneous equation models or econometric models of optimising agents (Hansen and Singleton, 1982). The generalized method of moments (GMM) procedures may not provide consistent estimator, with non-linear models, because GMM objective functions can lead to several global minima. Examples can be found in Domínguez and Lobato (2004). Conditional moment restrictions cannot always be solved with least-squares or pseudo-maximum likelihood. Nonetheless, it is usually possible to solve them semi-parametrically. The approach proposed offers an inference free from assumption about the distribution of an error term, which is particularly well suited with unknown heteroscedasticity.

Kitamura et al. (2004), Donald et al. (2003) and Chang et al. (2015) developed empirical likelihood-based estimators for conditional moment restrictions. Kitamura et al.'s (2004) "*smoothed empirical likelihood function*" (SEL) is weighted by a kernel function. Donald et al. (2003) approach is based on splines and is a particular case of Chang et al.'s (2015) approach. They are both based on high dimensional constraints; that is, the number of constraints increases with the sample size, even when the size of the parameter is moderate. This makes them computationally heavy. Our approach is low dimensional

and has the merit of being solely based on a finite number of constraints given by the size of the parameter. The correct size is usually achieved with the empirical likelihood test statistics proposed. On the other hand, the size of the SEL test statistics may have the wrong size. Donald et al.'s (2003) empirical likelihood ratio function depends on the number of splines which affects the size. The approach proposed share some similarities with Domínguez and Lobato's (2004) integrated regression technique.

We show that our estimator is $\sqrt{n}$-consistent and its empirical likelihood ratio function is asymptotically pivotal. It is less computationally intensive than Kitamura et al.'s (2004) and Donald et al.'s (2003) approaches, and it can be easily implemented with standard empirical likelihood packages. The proposed test is also more powerful than Domínguez and Lobato's (2004) wald test. The advantage of the proposed approach is its simplicity and the fact that it does not require local smoothing, or a bandwidth selection. The optimisation problem is low dimensional compared to other empirical likelihood-based methods (Kitamura et al., 2004; Donald et al., 2003). However, the proposed estimator does not reach the semiparametric efficiency bound, but an ad-hoc solution is proposed in Section 4.2.

The approach is suitable with regression models with endogenous covariates, when we have an unknown non-linear relationship between the instrumental variables and endogenous covariates. In this case, "*two-stage least squares*" (2SLS) may produce inefficient estimates, because the first stage may not be based on the correct relationship between the instrumental variables and the endogenous covariates. When this relationship is non-linear, we may have a weak correlation and inconsistent 2SLS estimators. Examples can be found in Section S2.1 of the Supplement. With the approach proposed, this relationship does not need to be known. It is also suitable with unknown heteroscedasticity, which usually gives inefficient "*ordinary least squares*" (OLS) estimators and wrong sizes with OLS Wald test. It can be also used with hard-to-estimate transformation model (e.g. Horowitz, 2009, Ch.6), such as Box-Cox transformation models.

In Section 2, we define the moment restriction considered and derive an unconditional moment condition which identifies the parameter. The empirical likelihood approach proposed is given in Section 3. The main results are the asymptotic properties of the approach proposed, which can be found in Section 4. An empirical likelihood model specification test is proposed in Section 5. Section 6 reports briefly some simulation results which support our findings. Detailed and additional simulation studies can be found in Sections S2 and S3 of the supplement. Proofs can be found in Appendix A and in Section S1 of the supplement.

2. CONDITIONAL MOMENT RESTRICTIONS MODELS

Let $Y \in \mathbb{R}^{d_Y}$ and $Z \in \mathbb{R}^{d_Z}$ denote two random vectors. Let Y contains the response variables and some covariates which can be exogenous or endogenous. The vector Z may contain instrumental variables or some of the exogenous variables within Y. The data are n independent realisations $\{(Y_i', Z_i')' : i = 1, \ldots, n\}$. Hereafter, A' shall denote the transpose of A.

Consider non-linear models defined by '*conditional moment restrictions*' given by

$$E\big[\rho(\theta) \mid Z\big] = 0_{d_\rho} \quad a.s., \quad \text{if and only if } \theta = \theta_0, \quad \text{where } \rho(\theta) = \varrho(Y, \theta) \tag{2.1}$$

and $\varrho(\cdot, \cdot)$ is Borel measurable function on $\mathbb{R}^{d_\rho} \times \Theta$, $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and 0_r denotes a $r \times 1$ vector of zeros. We assume $d_\theta < \infty$. Here, $\varrho(\cdot)$ is some given differentiable function. We

assume that θ_0 can be identified by (2.1); that is, for each $\theta \neq \theta_0$, we have $E[\rho(\theta)|Z] \neq 0_{d_\rho}$ with $\mathbb{P}(Z) \neq 0$.

In order to solve (2.1), the customary approach consists in using an "*instrument matrix*" of functions of Z and θ, denoted $\mathcal{A}(Z,\theta) \in \mathbb{R}^{q \times d_\rho}$, with $q \geqslant d_\theta$ (e.g. Newey, 1993); such that (2.1) implies

$$E\big[\mathcal{A}(Z,\theta_0)\rho(\theta_0)\big] = 0_q. \tag{2.2}$$

The solution to (2.2) can be obtained by minimising a quadratic form based on the sample moments (e.g. Chamberlain, 1987; Robinson, 1987; Newey, 1990, 1993). However, a solution may be inconsistent or not unique (Domínguez and Lobato, 2004; Newey, 1993) because the solution to (2.2) may be different from θ_0. Indeed, (2.2) is necessary but may not be sufficient for (2.1). Therefore, methods of moments based on (2.2) requires an identification assumption which states that (2.2) is sufficient for (2.1) (e.g. Newey, 1993, Assumption 4.2). In other words, it is necessary to assume that (2.2) identifies globally the parameters and this assumption does not necessarily holds (Domínguez and Lobato, 2004). The approach proposed does not have this drawback, and does not rely on such identification assumption.

Equation (2.1) is equivalent to the continuum of moment conditions (see Bierens, 1982)

$$E\big[\rho(\theta)\exp(2\pi\mathrm{i}\eta'Z)\big] = 0 \quad \forall \eta \in \mathbb{R}^{d_Z} \quad \text{iff} \quad \theta = \theta_0, \tag{2.3}$$

where i denotes the imaginary unit. Thus,

$$M(\theta) := \int \Big\| E\big[\rho(\theta)\exp(2\pi\mathrm{i}\eta'Z)\big] \Big\|^2 \widehat{W}(\eta)d\eta = 0, \quad \text{iff} \quad \theta = \theta_0, \tag{2.4}$$

for some function $\widehat{W}(\cdot)$, which is such that $\widehat{W}(\eta) > 0$, $\forall \eta \in \mathbb{R}^{d_Z}$. Here, $\|\cdot\|$ denotes the Frobenius norm. We propose to use the Fourier transform; that is,

$$\widehat{W}(\eta) := \int W(z)\exp(-2\pi\mathrm{i}\eta'z)\,dz,$$

where $W(\bullet) : \mathbb{R}^{d_Z} \to \mathbb{R}$, is any symmetric function which have a strictly positive integrable Fourier transform. For example, this could be the Gaussian function (6.44). Bierens (1982) shows that (2.3) only needs to hold for frequencies η within the neighbourhood of zero. Indeed, the function $\widehat{W}(\cdot)$ gives more weight to frequencies close to zero, as in Bierens (1982, p.111).

Since (2.3) and (2.4) imply $M(\theta) \geqslant 0$ for all $\theta \in \Theta$, with $M(\theta_0) = 0$, we have that $M(\theta)$ identifies θ_0; in other words,

$$\theta_0 = \arg\min_{\theta \in \Theta} M(\theta). \tag{2.5}$$

The function $W(\bullet)$ is not used for local smoothing, and does not require a bandwidth selection. It simply ensures that (2.5) holds.

The function $M(\theta)$ can be re-formulated in a more convenient way. Let $(Y_i', Z_i')'$ and $(Y_j', Z_j')'$ $(i \neq j)$ be two independent copies of $(Y', Z')'$. Let $\rho_i(\theta) := \varrho(Y_i, \theta)$. By using

the inverse Fourier transform, equation (2.4) reduces to

$$\begin{aligned} M(\theta) &= \int E\Big[\rho_i(\theta)\exp(-2\pi\mathfrak{i}\eta' Z_i)\Big]' E\Big[\rho_j(\theta)\exp(2\pi\mathfrak{i}\eta' Z_j)\Big]\widehat{W}(\eta)d\eta \\ &= \int E\Big[\rho_i(\theta)'\rho_j(\theta)\exp\left(2\pi\mathfrak{i}\eta'(Z_j - Z_i)\right)\Big]\widehat{W}(\eta)d\eta \\ &= E\Big[\rho_i(\theta)'\rho_j(\theta)\int \widehat{W}(\eta)\exp\left(2\pi\mathfrak{i}\eta'(Z_j - Z_i)\right)d\eta\Big] \\ &= E\Big[\rho_i'(\theta)\rho_j(\theta)\,W(Z_i - Z_j)\Big]. \end{aligned} \tag{2.6}$$

Note that $W(Z_i - Z_j)$ are bounded almost everywhere, for $i \neq j$, because $\widehat{W}(\eta)$ is integrable. This prevents $M(\theta) \to 0$ for $\theta \neq \theta_0$.

