

Research Article

Moment Conditions Selection Based on Adaptive Penalized Empirical Likelihood

Yunquan Song^{1,2}

¹ China University of Petroleum, Qingdao 266580, China

² Shandong University Qilu Securities Institute for Financial Studies, Shandong University, Jinan 250100, China

Correspondence should be addressed to Yunquan Song; math1212@163.com

Received 30 March 2014; Revised 29 May 2014; Accepted 5 June 2014; Published 13 July 2014

Academic Editor: Caihong Li

Copyright © 2014 Yunquan Song. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Empirical likelihood is a very popular method and has been widely used in the fields of artificial intelligence (AI) and data mining as tablets and mobile application and social media dominate the technology landscape. This paper proposes an empirical likelihood shrinkage method to efficiently estimate unknown parameters and select correct moment conditions simultaneously, when the model is defined by moment restrictions in which some are possibly misspecified. We show that our method enjoys oracle-like properties; that is, it consistently selects the correct moment conditions and at the same time its estimator is as efficient as the empirical likelihood estimator obtained by all correct moment conditions. Moreover, unlike the GMM, our proposed method allows us to carry out confidence regions for the parameters included in the model without estimating the covariances of the estimators. For empirical implementation, we provide some data-driven procedures for selecting the tuning parameter of the penalty function. The simulation results show that the method works remarkably well in terms of correct moment selection and the finite sample properties of the estimators. Also, a real-life example is carried out to illustrate the new methodology.

1. Introduction

As Xie et al. [1] show, growing attention is being paid to the fields of artificial intelligence (AI) and data mining as tablets and mobile application and social media dominate the technology landscape. Moment conditions often appear in the study of artificial intelligence (AI) and data mining. We all know that empirical likelihood is a very practical tool for the study of moment conditions. When a parametric likelihood function is not specified for a model, estimating equations may provide an alternative instrument for statistical inference. For example, let Z_1, \dots, Z_n be independent and identically distributed random vectors from a distribution, and let $\theta \in \mathbb{R}^p$ be a vector of unknown parameters. Suppose that the information of the distribution is available in the form of an unbiased estimating function $g(z; \theta) = \{g_1(z; \theta), \dots, g_r(z; \theta)\}^T$ satisfying $E\{g(Z; \theta)\} = 0$ and $r \geq p$. When $r = p$, θ can be estimated by solving the estimating equations $0 = n^{-1} \sum_{i=1}^r g(Z_i; \theta)$. Allowing $r > p$ provides a useful device to combine available information

for improving estimation efficiency, but directly solving $0 = n^{-1} \sum_{i=1}^r g(Z_i; \theta)$ may not be feasible.

The generalized method of moments (GMM) and empirical likelihood (EL) are two popular methodologies for estimating the parameters in the structural equations. As was introduced by Hansen [2], the GMM estimator $\hat{\theta}_n$ is defined as

$$\hat{\theta}_G = \arg \min_{\theta \in \Theta} \left[\frac{\sum_{i=1}^n g(Z_i; \theta)}{\sqrt{n}} \right]^T W_n \left[\frac{\sum_{i=1}^n g(Z_i; \theta)}{\sqrt{n}} \right], \quad (1)$$

where Θ is the parameter space where θ_0 lies and W_n is a given $q \times q$ weight matrix. Unlike the GMM, the EL uses likelihood to optimally combine information given in the estimating equations. More specifically, the estimator $\hat{\theta}_E$ is defined by maximizing the following empirical likelihood:

$$L(\theta) = \sup \left\{ \prod_{i=1}^n n \omega_i : \omega_i \geq 0, \sum_{i=1}^n \omega_i = 1, \sum_{i=1}^n \omega_i g(Z_i; \theta) = 0 \right\}. \quad (2)$$

The estimator $\hat{\theta}_E$ is optimal in the sense of Godambe and Heyde [3]. It is known that maximizing (2) is equivalent to minimizing the empirical likelihood ratio

$$l(\theta) = -[\log\{L(\theta) - n \log(n)\}] = \sum_{i=1}^n \log\{1 + \lambda_{\theta}^T g(Z_i; \theta)\}, \quad (3)$$

where λ_{θ} satisfies $n^{-1} \sum_{i=1}^n g(Z_i; \theta)\{1 + \lambda_{\theta}^T g(Z_i; \theta)\}^{-1} = 0$.