The reformulation (2.6) can be found in Lavergne and Patilea (2013) and in Berger and Patilea (2020) for endogenous selection. Domínguez and Lobato (2004) proposed to minimise $\widetilde{M}(\theta) := \int \|E[\rho(\theta)I(Z \leqslant \eta)]\|^2 dP_Z(\eta)$ instead of (2.4), where P_Z denotes the probability distribution of Z. An alternative expression similar to (2.6), is $\widetilde{M}(\theta) = E[\rho_i'(\theta)\rho_j(\theta)\widetilde{W}(Z_i, Z_j)]$, where $\widetilde{W}(Z_i, Z_j) := E[I(Z_i \vee Z_j \leqslant Z) \mid Z_i, Z_j] = 1 - P_Z(Z_i \vee Z_j)$. The function $\widetilde{W}(\bullet, \bullet)$ does not belong to the class of functions $W(\bullet)$ considered here, because $\widetilde{W}(Z_i, Z_j)$ is not a function of $Z_i - Z_j$. Furthermore, $\widetilde{W}(Z_i, Z_j)$ has to be estimated, unlike $W(\bullet)$.

Under mild usual conditions which ensure differentiation under the integral sign and that the map $\theta \mapsto M(\theta)$ is convex in the neighbourhood of θ_0, (2.5) allows us to replace the optimisation problem (2.5) by the unconditional moment restriction,

$$\frac{\partial M(\theta)}{\partial \theta} = E\big[g_{ij}(\theta) + g_{ji}(\theta)\big] = 0_{d_\theta} \qquad \text{if and only if } \theta = \theta_0, \tag{2.7}$$

where

$$g_{ij}(\theta) := \frac{\partial \rho_i(\theta)}{\partial \theta}\rho_j(\theta)W_{ij} \in \mathbb{R}^{d_\theta}, \tag{2.8}$$

$$W_{ij} := \begin{cases} W(Z_i - Z_j) & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

Here, $\partial/\partial\theta$ stands for the $d_\theta \times 1$ vector of partial derivatives. We impose $W_{ij} = 0$ for $i = j$, because (2.6) is based upon two independent copies of $(Y', Z')'$. Since $E[g_{ij}(\theta)] = E[g_{ji}(\theta)]$, Equation (2.7) is equivalent to

$$\dot{M}(\theta) := E\big[g_{ji}(\theta)\big] = 0_{d_\theta}, \qquad \text{if and only if } \theta = \theta_0. \tag{2.9}$$

Equation (2.9) identifies globally the parameter of interest.

3. EMPIRICAL LIKELIHOOD ESTIMATOR

The estimator proposed is the solution to an empirical equivalent of $\dot{M}(\theta)$, based on empirical likelihood. Consider the "*empirical likelihood ratio function*" defined by

$$\mathcal{R}(\theta) := \max_{p_i : i=1,\ldots,n}\left(\prod_{i=1}^n np_i \;:\; p_i > 0,\ \sum_{i=1}^n p_i = 1,\ \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n p_i\, g_{ji}(\theta) = 0_{d_\theta}\right), \tag{3.10}$$

where $g_{ji}(\theta)$ is defined by (2.8). Within (3.10), it is important to use $p_i\, g_{ji}(\theta)$ and not $p_i\, g_{ij}(\theta)$, in order for Condition 4.6 to hold. The main difference between (3.10) and the

customary empirical likelihood approach (Owen, 1988; Qin and Lawless, 1994) is the double sum within the constraint, which creates some asymptotic hurdles to overcome. For example, the double sum at $\theta = \theta_0$ when $p_i = n^{-1}$ needs to be $O_p(n^{-\frac{1}{2}})$, which is one of the basic requirements for the asymptotic properties of empirical likelihood. This is straightforward with single sums; but not with double sums (see Section 4).

We assume that θ in (3.10) is such that 0_{d_θ} is an inner point of the convex hull formed by the $\{\widehat{h}_i(\theta) : i = 1, \ldots, n\}$, where

$$\widehat{h}_i(\theta) := \frac{1}{n}\sum_{j=1}^{n} g_{ji}(\theta)\cdot \tag{3.11}$$

The strict concavity of the objective function $\sum_{i=1}^{n}\log(np_i)$ implies that there exists a unique solution to constraint within (3.10). It can be shown that (e.g. Owen, 1988)

$$\mathcal{R}(\theta) = \prod_{i=1}^{n} np_i(\theta), \tag{3.12}$$

where

$$p_i(\theta) := n^{-1}\Big(1 + t'(\theta)\widehat{h}_i(\theta)\Big)^{-1}, \tag{3.13}$$

The Lagrange multiplier $t(\theta)$ satisfies the constraints within (3.10). It can be computed using a modified Newton-Raphson approach (e.g. Polyak, 1987).

The approach proposed has the advantage of reducing the dimensionality of the optimisation problem. Indeed, the dual optimisation (3.13) yield to a lower-dimensional Lagrange multiplier than other empirical likelihood-based methods. For example, Kitamura et al.'s (2004) approach involves n multipliers of dimension d_ρ. Donald et al.'s (2003) splines-based empirical likelihood function is based on kd_ρ multipliers, where k is the number of splines which increases with n. Here, the multiplier $t(\theta)$ is a d_θ-vector; that is, the number of components of $t(\theta)$ is given by the size of the parameter, which has the advantage of not increasing with n.

The '*maximum empirical likelihood point estimator*' of θ_0 is defined by

$$\widehat{\theta} := \arg\max_{\theta\in\Theta} \mathcal{R}(\theta)\cdot \tag{3.14}$$

It can be shown that $\widehat{\theta}$ is also the solution to

$$\dot{M}_n(\theta) = 0_{d_\theta}, \tag{3.15}$$

with

$$\dot{M}_n(\theta) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ji}(\theta) \tag{3.16}$$

is the empirical equivalent of (2.9). It can be used with non-linear models, when the GMM objective function does not provide consistent estimator or lead to several global minima. This estimator can be directly computed from (3.15), without invoking (3.10). The function (3.10) will be used to derive the test statistics (3.17).

Unlike Kitamura et al. (2004), we use the traditional empirical likelihood function. Hence, usual empirical likelihood packages can be used with the estimating function (3.11). With Kitamura et al.'s (2004) and Donald et al.'s (2003) approaches, the number

of constraints increases with the sample size, which makes them high dimensional. Our approach is based on a fixed number of constraints d_θ which is unrelated to n, d_ρ or d_Z, unlike Kitamura et al.'s (2004) and Donald et al.'s (2003) estimators.

The estimator proposed possessed the analogous property of a likelihood-type estimator, namely $\sqrt{n}$-consistency and asymptotic normality (Theorem 4.1). In Section 4, we also show that the empirical likelihood ratio function converges in distribution to $\chi^2_{d_\theta}$-distribution under the null (Theorem 4.2). Thus, $-2\log\mathcal{R}(\theta_0)$ is indeed an ancillary pivotal statistics, which can be used as an usual parametric likelihood ratio for testing hypotheses about θ_0.

It is often necessary to test a scalar component of θ_0 or a sub-vector of θ_0 (e.g. when we compare two nested models). This can be achieved by profiling $-2\log\mathcal{R}(\theta)$. Let $\theta^\dagger$ denote a sub-vector of θ. Let $\psi \in \mathbb{R}^{d_\psi}$ is the component of θ which are not part of $\theta^\dagger$; say $\theta = (\theta^{\dagger\prime}, \psi')'$. by using the Theorem 4.3 in Section 4, we have that the '*profile empirical likelihood ratio function*' is pivotal; that is,

$$-2\max_{\psi\in\Psi}\log\mathcal{R}(\theta_0^\dagger,\psi) \xrightarrow{d} \chi^2_{d_{\theta^\dagger}} \tag{3.17}$$

where Ψ denotes the parameter space of ψ (see (4.35)). Hence, the left hand side of (3.17) is a function of $\theta_0^\dagger$ which can be used to test $\theta_0^\dagger$. Property (3.17) can be used for model building by treating the profile empirical likelihood ratio function as a traditional log-likelihood ratio statistics.

4. MAIN RESULTS

In Section 4.1, we outline the regularity conditions needed for the asymptotic properties. Consistency and efficiency is discussed in Section 4.2. In Section 4.3, we have she asymptotic results of the empirical likelihood ratio statistics. Note that Qin and Lawless's (1994) asymptotic results do not directly apply, because the constraint within (3.10) is a double sum, rather than being a single sum, as in Qin and Lawless (1994).

4.1. Regularity conditions

In what follow, b_n denotes an arbitrary sequence such that $b_n \to 0$ and $nb_n^2 \to \infty$ and $\mathcal{B}_n := (\theta : \|\theta - \theta_0\| \leqslant b_n)$ is a ball around θ_0. Consider the following mild regularity conditions.

CONDITION 4.1. *We have $E[\|g_{ij}(\theta)\|^4] < \infty$, for all $\theta \in \mathcal{B}_n$, i and j.*

CONDITION 4.2. *(a) $\theta \mapsto \dot{\ddot{M}}_n(\theta)$ is continuous on Θ a.s. (b) $\|\partial\dot{\ddot{M}}_n(\theta)_k/\partial\theta\| = O_p(1)$ uniformly, where $\dot{\ddot{M}}_n(\theta)_k$ denotes the k-th row of $\dot{\ddot{M}}_n(\theta)$, where*

$$\dot{\ddot{M}}_n(\theta) := \frac{\partial\dot{M}_n(\theta)}{\partial\theta} \in \mathbb{R}^{d_\theta\times d_\theta} \tag{4.18}$$

and Θ denotes the compact parameter space of θ_0.

CONDITION 4.3. *There exists m_1, m_2 and n_0 such that for $n > n_0$, we have that $\mathbb{P}\big(0 < m_1 \leqslant \|\dot{\ddot{M}}_n(\theta_0)\| \leqslant m_2 < \infty\big) \to 1$.*

CONDITION 4.4. *There exists* λ^M *and* n_0 *such that for* $n > n_0$*, we have that* $\mathbb{P}\big(\lambda^M_{\min}(\theta_0) \geqslant \lambda^M > 0\big) \to 1$*, where* $\lambda^M_{\min}(\theta_0)$ *denotes the smallest eigenvalue of* $\ddot{M}_n(\theta_0)'\ddot{M}_n(\theta_0)$.

CONDITION 4.5. *There exists* $\lambda^{\Omega^\star}$ *and* n_0 *such that for* $n > n_0$*, we have that* $\mathbb{P}\big(\lambda^{\Omega^\star}_{\min}(\theta_0) \geqslant \lambda^{\Omega^\star} > 0\big) \to 1$*, where* $\lambda^{\Omega^\star}_{\min}(\theta_0)$ *denotes the smallest eigenvalue of*

$$\Omega^\star_n(\theta_0) := \frac{1}{n}\sum_{i=1}^n h^\star_i(\theta_0) \otimes h^\star_i(\theta_0), \tag{4.19}$$

where

$$h^\star_i(\theta_0) := E\big[g_{ji}(\theta_0)\big|Z_i, Y_i\big] \tag{4.20}$$

and $g_{ji}(\theta)$ *is defined by* (2.8)*. Here,* $\otimes$ *denotes the outer product.*

CONDITION 4.6. *There exists* λ^Ω *and* n_0 *such that for* $n > n_0$*, we have that* $\mathbb{P}\big(\inf_{\theta\in\mathcal{B}_n} \lambda^\Omega_{\min}(\theta) \geqslant \lambda^\Omega > 0\big) \to 1$*, where* $\lambda^\Omega_{\min}(\theta_0)$ *denotes the smallest eigenvalue of*

$$\Omega_n(\theta_0) := \frac{1}{n}\sum_{i=1}^n \widehat{h}_i(\theta_0) \otimes \widehat{h}_i(\theta_0) \in \mathbb{R}^{d_\theta \times d_\theta}. \tag{4.21}$$

For Condition 4.6 to hold, it is necessary to use $p_i\, g_{ji}(\theta)$ within (3.10). Indeed, Condition 4.6 would not hold if we use $p_i\, g_{ij}(\theta)$ instead of $p_i\, g_{ji}(\theta)$, because in this case, the quantities $\widehat{h}_i(\theta_0)$ would be based on $g_{ij}(\theta_0)$ and would converge to zero. Indeed,

$$\frac{1}{n}\sum_{j=1}^n g_{ij}(\theta_0) = \frac{\partial\rho_i(\theta_0)}{\partial\theta_0}\frac{1}{n}\sum_{j=1}^n \rho_j(\theta_0)W_{ij} \to \frac{\partial\rho_i(\theta_0)}{\partial\theta_0}E\big[\varrho(Y,\theta_0) \mid Z_i\big] = 0\cdot$$

and the smallest eigenvalue of $\Omega_n(\theta_0)$ would converge to zero. This can be easily avoided by using $p_i\, g_{ji}(\theta)$ within (3.10), and consequently $g_{ji}(\theta_0)$ within $\widehat{h}_i(\theta_0)$.

CONDITION 4.7. *The maximum eigenvalue of* $\Omega_n(\theta)$ *is finite* $\forall\theta \in \mathcal{B}_n$.

CONDITION 4.8. *There exists* $\lambda^{\bar\Omega^\star}$ *and* n_0 *such that for* $n > n_0$*, we have that* $\mathbb{P}\big(\lambda^{\bar\Omega^\star}_{\min}(\theta_0) \geqslant \lambda^{\bar\Omega^\star} > 0\big) \to 1$*, where* $\lambda^{\bar\Omega^\star}_{\min}(\theta_0)$ *denotes the smallest eigenvalue of*

$$\Omega^\star(\theta_0) := E\big[h^\star_i(\theta_0) \otimes h^\star_i(\theta_0)\big]\cdot \tag{4.22}$$

CONDITION 4.9. $\|E[g_{ji}(\theta_0) \otimes g_{ji}(\theta_0)]\| < \infty$*, where* $g_{ji}(\theta)$ *is defined by* (2.8).

Condition 4.2-4.4 are mild conditions on the Hessian, closely related to the one found in Qin and Lawless (1994). Conditions 4.4–4.8 are non-singularity conditions.

4.2. Consistency and semi-parametric efficiency

LEMMA 4.1. *Under Condition 4.1 and Conditions 4.3–4.9, we have that* $\|\widehat{\theta} - \theta_0\| = o_p(1)$

The proof of Lemma 4.1 can be found in Appendix A. This lemma is used to establish the following theorem.

THEOREM 4.1. *Under Condition 4.1–4.9, we have that*

$$n^{\frac{1}{2}}\big\|\widehat{\theta}-\theta_0\big\| = O_p(1), \tag{4.23}$$

$$n^{\frac{1}{2}}(\widehat{\theta}-\theta_0) \xrightarrow{d} N(0_{d_\theta},\Sigma), \tag{4.24}$$

where

$$\Sigma := \Big(\dot{\ddot{M}}'(\theta_0)\Omega^\star(\theta_0)^{-1}\dot{\ddot{M}}(\theta_0)\Big)^{-1},$$

$\Omega^\star(\theta_0)$ *is defined by* (4.22) *and* $\dot{\ddot{M}}(\theta_0)$ *is the second derivative of* $M(\theta)$ *defined by* (2.6).

The proof can be found in Appendix A.

Under conditions outlined in Newey (1993), semi-parametric efficiency can be obtained by adding an additional constraint to the empirical likelihood function; that is, consider an "*adjusted empirical likelihood ratio function*" given by

$$\mathcal{R}^\circ(\theta,\theta^\circ) := \max_{p_i:i=1,\ldots,n}\Big(\prod_{i=1}^n np_i \;:\; np_i>\frac{1}{2},\ \sum_{i=1}^n p_i=1,\ \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n p_i\, g_{ji}(\theta)=0_{d_\theta},$$
$$\sum_{i=1}^n p_i\,\widehat{\mathcal{A}}_i^\circ(\theta)\Big(\frac{\partial\rho_i'(\theta)}{\partial\theta}(\theta^\circ-\theta)+\rho_i(\theta)\Big)=0_{d_\theta}\Big), \tag{4.25}$$

where $\widehat{\mathcal{A}}_i^\circ(\theta)$ is an estimator of the optimal instrument (e.g. Newey, 1993) given by

$$\mathcal{A}^\circ(Z_i,\theta) := E\Big[\frac{\partial\rho_i(\theta)}{\partial\theta}\Big|Z_i\Big]E\big[\rho_i(\theta)\rho_i'(\theta)|Z_i\big]^{-1}. \tag{4.26}$$

Since the estimator $\widehat{\mathcal{A}}_i^\circ(\theta)$ involves estimating conditional expectations, regressions or nearest neighbour estimator needs to be used (Newey, 1990, 1993; Robinson, 1991). For example, we can use the Nadaraya (1964) and Watson's (1964) estimator described in Section S2.1 of the Supplement.

Let $(\widehat{\theta}',\widehat{\theta}^{\circ\prime})' := \arg\max_{\theta,\theta^\circ\in\Theta}\mathcal{R}^\circ(\theta,\theta^\circ)$. Thus, the maximum empirical likelihood estimator is

$$\widehat{\theta}^\circ := \big(0_{d_\theta\times d_\theta}, I_{d_\theta\times d_\theta}\big)\arg\max_{\theta,\theta^\circ\in\Theta}\mathcal{R}^\circ(\theta,\theta^\circ), \tag{4.27}$$

where $0_{d_\theta\times d_\theta}$ denotes the $d_\theta\times d_\theta$ zero matrix and $I_{d_\theta\times d_\theta}$ is the $d_\theta\times d_\theta$ identity matrix. The estimator $\widehat{\theta}^\circ$ is the solution to

$$\frac{1}{n}\sum_{i=1}^n\widehat{\mathcal{A}}_i^\circ(\widehat{\theta})\Big(\frac{\partial\rho_i'(\widehat{\theta})}{\partial\widehat{\theta}}(\widehat{\theta}^\circ-\widehat{\theta})+\rho_i(\widehat{\theta})\Big)=0_{d_\theta}, \tag{4.28}$$

where $\widehat{\theta}$ defined by the first constraint within (4.25); that is, $\widehat{\theta}$ is the solution to (3.16).

Equation (4.28) is one Newton-Raphson iteration towards the solution of $n^{-1}\sum_{i=1}^n\widehat{\mathcal{A}}_i^\circ(\theta)\rho_i(\theta)=0_q$, as in Newey (1993). The estimator $\widehat{\theta}^\circ$ is based on an initial $\sqrt{n}$-consistent estimator $\widehat{\theta}$, rather than an initial arbitrary estimator that may be inconsistent, as in Newey (1993). The $\sqrt{n}$-consistency of $\widehat{\theta}$ implies that this iterative step gives an efficient estimator. This relies on additional regularity conditions which can be found in Newey (1993). The key condition is that the unconditional restrictions (2.2) with $\mathcal{A}(Z,\theta_0)=\mathcal{A}^\circ(Z,\theta)$ identifies globally θ_0, which may not be true, as pointed out by Domínguez and Lobato (2004) (see Section 2). It also relies on a well-behaved estimator

$\widehat{\mathcal{A}}_i^\circ(\theta)$ of (4.26), which may be difficult to derive. Examples can be found in Section S3.2 of the Supplement.

The dual optimisation induced by (4.25) also yield to a low dimensional Lagrange multiplier of dimension $2d_\theta$. However, local smoothing is needed to estimate (4.26), for example, when the Nadaraya (1964) and Watson's (1964) estimator is used. This is disadvantage over (3.10) which is free of smoothing parameters. Indeed, smoothing parameters are often required to reach the semiparametric efficiency.