Both the GMM and the EL have been successfully used for parameter estimation and variable selection in general estimating equations. The statistical properties of the GMM and the EL estimators rely heavily on the quality of these moment conditions. The strong and valid moment conditions can help to reduce finite-sample bias and improve efficiency of the GMM and the EL estimators. However, when some moment conditions are misspecified, the GMM and the EL estimators may be inconsistent. In this paper, we are interested in estimating some unknown parameter θ_0 identified by a set, set-1, of some moment restrictions which can be used to estimate θ_0 consistently. Meanwhile, it is supposed that there is another set, set-2, of possibly misspecified moment conditions. When the moment conditions in set-2 (or some of them) are correctly specified, including them into estimation equations can improve the asymptotic efficiency of the estimator for θ_0 . However, if they are misspecified, then using these moment conditions will lead to inconsistent estimation. Hence, whenever an empirical researcher has a set of moment conditions and there is no prior information about their validity, it is important to have some procedures to select the correctly specified moment conditions in that set and include them in the estimation equations.

Note that both the GMM estimators and the EL estimators are defined through moment restrictions. They generally have the same asymptotic distributions, but possibly different higher order asymptotic properties; see Newey and Smith [4] and Schennach [5]. As discussed in Newey and Smith [4], the small sample performance of the GMM is poor in some applications and the EL has advantages over the GMM estimators. First, unlike GMM, the asymptotic bias of the EL estimator does not grow with the number of moment restrictions. Consequently, with many moment conditions, the bias of EL will be less than the bias of GMM. The relatively low asymptotic bias of the EL indicates that it is an important alternative to the GMM. Second, unlike the GMM, the EL does not require weight matrix estimation and is invariant to nonsingular linear transformations of the moment conditions. The third theoretical advantage of EL is that after it is bias corrected, it is higher efficient relative to the GMM bias corrected estimators. The reason is that the biased corrected EL estimators inherit the higher order property of maximum likelihood estimators.

Inspired by the idea of Liao [6] and considering the above advantages of the EL estimators relative to the GMM estimators, we propose a novel method for moment selection and parameter estimation simultaneously. The new method attaches a penalty function to the EL criterion and the resulting estimator of θ_0 is then called the EL shrinkage estimator. Our method embeds the moment selection in EL

estimation and once a certain moment condition is selected, it will be automatically included into estimating θ_0 . Hence, our method not only selects the correct moment conditions in the set-2 in one step but also deals with the moment selection and efficient estimation simultaneously. Under some regularity conditions, we show that the EL shrinkage estimator of θ_0 is root- n consistent and asymptotically normal. Moreover, we show that consistent moment selection is automatically achieved in the penalized EL estimation and the EL shrinkage estimator of θ_0 is asymptotically oracle-efficient (i.e., as efficient as the oracle EL estimator based on all valid moment conditions). Unlike the GMM, our proposed method allows us to carry out confidence regions for parameters included in the model without estimating the covariances of the estimators.

The rest of the paper is organized as follows. In Section 2, based on the EL and penalty method, the parameter estimation and moment condition selection are introduced. The theoretical properties of the EL shrinkage estimators and the empirical likelihood ratio are presented in Section 3. Section 4 provides simple and data-driven procedures of selecting the tuning parameters. Simulation studies and a real-life example are given in Section 5. Proofs and the technical derivations are included in the appendix.

2. Methodology

Suppose that we are interested in estimating some unknown parameter θ_0 identified by the following moment restrictions:

$$E[g_q(Z, \theta_0)] = 0, \quad (4)$$

where Z is a d_z -dimensional random vector, θ_0 is a d_{θ} -dimensional parameter vector, the subscript q of $g_q(\cdot, \cdot)$ denotes the number of moment conditions, and $g_q(\cdot, \cdot) : R^{d_z} \times R^{d_{\theta}} \rightarrow R^q$. The moment conditions in (4) can be used to estimate θ_0 consistently. Suppose there is another set of possibly misspecified moment conditions as

$$E[g_k(Z, \theta_0)] \stackrel{?}{=} 0, \quad (5)$$

where “ $\stackrel{?}{=}$ ” signifies that equality may hold for some elements but not others, the subscript k of $g_k(\cdot, \cdot)$ denotes the number of moment conditions, and $g_k(\cdot, \cdot) : R^{d_z} \times R^{d_{\theta}} \rightarrow R^k$. The goal of this paper is to consistently select the correct moment conditions in the set-2 and automatically include them into the empirical likelihood estimation to improve the efficiency of estimating θ_0 .