In Section S3 of the supplement, a simulation study reveals that $\widehat{\theta}^\circ$ can out-perform $\widehat{\theta}$ in some cases, but can also be less efficient than $\widehat{\theta}$ in others. Thus, $\widehat{\theta}^\circ$ should be used cautiously, and only when $\widehat{\mathcal{A}}_i^\circ(\theta)$ is a good estimator. The difference $(\widehat{\theta}^\circ - \widehat{\theta})$ should be small and could be used as a diagnostic. For example, a estimate $\widehat{\theta}^\circ$ is not reliable, when $\widehat{\theta}^\circ$ is not within an $1-\alpha$ confidence region; that is, when $-2\log\mathcal{R}(\widehat{\theta}^\circ)$ is larger than the $(1-\alpha)$-quantile of a $\chi^2_{d_\theta}$-distribution.

4.3. Empirical likelihood ratio and testing

LEMMA 4.2. *Under Condition 4.1 and Conditions 4.5 and 4.6, and for any random matrix $\widehat{A}$ such that $\|\widehat{A}\| < \infty$, we have that*

$$\dot{M}_n^{\star\prime}(\theta_0)\,\widehat{A}\,\Omega_n(\theta_0)^{-1}\dot{M}_n^\star(\theta_0) = \dot{M}_n^{\star\prime}(\theta_0)\,\widehat{A}\,\Omega_n^\star(\theta_0)^{-1}\dot{M}_n^\star(\theta_0) + O_p(n^{-\frac{3}{2}}), \tag{4.29}$$

where $\Omega_n^\star(\theta)$ is defined by (4.19) *and*

$$\dot{M}_n^\star(\theta) := \frac{1}{n}\sum_{i=1}^n h_i^\star(\theta)\cdot \tag{4.30}$$

Here, $h_i^\star(\theta)$ is defined by (4.20).

LEMMA 4.3. *Under Condition 4.1, we have $t(\theta_0) = O_p(n^{-\frac{1}{2}})$ and $t(\theta_0) = \Omega_n(\theta_0)^{-1}\dot{M}_n^\star(\theta_0) + O_p(n^{-1})$.*

The proof of Lemmas 4.2 and 4.3 can be found in the supplement. Note that $\dot{M}_n^\star(\theta_0)$ is a single sum, which is a key feature to derive the asymptotic distribution, based on the approximation of $t(\theta_0)$ in Lemma 4.3 (see the proof of Theorem 4.2 and equation (4.35)).

THEOREM 4.2. *Under Condition 4.1–4.9, we have that*

$$-2\log\mathcal{R}(\theta_0) \xrightarrow{d} \chi^2_{d_\theta}, \tag{4.31}$$

where $\chi^2_{d_\theta}$ denotes the χ^2-distribution with d_θ degrees of freedom.

Proof of Theorem 4.2 This proof is based on Lemmas S2 and S3, which can be found in the supplement. Lemma S3 implies that the conditions of Lemma S2 holds with b_n and θ respectively replaced by $n^{-\frac{1}{2}}$ and θ_0. Thus, by using Lemma S2, we obtain

$$-2\log\mathcal{R}(\theta_0) = n\,t'(\theta_0)\Omega_n(\theta_0)t(\theta_0) + O_p(n^{-\frac{1}{2}})\cdot \tag{4.32}$$

By substituting into (4.32), the expression of $t(\theta_0)$ within Lemma 4.3, we obtain

$$-2\log\mathcal{R}(\theta_0) = n\Big(\dot{M}_n^{\star\prime}(\theta_0)\Omega_n(\theta_0)^{-1}\dot{M}_n^{\star}(\theta_0) + 2\dot{M}_n^{\star\prime}(\theta_0)O_p(n^{-1}) + O_p(n^{-1})'\Omega_n(\theta_0)O_p(n^{-1})\Big) + O_p(n^{-\frac{1}{2}})\cdot \tag{4.33}$$

Now, Lemma S3 implies $\dot{M}_n^{\star\prime}(\theta_0)O_p(n^{-1}) = O_p(n^{-\frac{3}{2}})$. Since the minimum eigenvalue of $\Omega_n(\theta_0)$ is bounded away from zero, we have that $O_p(n^{-1})'\Omega_n(\theta_0)O_p(n^{-1}) = O_p(n^{-2})$. Thus, (4.33) reduces to

$$-2\log\mathcal{R}(\theta_0) = n\,\dot{M}_n^{\star\prime}(\theta_0)\Omega_n(\theta_0)^{-1}\dot{M}_n^{\star}(\theta_0) + O_p(n^{-\frac{1}{2}})\cdot$$

Now, by using Lemma 4.2 with $\widehat{A}$ being the identity matrix, we obtain

$$-2\log\mathcal{R}(\theta_0) = n\,\dot{M}_n^{\star\prime}(\theta_0)\Omega_n^{\star}(\theta_0)^{-1}\dot{M}_n^{\star}(\theta_0) + O_p(n^{-\frac{1}{2}})\cdot \tag{4.34}$$

The key feature of (4.34) is that the right hand side is a quadratic form with $\dot{M}_n^{\star}(\theta)$ being a single sum, despite that (3.16) is an double sum. Since $E[\dot{M}_n^{\star}(\theta_0)] = 0$ and $n^{-1}E[\Omega_n^{\star}(\theta_0)] = V[\dot{M}_n^{\star}(\theta_0)]$, standard central limit theorem implies that $n^{\frac{1}{2}}\Omega_n^{\star}(\theta_0)^{-\frac{1}{2}}\dot{M}_n^{\star}(\theta_0)$ converges (in distribution) to a standard multivariate normal distribution. Hence, (4.34) converges in distribution to $\chi^2_{d_\theta}$-distribution and (4.31) follows. □

THEOREM 4.3. *Let $\widetilde{\theta}_0 = (\theta_0^{\dagger\prime}, \psi_M')'$, where $\psi_M := \arg\max_{\psi\in\Psi}\log\mathcal{R}(\theta_0^{\dagger}, \psi)$ and $\theta_0^{\dagger}$ denotes a sub-vector of θ_0 and $\psi \in \mathbb{R}^{d_\psi}$ is the remaining sub-vector. Under Conditions 4.1–4.9, we have that*

$$-2\log\mathcal{R}(\widetilde{\theta}_0) = n\,\dot{M}_n^{\star\prime}(\theta_0)(I_{d_\theta\times d_\theta} - A_0)\Omega_n^{\star}(\theta_0)^{-1}\dot{M}_n^{\star}(\theta_0) + O_p(n^{-\frac{1}{2}}) \xrightarrow{d} \chi^2_{d_{\theta^\dagger}}, \tag{4.35}$$

where $I_{d_\theta\times d_\theta}$ denotes the $d_\theta \times d_\theta$ identity matrix and

$$A_0 := \Omega_n(\theta_0)^{-1}\nabla_n\Big(\nabla_n\Omega_n(\theta_0)^{-1}\nabla_n'\Big)^{-1}\nabla_n',$$
$$\nabla_n := \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\widehat{h}_i(\theta)}{\partial\psi}\Big|_{\theta^\dagger=\theta_0^\dagger,\psi=\psi_0}. \tag{4.36}$$

The proof of Theorem 4.3 can be found in Appendix A.

With (4.25), testing can be based on $-2\max_{\theta\in\Theta}\log\mathcal{R}^\circ(\theta,\theta^\circ)$ which converges in distribution to χ^2-distribution with $d_{\theta^\dagger}$ degree of freedom, when $\theta^\circ = \theta_0$. The proof is analogous to the proof of Theorem 4.3, and involves regularity conditions on $\widehat{\mathcal{A}}_i^\circ(\theta)$, as in Newey (1993).

4.4. Local power and test consistency

We establish an asymptotic expression for the local power. We also show that the empirical likelihood ratio test is consistent.

LEMMA 4.4. *Let $\widetilde{\theta} = \theta_0 + Lb_n$, for some $\|L\| < \infty$, where b_n denotes an arbitrary*

sequence such that $nb_n^2 \to \infty$ *and* $b_n \to 0$. *If* $(n \min_{k=1,\dots,n} p_k(\widetilde{\theta}))^{-1} = O_p(1)$, *we have that*

$$-2\log \mathcal{R}(\widetilde{\theta}) = n\dot{M}_n(\widetilde{\theta})'\Omega_n(\widetilde{\theta})^{-1}\dot{M}_n(\widetilde{\theta}) + O_p(nb_n^3)\cdot \tag{4.37}$$

The proof of Lemma 4.4 can be found in the supplement. Corollary 4.1 shows that the empirical likelihood ratio test of the hypothesis $H_0 : \theta = \widetilde{\theta}$ is consistent against the alternative $H_A : \theta \neq \widetilde{\theta}$, because the power tends to 1, as $n \to \infty$.

COROLLARY 4.1. *With* $\widetilde{\theta}$ *defined as in Lemma 4.4, we have that there exists a sequence* $r_n \to \infty$, *such that* $\mathbb{P}(-2\log \mathcal{R}(\widetilde{\theta}) \geqslant r_n) \to 1$, *as* $n \to \infty$.

The proof of Corollary 4.1 can be found in the supplement.

By using Lemma 4.4, we can derive the local power of the test $H_0 : \theta = \breve{\theta}$, with $\breve{\theta} = \theta_0 + Ln^{-\frac{1}{2}}$ against the alternative hypothesis based on the correct value of the parameter. By substituting b_n by $n^{-\frac{1}{2}}$, Lemma 4.4 implies

$$-2\log \mathcal{R}(\breve{\theta}) = n\dot{M}_n(\breve{\theta})'\Omega_n(\breve{\theta})^{-1}\dot{M}_n(\breve{\theta}) + O_p(n^{-\frac{1}{2}})\cdot \tag{4.38}$$

Since $\Omega_n(\breve{\theta})^{-1} - E\big[\Omega_n(\breve{\theta})\big]^{-1} = \Omega_n(\breve{\theta})^{-1}\big[E\big[\Omega_n(\breve{\theta})\big] - \Omega_n(\breve{\theta})\big]E\big[\Omega_n(\breve{\theta})\big]^{-1}$, we have that

$$\begin{aligned}\Big\|\dot{M}_n(\breve{\theta})'\big[\Omega_n(\breve{\theta})^{-1} - E\big[\Omega_n(\breve{\theta})\big]^{-1}\big]\dot{M}_n(\breve{\theta})\Big\| \leqslant \big\|\Omega_n(\breve{\theta})^{-1}\big\|\big\|E\big[\Omega_n(\breve{\theta})\big]^{-1}\big\| \\ \Big\|E\big[\Omega_n(\breve{\theta})\big] - \Omega_n(\breve{\theta})\Big\|\big\|\dot{M}_n(\breve{\theta})\big\|\cdot\end{aligned} \tag{4.39}$$

Note that (S.7) implies $\big\|\dot{M}_n(\breve{\theta})\big\| = O_p(n^{-\frac{1}{2}})$. Thus, under $\Omega_n(\breve{\theta}) - E\big[\Omega_n(\breve{\theta})\big] = O_p(n^{-\frac{1}{2}})$, we have that the right hand side of (4.39) is $O_p(n^{-\frac{3}{2}})$. Hence, (4.38) implies

$$-2\log \mathcal{R}(\breve{\theta}) = n\dot{M}_n(\breve{\theta})'E\big[\Omega_n(\breve{\theta})\big]^{-1}\dot{M}_n(\breve{\theta}) + O_p(n^{-\frac{1}{2}})\cdot$$

Since, $n^{-1}E\big[\Omega_n(\breve{\theta})\big] \doteq V[\dot{M}_n(\breve{\theta})]$, we can conjecture that $n^{\frac{1}{2}}E\big[\Omega_n(\breve{\theta})\big]^{-\frac{1}{2}}\dot{M}_n(\breve{\theta})$ converges (in distribution) to a multivariate normal distribution with a covariance matrix $I_{d_\theta \times d_\theta}$, we have that $-2\log \mathcal{R}(\breve{\theta})$ has a limit χ^2-distribution with d_θ degree of freedom and a non-centrality parameter $\tau(\breve{\theta}) := 4n\mu(\breve{\theta})'\mu(\breve{\theta})$, where $\mu(\breve{\theta}) := E\big[\Omega_n(\breve{\theta})\big]^{-\frac{1}{2}}E\big[\dot{M}_n(\breve{\theta})\big]$. Taylor's theorem and $E[\dot{M}_n(\theta_0)]$ imply $E[\dot{M}_n(\breve{\theta})] = E[\ddot{M}_n(\theta_0)]Ln^{-\frac{1}{2}} + O(n^{-1})$; where $\ddot{M}_n(\theta)$ is defined by (4.18). Thus, the non-centrality parameter can be approximated by

$$\tau(\breve{\theta}) \doteq L'S_n(\theta_0)L,$$

where

$$S_n(\theta_0) := E[\ddot{M}_n(\theta_0)]'E\big[\Omega_n(\breve{\theta})\big]^{-1}E[\ddot{M}_n(\theta_0)]\cdot$$

Finally, an asymptotic approximation for the local power is

$$\beta(\breve{\theta}) \doteq 1 - F\big(\chi^2_{d_\theta;1-\alpha}, d_\theta, \tau(\breve{\theta})\big), \tag{4.40}$$

where $F(\cdot, d_\theta, \tau(\breve{\theta}))$ is the distribution function of a non-central χ^2-distribution with d_θ degree of freedom and a non-centrality parameter $\tau(\breve{\theta})$. Here, $\chi^2_{d_\theta;1-\alpha}$ is the $(1-\alpha)$-quantile of a central $\chi^2_{d_\theta}$-distribution. Since $\tau(\breve{\theta}) \geqslant 0$, we have that $\beta(\breve{\theta})$ is indeed larger than a type I error α. The asymptotic power increases with $\tau(\breve{\theta})$, as expected. The matrix

$S_n(\theta_0)$ characterises the curvature of $-2\log\mathcal{R}(\breve{\theta})$. A large curvature increases the values of $\tau(\breve{\theta})$ which implies a greater power and smaller confidence region. For power analysis, (4.40) can be estimated by replacing the matrix $S_n(\theta_0)$ by $\dot{M}_n^{'}(\widehat{\theta})\Omega_n(\widehat{\theta}+Ln^{-\frac{1}{2}})^{-1}\dot{M}_n(\widehat{\theta})$. Simulation results related to the estimation of (4.40) can be found in Section S2 of the supplement.

5. AN EMPIRICAL LIKELIHOOD MODEL SPECIFICATION TEST

Suppose we wish to test

$$H_0 : \exists\,\theta\in\Theta\ :\ E\big[\varrho(Y,\theta)\mid Z\big]=0_{d_\rho},\ a.s.$$

against $H_A : \mathbb{P}[E[\varrho(Y,\theta)\mid Z]=0_{d_\rho}]<1,\ \forall\theta\in\Theta$. By using (2.3), we see that H_0 holds if $\exists\,\theta\in\Theta$ such that $E[\rho(\theta)\exp(2\pi\imath\eta'Z)]=0,\ \forall\eta\in\mathbb{R}^{d_Z}$, or equivalently if

$$\exists\,\theta\in\Theta:\quad E\big[\rho(\theta)\otimes\digamma(2\pi\eta'Z)\big]=0_{d_\rho\times 2},$$

$\forall\eta\in\mathbb{R}^{d_Z}$, where

$$\digamma(x):=\big(\cos(x),\sin(x)\big)'\in\mathbb{R}^2.$$

Consider

$$\widetilde{\mathcal{R}}(\theta,\eta):=\max_{p_i:i=1,\ldots,n}\Bigg(\prod_{i=1}^n np_i\ :\ p_i>0,\ \sum_{i=1}^n p_i=1,\ \sum_{i=1}^n p_i\,\rho_i(\theta)\otimes\digamma(2\pi\eta'Z_i)=0_{d_\rho\times 2}\Bigg).$$

Following Qin and Lawless (1994), it can be shown that under H_0 and for a given η

$$-2\log\widetilde{\mathcal{R}}(\theta_0,\eta)\xrightarrow{d}\chi^2_{2d_\rho}.$$

Since $\widehat{\theta}$ is a $\sqrt{n}$ -consistent estimator of θ_0, we have that $\log\widetilde{\mathcal{R}}(\widehat{\theta},\eta)$ and $\log\widetilde{\mathcal{R}}(\theta_0,\eta)$ have the same asymptotic distribution under H_0 (see Theorem S1 in the supplement); that is,

$$-2\log\widetilde{\mathcal{R}}(\widehat{\theta},\eta)\xrightarrow{d}\chi^2_{2d_\rho},\quad\forall\eta. \tag{5.41}$$

We propose to use the following test statistics.

$$\mathcal{S}(\widehat{\theta}):=\max_{\eta\in\Gamma_\varepsilon}\Big(-2\log\widetilde{\mathcal{R}}(\widehat{\theta},\eta)\Big), \tag{5.42}$$

where Γ_ε denotes a finite set of d_Γ vectors randomly chosen within a d_Z-ball of dimension ε centred at zero. By using (5.41), the asymptotic distribution of $\mathcal{S}(\widehat{\theta})$ under H_0, is given by the distribution of the variable $\mathcal{X}^2_{\max}:=\max(\mathcal{X}^2_i:\ i=1,\ldots,d_\Gamma)$, where $\mathcal{X}^2_1$, …, $\mathcal{X}^2_i$, …, $\mathcal{X}^2_{\dim(\Gamma_\varepsilon)}$ denote d_Γ independent χ^2-distributed random variables with $2d_\rho$ degrees of freedom. Thus, $\mathcal{X}^2_{\max}$ follows a Gumbel distribution, as $d_\Gamma\to\infty$ (Embrechts et al., 1997, Ch.3), and the asymptotic p-value is

$$\text{p-value}=1-\exp\Big(-\exp\Big(\frac{1}{2}\mathcal{S}(\widehat{\theta})-\log(d_\Gamma)+(1-d_\rho)\log\log(d_\Gamma)+\log\Gamma(d_\rho)\Big)\Big). \tag{5.43}$$

We reject H_0 if the p-values is less than a nominal size α. On the other hand, under H_A, the quantity $n^{-1}\sum_{i=1}^n\rho_i(\widehat{\theta})\otimes\digamma(2\pi\eta'Z_i)$ would be significantly different from zero for some η, leading to large values for $\mathcal{S}(\widehat{\theta})$, which increases the probability of rejecting H_0.

The test depends on the arbitrarily chosen constants ε and d_Γ, specifying the size of the d_Z-ball and the number of vectors within it. The p-value is automatically adjusted for d_Γ. Finding a suitable decision rule to determine ε and d_Γ is beyond the scope of

this paper. The simulation study in Section 6 and in Section S2.6 of the supplement, shows that the proposed specification test has the right size and acceptable power, with large sample sizes. The effect of ε and d_Γ on the size and the power is minimal when the sample size is large. For small sample size, having small values for ε and d_Γ seems more suitable.

6. SIMULATION STUDY

In Section S2 of the supplement, we have the detailed simulation study, additional simulation results of linear models with endogenous covariates, as well as simulation results related to Kitamura et al.'s (2004) smoothed empirical likelihood (SEL) approach.

For the simulation studies, we shall use the Gaussian function for $W(\bullet)$; that is,

$$W(z) := \exp\bigl(-\mu \|z\|^2\bigr), \tag{6.44}$$

where μ is strictly positive.