To incorporate moment selection into the estimation procedure, we first introduce a set of auxiliary unknown parameters β_0 and reparametrize the moment conditions in the set-2 as

$$E[g_k(Z, \theta_0) - \beta_0] = 0. \quad (6)$$

From (6), we see that if the j th ($j = 1, \dots, k$) moment condition in (5) is correctly specified (or misspecified), then $\beta_{0,j} = 0$ (or $\beta_{0,j} \neq 0$). Hence, the zero/nonzero components in β_0

can be used to identify the correctly specified/misspecified moment conditions in the set-2 and consistent moment selection is equivalent to consistent selection of the zero components in β_0 . We thus stack the moment conditions in (4) and (6) to get

$$E[\rho(Z, \theta_0, \beta_0)] \equiv E\left[\begin{pmatrix} g_q(z, \theta_0) \\ g_k(z, \theta_0) - \beta_0 \end{pmatrix}\right] = 0. \quad (7)$$

Let $\{Z_i\}_{i \leq n}$ be a sample of Z . The EL shrinkage estimator $(\hat{\theta}_n^s, \hat{\beta}_n^s)$ of (θ_0, β_0) is defined as

$$\begin{aligned} (\hat{\theta}_n^s, \hat{\beta}_n^s) = \arg \min_{(\theta, \beta) \in \Theta \times \mathcal{B}} \sum_{i=1}^n \log \{1 + \lambda_{\theta}^T \rho(Z_i, \theta, \beta)\} \\ + n \sum_{j=1}^k P_{\tau_n}(\beta_j), \end{aligned} \quad (8)$$

where λ_{θ} satisfies $n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \{1 + \lambda_{\theta}^T \rho(Z_i, \theta, \beta)\}^{-1} = 0$, $\Theta \times \mathcal{B}$ is the parameter space where (θ_0, β_0) lies, and τ_n is the tuning parameter in some general penalty function $P_{\tau_n}(\cdot)$. The success of our method in simultaneous moment selection and efficient estimation relies on the ‘‘oracle properties’’ of the shrinkage techniques. That is to say, if $\beta_{0,j} = 0$, for some $j \in \{1, \dots, k\}$, our method will estimate $\beta_{0,j}$ as zero with probability approaching 1 (w.p.a.1.). When $\beta_{0,j}$ is estimated as zero w.p.a.1., the information contained in the j th moment condition of (5) is automatically used in estimating θ_0 w.p.a.1. On the other hand, the nonzero components in β_0 are consistently estimated and their estimators are nonzero w.p.a.1. Hence, our method can consistently distinguish the zero and nonzero components in β_0 and is consistent in moment selection. Moreover, it estimates θ_0 as if we knew all potentially correct moment conditions in the set-2.

There are many popular choices for the penalty function $P_{\tau_n}(\cdot)$. For example, the bridge penalty is defined as $P_{\tau_n}(\beta) = \tau_n |\beta|^{\gamma}$, where $\gamma \in (0, 1)$; the adaptive Lasso penalty is defined as $P_{\tau_n}(\beta) = \tau_n \hat{\omega}_{\beta} |\beta|$, where $\hat{\omega}_{\beta} = |\hat{\beta}_n|^{-\omega}$ ($\omega > 0$) and $\hat{\beta}_n$ is some first-step consistent estimator of β_0 ; and the smoothly clipped absolute deviation (SCAD) penalty is defined as

$$\begin{aligned} P_{\tau_n}(\beta) = \tau_n |\beta| I(0 \leq |\beta| \leq \tau_n) + \frac{a\tau_n |\beta| - (\beta^2 + \tau_n^2) / 2}{a - 1} \\ \times I(\tau_n \leq |\beta| \leq a\tau_n) + \frac{(a + 1)\tau_n^2}{2} I(|\beta| \geq a\tau_n), \end{aligned} \quad (9)$$

where a is some positive real number strictly larger than 2. The above penalty functions differ in their empirical implementations, although the related EL shrinkage estimators may have the same asymptotic properties (see the results in Section 3). We focus on the SCAD penalty in this paper.

3. Asymptotic Theory

This section establishes the oracle property of the adaptive empirical likelihood (EL) shrinkage estimator. We state our theorems here, but their proofs are relegated to the appendix.