Consider Domínguez and Lobato's (2004) non-linear model, given by

$$\mathcal{Y} = \theta_0^2 Z + \theta_0 Z^2 + u, \tag{6.45}$$

with $\theta_0 = 1.25$, $Z \sim \mathcal{N}(\mu_Z, 1)$ independent of $u \sim \mathcal{N}(0,1)$. Here, $\mu_Z = 0$ or 1. When $\mu_Z = 1$, the optimal instrument cannot identify θ_0, but it does with $\mu_Z = 0$ (Domínguez and Lobato, 2004). Here, $\mu = 0.5$. The sample size is $n = 100$. We use 1000 replicates. In Table 1, we have the observed local power based on the EL test based on (4.31) (column EL) and on Domínguez and Lobato's (2004) Wald statistics (columns DL). The observed sizes are close to 5%, but the size of DL with $\mu_Z = 0$ is about 7%. We clearly see that the EL test is more powerful.

Table 1. Observed local power (%). $H_0 : \theta = \theta_0 + Ln^{-\frac{1}{2}}$; where $L = \ell\, 1_{d_\theta}$. Nominal level = 5%. $n = 100$. 1000 replicates.

	$\mu_Z = 0$		$\mu_Z = 1$	
ℓ	EL	DL	EL	DL
-1.0	60.6	32.2	99.7	80.7
-0.5	21.1	14.7	72.5	33.2
0.0	6.2	7.1	5.8	4.9
0.5	22.7	11.8	74.9	32.6
1.0	67.2	31.0	99.7	83.8

In Section S3.2 of the supplement, we compare the root mean-squared errors for different models. For example, under (6.45), the proposed estimators (3.14) and (4.27) are more efficient than SEL, SBEL and DL, with DL being the less efficient.

Consider Bierens's (1982) example to illustrate the model specification test proposed in Section 5; namely,

$$\mathcal{Y} = Z^{(1)} + Z^{(2)} + Z^{(1)}Z^{(2)} + \epsilon_1, \tag{6.46}$$

$$\mathcal{Y} = Z^{(1)} + Z^{(2)} + \epsilon_2; \tag{6.47}$$

where $Z^{(1)} := (Z^2 - 1)/\sqrt{2}$, $Z \sim \mathcal{N}(0,1)$, $Z^{(2)} \sim \mathcal{N}(1,1)$, $\epsilon_1 \sim \mathcal{N}(0, \sigma_\epsilon^2 = 0.04)$ and

$\epsilon_2 \sim \mathcal{N}(0, \sigma^2_\epsilon = 2.04)$. We wish to test

$$H_0 \colon \exists\, \theta = (\alpha, \beta^{(1)}, \beta^{(2)})' \in \Theta \colon E\big[\mathcal{Y} - \alpha - \beta^{(1)} Z^{(1)} - \beta^{(2)} Z^{(2)} \mid (Z^{(1)}, Z^{(2)})'\big] = 0_{d_\rho},\ a.s.$$

Under model (6.46), H_0 is false and under (6.47), H_0 is true. The number of vectors within Γ_ε is given by $d_\Gamma := \lceil \pi \varepsilon^2 \rceil$, where ε is the radius of a 2-ball. We shall consider d_Z-balls of sizes $\varepsilon = 5$, 10, 20 and 30.

In Table 2, we report observed rejection rates of H_0, under models (6.46) and (6.47), with a nominal size 0.05. Under model (6.46), H_0 is false and we observe large rates, especially with large value of ε. With the model (6.47), the rates are close to the nominal level, with $n = 500$ and . Larger rates are observed with $n = 100$. A small value for ε is enough to achieve a decent size. Nevertheless, large values of ε increases the power, for small sample sizes.

Table 2. Rejection rates (%) of H_0, under (6.46) and (6.47). Nominal size = 0.05. 1000 replicates.

	Model (6.46): H_0 false.				Model (6.47): H_0 true.			
n	$\varepsilon = 5$	$\varepsilon = 10$	$\varepsilon = 20$	$\varepsilon = 30$	$\varepsilon = 5$	$\varepsilon = 10$	$\varepsilon = 20$	$\varepsilon = 30$
100	51.6	56.9	65.0	73.3	8.7	10.5	14.5	18.2
200	49.4	49.1	55.4	55.5	7.0	6.7	6.3	8.4
500	55.5	50.5	46.8	49.5	7.8	5.0	8.1	4.6

7. EMPIRICAL EXAMPLE

The 1998–99 UK Family Expenditure Survey (FES) is a random sample stratified by region which contains $n = 6630$ private households drawn from the Post Office's list of addresses (e.g. Goodman and Webb, 1994). We use the FES data to fit a basic income-determination model between total household expenditure ($\mathcal{Y}$) and income ($\mathcal{X}$).

$$\mathcal{Y} = \alpha_0 + \beta_0 \mathcal{X} + \epsilon, \quad \text{with } E[\epsilon \mid Z] = 0 \text{ and } Z := \mathcal{X} - \mathcal{Y}.$$

The income $\mathcal{X}$ is potentially endogenous and Z is treated as an instrumental variable. To protect the confidentiality, the scale of the variables was changed and the name and size of the regions are not revealed.

First, we test if Z is instrumental; that is, we test

$$H_0^Z : \exists\, \alpha,\, \beta :\ E\big[\mathcal{Y} - \alpha - \beta \mathcal{X} \mid Z\big] = 0_{d_\rho},\ a.s. \tag{7.48}$$

In Table 3, we report of the p-values (5.43) of the 12 regions, for testing (7.48), with the approach of Section 5. Here, $\varepsilon = 30$ and $d_Z = 3000$. The regions haven been sorted according to their p-value and re-labelled. Coincidentally, we do not reject H_0^Z at 5% level, for 6 out of the 12 regions.

In Figure 1, we have the estimates (3.14) and those of Domínguez and Lobato's (2004) (DL) and 2SLS, for the first 6 regions, with p-values less than 5% in Table 3. The vertical bars represents the 95% confidence intervals of EL, DL and on 2SLS. We also report the values of (4.27). The quantity μ within (6.44) is given by $\mu = (2\,\widehat{\sigma}_z^2)^{-1}$, where $\widehat{\sigma}_z$ is the observed standard deviation of the variable Z.

The OLS estimates are biased, because of the endogeneity of $\mathcal{X}$. The confidence interval tends of EL and DL overlap, but DL tends to give wider confidence intervals, because

Table 3. p-values (5.43) for testing (7.48). $\mu = (2\,\widehat{\sigma}_z^2)^{-1}$.

Region	p-value	Region	p-value	Region	p-value	Region	p-value
1	0.40	4	0.13	7	0.04	10	0.01
2	0.30	5	0.11	8	0.04	11	0.00
3	0.26	6	0.08	9	0.03	12	0.00

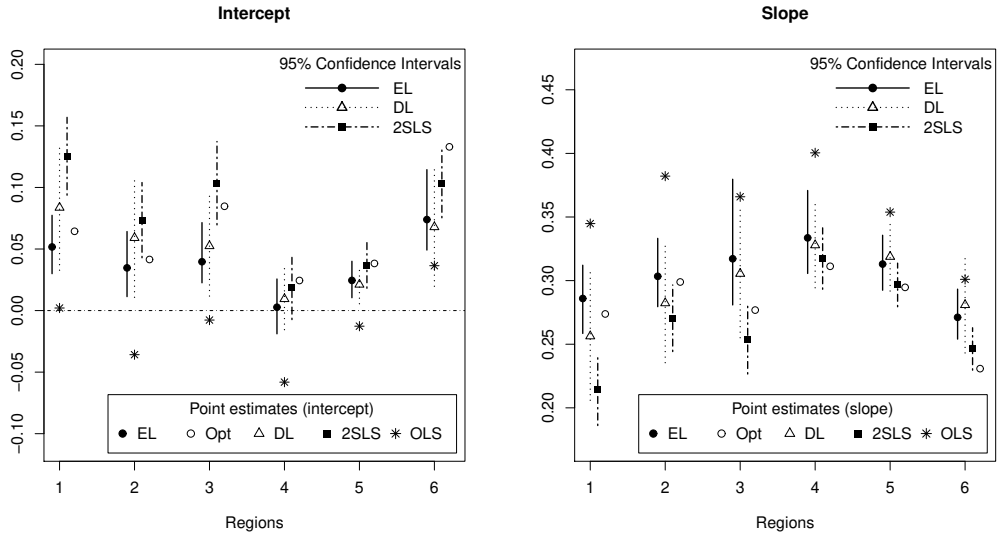

Figure 1. Point estimates (intercept and slope) and 95% confidence intervals. 1998–99 UK Family Expenditure Survey data. EL: proposed empirical likelihood (3.14). Opt: estimate (4.27). DL: Domínguez and Lobato (2004). 2SLS: two-stage least squares. OLS: ordinary least squares. $\mu = (2\,\widehat{\sigma}_z^2)^{-1}$.

the DL point estimator can be less accurate than EL. We notice significant difference between the proposed empirical likelihood (EL) approach and 2SLS. Confidence intervals may or may not overlap. The EL intervals can be asymmetric due to the skewness of the data. We mostly obtain smaller intercepts and larger slopes for EL compared to 2SLS. The EL estimates tends to be between the OLS and 2SLS estimates. The estimator (4.27) (Opt) based on the optimal instrument, is usually close to EL, but drifts slightly in the direction of the 2SLS estimates. Since $\mathcal{X}$ and Z are skewed, the relationship between $\mathcal{X}$ and Z is heteroscedastic and/or may not be linear. As a result a bias can be introduced within the fitted values of the first stage of 2SLS. This tends to underestimate the slope and overestimate the intercept, as observed in Figure 1.

8. CONCLUDING REMARKS

The empirical likelihood estimator proposed has several advantages over its competitors, namely Domínguez and Lobato's (2004), Donald et al.'s (2003) and Kitamura et al.'s (2004) approaches. The point estimator proposed is a solution to a finite number of estimating equations, and can be simply computed with usual empirical likelihood packages. Donald et al.'s (2003) and Kitamura et al.'s (2004) approaches are based on high dimen-

sional constraints, which may creates some instability in the estimates. The simulation studies in Section S3 of the supplement, show that the estimator proposed is as accurate as its competitors. It may even be more precise with over-identified moment restrictions. We also proposed an empirical likelihood test for model specification. The simulation study in the supplement shows that this test has an acceptable size.

We showed that the self-normalising property holds. We provide an expression for the local power and show that the proposed test is asymptotically consistent. The simulation studies shows that the test proposed has usually the right size and has acceptable power. Domínguez and Lobato's (2004) Wald test may be less powerful, and its point estimator can be less accurate. The empirical likelihood ratio function test statistics proposed by Donald et al.'s (2003) and Kitamura et al.'s (2004) may not have the right sizes.

In Section 4.2, we also propose an efficient estimator based on an empirical likelihood ratio function adjusted with an additional constraint. This leads to an efficient estimator, under some specific conditions. This adjusted estimator should be used with caution, because our simulation study shows that it can be inefficient, when these conditions are not met.

ACKNOWLEDGEMENTS

The author is grateful to Professor Valentin Patilea (ENSAI, France) for suggesting suitable examples and regularity conditions, and to Professor Peter Phillips (Yale University) for advising to add Section 5. The author wish to thanks Elodie Maignan (Polytech Clermont-Ferrand, France) for preliminary numerical investigation of Kitamura et al.'s (2004) approach.

REFERENCES

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Amemiya, T. (1977). The maximum likelihood and the nonlinear three stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica 45*, 955–968.

Berger, Y. G. and V. Patilea (2020). A semi-parametric empirical likelihood approach for conditional estimating equations under endogenous selection. *Submitted manuscript*.

Bierens, H. (1982). Consistent model specification tests. *Journal of Econometrics 20*(1), 105–134.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*, 305–334.

Chang, J., S. X. Chen, and X. Chen (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics 185*, 283–304.

Chen, S. and I. Van Keilegom (2009). A review on empirical likelihood methods for regression. *Test 18*, 415–447.

Chen, X. and D. Pouzo (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics 152*, 46–60.

Domínguez, M. A. and I. N. Lobato (2004). Consistent estimation of models defind by conditional moment restrictions. *Econometrika 72*(5), 1601–1615.

Donald, S., G. Imbens, and W. Newey (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics 117*, 55–93.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modelling Extremal Events for Insurance and Finance*. Berlin Heidelberg: Springer-Verlag.

Goodman, A. and S. Webb (1994). For richer, for poorer: the changing distribution of income in the uk, 1961-91. *Fiscal Studies 15*, 29–62.

Hansen, L. P. and K. J. Singleton (1982). Generalized instrumental variable estimation of nonlinear rational expectations models. *Econometrica 50*(4), 1269–1286.

Horowitz, J. (2009). *Semiparametric and Nonparametric Methods in Econometrics: Springer Series in Statistics*. Dordrecht Heidelberg London New York: Springer.

Kitamura, Y., G. Tripathi, and H. Ahn (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica 72*(6), 1667–1714.

Lavergne, P. and V. Patilea (2013). Smooth minimum distance estimation and testing in conditional moment restrictions models: uniform in bandwidth theory. *Journal of Econometrics 177*(1), pp. 47–59.

Nadaraya, E. A. (1964). On estimating regression. *Theory of probability and its applications 9*(1), 141–142.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica 58*, 809–837.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In G. Maddala, C. Rao, and H. Vinod (Eds.), *Sample Surveys: Inference and Analysis*, Volume 11 of *Handbook of Statistics*, pp. 2111–2245. Amsterdam: Elsevier.

Owen, A. B. (1988, June). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–249.

Owen, A. B. (2001). *Empirical Likelihood*. New York: Chapman & Hall.

Polyak, B. T. (1987). *Introduction to Optimization*. New York: Optimization Software, Inc., Publications Division.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *Annals of Statistics 22*(1), pp. 300–325.

Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica 55*, 875–891.

Robinson, P. M. (1991). Best nonlinear three-stage least squares estimation of certain econometric models. *Econometrica 59*(3), 755–786.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

Smith, R. J. (2007). Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics 138*, 430–460.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā, Series A 26*(4), 359–372.

APPENDIX A: PROOFS OF RESULTS

Proof of Lemma 4.1

Since equation (4.34) is not based on any results derived within the current proof, $-2\log\mathcal{R}(\theta_0) = O_p(1)$ and $\forall e_1$, and there exists $\tau < \infty$ such that

$$\mathbb{P}\big(-2\log\mathcal{R}(\theta_0) \leqslant \tau\big) > 1 - e_1. \tag{A.1}$$

Let $\widetilde{\theta} = \theta_0 + Lb_n$, for some $\|L\| = 1$, where b_n denotes an arbitrary sequence such that $nb_n^2 \to \infty$ and $b_n \to 0$. Since $\widehat{\theta}$ minimises $-2\log\mathcal{R}(\theta)$ and $\widehat{\theta}$ is assumed unique, we have

that $-2\log\mathcal{R}(\widetilde{\theta}) \geqslant -2\log\mathcal{R}(\theta_0)$ implies $\widehat{\theta} \in (\theta : \|\theta - \theta_0\| \leqslant b_n)$; that is, $\|\widehat{\theta} - \theta_0\| \leqslant b_n$. Thus,

$$\mathbb{P}\big(\|\widehat{\theta} - \theta_0\| \leqslant b_n\big) \geqslant \mathbb{P}\big(-2\log\mathcal{R}(\widetilde{\theta}) \geqslant -2\log\mathcal{R}(\theta_0)\big). \tag{A.2}$$

We have that Corollary 4.1 and $-2\log\mathcal{R}(\theta_0) \leqslant \tau < \infty$ imply $-2\log\mathcal{R}(\widetilde{\theta}) \geqslant -2\log\mathcal{R}(\theta_0)$ for n large enough, because $-2\log\mathcal{R}(\widetilde{\theta}) \geqslant 0$ and $-2\log\mathcal{R}(\theta_0) \geqslant 0$. Thus, (A.2) implies

$$\mathbb{P}\big(\|\widehat{\theta} - \theta_0\| \leqslant b_n\big) \geqslant \mathbb{P}\Big(\Big(-2\log\mathcal{R}(\widetilde{\theta}) \geqslant r_n\Big) \cap \Big(-2\log\mathcal{R}(\theta_0) \leqslant \tau\Big)\Big). \tag{A.3}$$

Now, Corollary 4.1 and (A.1) imply that the right hand side of (A.3) tends to one. Thus, for n large enough, $\mathbb{P}(\|\widehat{\theta} - \theta_0\| \leqslant b_n) \to 1$; that is, $\|\widehat{\theta} - \theta_0\| = o_p(1)$. □

Proof of Theorem 4.1 Under Conditions 4.2, Young's theorem (Serfling, 1980, p.45) implies

$$\dot{M}_n(\widehat{\theta}) = \dot{M}_n(\theta_0) + \ddot{M}_n(\theta_0)(\widehat{\theta} - \theta_0) + \|\widehat{\theta} - \theta_0\|^2 O_p(1),$$

where $\ddot{M}_n(\theta)$ is defined by (4.18). We have $\dot{M}_n(\widehat{\theta}) = 0$, because $\widehat{\theta}$ is the solution to (3.16). Thus, using (S.7), we have

$$\ddot{M}_n(\theta_0)(\widehat{\theta} - \theta_0) = O_p(n^{-\frac{1}{2}}) + \|\widehat{\theta} - \theta_0\|^2 O_p(1). \tag{A.4}$$

Thus, Lemma 4.1, Condition 4.3 and (A.4) imply (4.23).

Note that $\|t(\theta_0)\| = O_p(n^{-\frac{1}{2}})$ follows from Lemma S1 with $\theta = \theta_0$ and $b_n = n^{-\frac{1}{2}}$.

By using (3.13), the constraint within (3.10) reduces to

$$\sum_{i=1}^{n} \frac{\widehat{h}_i(\widehat{\theta})}{1 + t'(\widehat{\theta})\widehat{h}_i(\widehat{\theta})} = 0_{d_\theta}. \tag{A.5}$$

By substituting (3.13) into (3.12), we obtain

$$\log\mathcal{R}(\widehat{\theta}) = -\sum_{i=1}^{n} \log\big(1 + t'(\widehat{\theta})\widehat{h}_i(\widehat{\theta})\big).$$

Since $\widehat{\theta}$ is the solution to $n^{-1}\partial\log\mathcal{R}(\theta)/\partial\theta = 0$, we have that

$$\sum_{i=1}^{n} p_i(\widehat{\theta})\Big(\frac{\partial t'(\widehat{\theta})}{\partial\widehat{\theta}}\widehat{h}_i'(\theta) + t'(\widehat{\theta})\frac{\partial\widehat{h}_i(\widehat{\theta})}{\partial\widehat{\theta}}\Big) = t'(\widehat{\theta})\sum_{i=1}^{n} p_i(\widehat{\theta})\frac{\partial\widehat{h}_i(\widehat{\theta})}{\partial\widehat{\theta}} = 0_{d_\theta}, \tag{A.6}$$

because of (A.5). By combining equations (A.5) and (A.6), and by using (3.13), we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \frac{c_i(t(\widehat{\theta}), \widehat{\theta})}{1 + t'(\widehat{\theta})\widehat{h}_i(\widehat{\theta})} = 0_{d_\theta + d_\psi}, \tag{A.7}$$

where $c_i(t, \theta) := [\widehat{h}_i'(\theta), (\partial\widehat{h}_i'(\theta)/\partial\theta)t]'$. A Taylor approximation of the left-hand side of (A.7) around $(0_{d_\theta}', \theta_0')'$ gives

$$\frac{1}{n}\sum_{i=1}^{n} \frac{c_i(t(\widehat{\theta}), \widehat{\theta})}{1 + t'(\widehat{\theta})\widehat{h}_i(\widehat{\theta})} = \frac{1}{n}\sum_{i=1}^{n} c_i(0_{d_\theta}, \theta_0) + \widetilde{D}_0\big(t', (\widehat{\theta} - \theta_0)'\big)' + o_p(n^{-\frac{1}{2}}), \tag{A.8}$$

where

$$\widetilde{D}_0 := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial(t', \theta')'}\Big(\frac{c_i(t,\theta)}{1 + t'\widehat{h}_i(\theta)}\Big)\Big|_{\substack{t=0_{d_\theta} \\ \theta=\theta_0}} = \begin{pmatrix} -\Omega_n(\theta_0) & \ddot{M}_n(\theta_0) \\ \dot{M}_n'(\theta_0) & 0_{d_\theta \times d_\theta} \end{pmatrix}.$$

Here, $\ddot{M}_n(\theta)$ and $\Omega_n(\theta)$ are respectively defined by (4.18) and (4.21). Equation (S.2) implies $\Omega_n(\theta_0) = \Omega_n^\star(\theta_0) + O_p(n^{-\frac{1}{2}}) \to \Omega^\star(\theta_0)$ given by (4.22) Assuming that $\ddot{M}_n(\theta_0) \to \ddot{M}(\theta_0)$, the second derivative of $M(\theta)$ defined by (2.6). Thus,

$$\widetilde{D}_0 \to \begin{pmatrix} -\Omega^\star(\theta_0) & 2\ddot{M}(\theta_0) \\ 2\dot{M}'(\theta_0) & 0_{d_\theta \times d_\theta} \end{pmatrix}. \tag{A.9}$$

We also have that

$$\frac{1}{n}\sum_{i=1}^{n} c_i(0_{d_\theta}, \theta_0) = \big(\dot{M}_n(\theta_0), 0_{d_\theta}'\big)', \tag{A.10}$$

where $\dot{M}_n(\theta)$ is defined by (3.16). Using the Schur complement, equations (A.7), (A.8), (A.9) and (A.10) imply

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) = \Big(\dot{M}'(\theta_0)\Omega^\star(\theta_0)^{-1}\dot{M}(\theta_0)\Big)^{-1}\dot{M}'(\theta_0)\Omega^\star(\theta_0)^{-1}n^{\frac{1}{2}}\dot{M}_n(\theta_0) + o_p(1).$$

Since $V[\dot{M}_n(\theta_0)] \doteq V[\dot{M}_n^\star(\theta_0)] = n^{-1}\Omega^\star(\theta_0)$, we have that (4.24) holds. □

Proof of Theorem 4.3 Since ψ_M is the empirical likelihood estimator after imposing the first components $\theta^\dagger$ to be equal to $\theta_0^\dagger$. The proof that led to Theorem 4.1 can be used to show

$$\|\psi_M - \psi_0\| = O_p(n^{-\frac{1}{2}}). \tag{A.11}$$

Condition 4.2 implies

$$\dot{M}_n(\widetilde{\theta}_0) = \dot{M}_n(\theta_0) + \ddot{M}_n(\theta_0)(\psi_M - \psi_0) + \|\psi_M - \psi_0\|^2 O_p(1),$$

where $\ddot{M}_n(\theta)$ is defined by (4.18). Thus, Lemmas S3 and S4, (A.11) and Condition 4.3 imply

$$\dot{M}_n(\widetilde{\theta}_0) = O_p(n^{-\frac{1}{2}}). \tag{A.12}$$

Equation (A.12) implies that we can use Lemma S1 with $\theta = \widetilde{\theta}_0$ and $b_n = n^{-\frac{1}{2}}$. Thus,

$$\|t(\widetilde{\theta}_0)\| = O_p(n^{-\frac{1}{2}}). \tag{A.13}$$

By using (3.13), the constraint within (3.10) reduces to

$$\sum_{i=1}^{n} \frac{\widehat{h}_i(\widetilde{\theta})}{1 + t'(\widetilde{\theta})\widehat{h}_i(\widetilde{\theta})} = 0_{d_\theta}, \quad \text{with } \widetilde{\theta} = (\theta_0^{\dagger\prime}, \psi')'. \tag{A.14}$$

By substituting (3.13) into (3.12), we obtain

$$\log \mathcal{R}(\widetilde{\theta}) = -\sum_{i=1}^{n} \log\big(1 + t'(\widetilde{\theta})\widehat{h}_i(\widetilde{\theta})\big). \tag{A.15}$$

We have that ψ_M is the solution to $n^{-1}\partial \log \mathcal{R}(\widetilde{\theta})/\partial\psi = 0$. This reduces to

$$\sum_{i=1}^{n} p_i(\theta)\Big(\frac{\partial t'(\widetilde{\theta})}{\partial \psi}\widehat{h}_i(\widetilde{\theta}) + t'(\widetilde{\theta})\frac{\partial \widehat{h}_i(\widetilde{\theta})}{\partial \psi}\Big) = t'(\widetilde{\theta})\sum_{i=1}^{n} p_i(\theta)\frac{\partial \widehat{h}_i(\widetilde{\theta})}{\partial \psi} = 0_{d_\psi}, \text{ when } \psi = \psi_M, \tag{A.16}$$

because of (A.14). By combining (A.14) and (A.16), and by using (3.13), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\frac{c_i(t(\widetilde{\theta}),\psi)}{1+t'(\widetilde{\theta})\widehat{h}_i(\widetilde{\theta})} = 0_{d_\theta+d_\psi}, \text{ when } \psi = \psi_M; \tag{A.17}$$

where $c_i(t,\psi) := [\widehat{h}_i'(\theta), (\partial \widehat{h}_i'(\widetilde{\theta})/\partial\psi)t]'$. A Taylor approximation of (A.17) around $(t',\psi')' = (0'_{d_\theta}, \psi_0')'$ gives

$$\frac{1}{n}\sum_{i=1}^{n}\frac{c_i(t,\psi)}{1+t'\widehat{h}_i(\widetilde{\theta})} = \frac{1}{n}\sum_{i=1}^{n} c_i(0_{d_\theta},\psi_0) + D_0\big(t', (\psi\text{-}\psi_0)'\big)' + \big\|\big(t', (\psi\text{-}\psi_0)'\big)\big\|^2 O_p(1), \tag{A.18}$$

where

$$D_0 := \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial(t',\psi')'}\Big(\frac{c_i(t,\psi)}{1+t'\widehat{h}_i(\widetilde{\theta})}\Big)\Big|_{t=0_{d_\theta};\psi=\psi_0} = \begin{pmatrix} -\Omega_n(\theta_0) & \nabla_n \\ \nabla_n' & 0_{d_\psi\times d_\psi}\end{pmatrix}. \tag{A.19}$$

Here, $\Omega_n(\theta)$ and ∇_n are respectively defined by (4.21) and (4.36).

Now, by substituting within (A.18), t and ψ respectively by $t(\widetilde{\theta}_0)$ and ψ_M, we have that (A.11), (A.13) and (A.17) imply that (A.18) reduces to

$$\begin{aligned}\big(t'(\widetilde{\theta}_0), (\psi_M-\psi_0)'\big)' &= -D_0^{-1}\frac{1}{n}\sum_{i=1}^{n} c_i(0_{d_\theta},\psi_0) + O_p(n^{-1}) \\ &= -D_0^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\widehat{h}_i'(\theta_0), 0'_{d_\psi}\Big)' + O_p(n^{-1})\cdot\end{aligned} \tag{A.20}$$

Now, by substituting (A.19) within (A.20), we obtain

$$t(\widetilde{\theta}_0) = (I_{d_\theta\times d_\theta} - A_0)\Omega_n(\theta_0)^{-1}\dot{M}_n(\theta_0) + O_p(n^{-1})\cdot \tag{A.21}$$

A Taylor expansion of (A.15), around $(t'(\widetilde{\theta}_0), \psi_M')' = (0'_{d_\theta}, \psi_0')'$ gives

$$-2\log\mathcal{R}(\widetilde{\theta}_0) = \sum_{i=1}^{n}\log(1+\widetilde{\delta}_i) = 2\sum_{i=1}^{n}\widetilde{\delta}_i - \sum_{i=1}^{n}\widetilde{\delta}_i^2 + 2\sum_{i=1}^{n}\widetilde{\varphi}_i, \tag{A.22}$$

where $\widetilde{\delta}_i := t'(\widetilde{\theta}_0)\widehat{h}_i(\theta_0)$ and $\widetilde{\varphi}_i$ is such that $\mathbb{P}(|\widetilde{\varphi}_i| \le \kappa|\widetilde{\delta}_i|^3, i\in s) \to 1$, for some finite $\kappa > 0$. Now, we follow the same derivation that lead to (S.44) from (S.40). By using (A.13) and (S.37) which holds with $\theta = \theta_0$, equation (A.22) reduces to

$$-2\log\mathcal{R}(\widetilde{\theta}_0) = \sum_{i=1}^{n}\widetilde{\delta}_i^2 + O_p(n^{-\frac{1}{2}}) = n\,t'(\widetilde{\theta}_0)\Omega_n(\theta_0)t(\widetilde{\theta}_0) + O_p(n^{-\frac{1}{2}})\cdot \tag{A.23}$$

Now, by substituting (A.21) within (A.23) and by using Lemma 4.2, we obtain (4.35) (see (S.46)–(S.48) for similar derivation). Finally, $-2\log\mathcal{R}(\widetilde{\theta}_0) \xrightarrow{d} \chi^2_{d_{\theta^\dagger}}$, because $(I_{d_\theta\times d_\theta} - A_0)$ is an idempotent matrix with trace $d_{\theta^\dagger}$. □