Let $\mathcal{S}_{\beta} \equiv \{j : \beta_{0,j} \neq 0, j = 1, \dots, k\}$ and $\mathcal{S}_{\beta,n} \equiv \{j : \hat{\beta}_{n,j} \neq 0, j = 1, \dots, k\}$ be the index set of the nonzero components in β_0 and $\hat{\beta}_n$, respectively. For ease of notation, we sort the elements in β_0 in the following way: $\beta'_0 = (\beta'_{0,-}, \beta'_{0,+})$, where $\beta_{0,-} = 0$ and $\beta_{0,+} \neq 0$. Let k_0 denote the number of valid moment conditions in the set-2. By definition, we know that $\beta_{0,-}$ and $\beta_{0,+}$ are k_0 and $k - k_0$ dimensional vectors, respectively. We define $\alpha' = (\theta', \beta')$ and

$$\begin{aligned} m(\alpha) &\equiv E[\rho(Z, \theta, \beta)] \\ &\equiv E\left(\begin{pmatrix} g_q(Z, \theta) \\ g_k(Z, \theta) - \beta \end{pmatrix}\right) \equiv \begin{pmatrix} G_q(\theta) \\ G_k(\theta) - \beta \end{pmatrix} \end{aligned} \quad (10)$$

for any $(\theta, \beta) \in \Theta \times \mathcal{B}$. We use $\|\cdot\|$ to denote the Euclidean norm in the Euclidean space.

We first present and discuss the sufficient conditions for consistency of $\hat{\alpha}_n$.

Assumption 1. (i) $E[g_l(z, \theta_0)g_l^T(z, \theta_0)]$ is positive definite for $l = q, k$;

(ii) $\partial g_l(z, \theta) / \partial \theta$ and $\partial^2 g_l(z, \theta) / \partial \theta \partial \theta^T$ are continuous in a neighborhood of the true value θ_0 for $l = q, k$;

(iii) $\|\partial g_l(z, \theta) / \partial \theta\|$, $\|\partial^2 g_l(z, \theta) / \partial \theta \partial \theta^T\|$, and $\|g_l(z, \theta)\|^3$ are bounded by some integrable function $H(z)$ in this neighborhood of the true value θ_0 for $l = q, k$, and the rank of $E[\partial g_{q+k}(z, \theta) / \partial \theta]$ is p ;

(iv) $E\{\sup_{(\theta, \beta) \in \Theta \times \mathcal{B}} (\|\rho(Z_i, \theta, \beta)\| (q + k)^{-1/2})^{\kappa}\} < \infty$ for some $\kappa > 10/3$ when n is large.

Assumption 1 is similar to those of Qin and Lawless [7]. We emphasize that the dimensionality $q + k$ cannot exceed n because the convex hull of $\{g_{q+k}(Z_i, \theta_0)\}_{i=1}^n$ is at most at a subset in \mathbb{R}^n .

Assumption 2. (i) $G_k(\theta)$ is continuous in θ and for any $\varepsilon > 0$ there exists some δ_{ε} such that

$$\inf_{\{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon\}} \|G_q(Z, \theta)\| > \delta_{\varepsilon}; \quad (11)$$

(ii) the following uniform law of large numbers (ULLN) holds:

$$\sup_{\theta \in \Theta} \left[n^{-1} \sum_{i=1}^n \{g_l(Z_i, \theta) - E[g_l(Z_i, \theta)]\} \right] = o_p(1) \quad (12)$$

for $l = q, k$;

(iii) $W_0 = (1/2)E[\rho(Z, \theta, \beta)\rho^T(Z, \theta, \beta)]$ is positive definite; W_n is a symmetric and real matrix and its eigenvalues are bounded from below and above by some fixed finite positive constants for all n ;

(iv) the penalty function $P_{\tau_n}(\cdot)$ is nonnegative and $P_{\tau_n}(\beta_{0,j}) = o_p(1)$ for $j = 1, \dots, k$.

Condition (11) in Assumption 2(i) is the identifiable uniqueness condition for θ_0 . By definition, $\beta_0 = G_k(\theta_0)$; thus, β_0 is locally uniquely identified under (11) and the continuity of $G_k(\theta)$. Assumptions 2(ii) and (iii) are two conditions whose

application range is very wide because it does not specify the data structure, the properties of the moment functions, and the form of the weight matrix W_n . It is clear that when W_n is an identity matrix this assumption holds automatically. We choose $W_n = (1/2n) \sum_{i=1}^n \rho(Z_i, \theta, \beta) \rho^T(Z_i, \theta, \beta)$ usually. Assumption 1(iv) implies that the shrinkage effect of the penalty function on the moment selection coefficients (i.e. $\beta_{0,+}$) converges in probability to zero as $n \rightarrow \infty$. It states that the nonzero parameters cannot converge to zero too fast. This is reasonable because otherwise the noise is too strong. This condition includes the case that $P_{\tau_n}(\beta_{0,j}) = 0$ for $j = 1, \dots, k$ as a special example.

Assumption 3. (i) The following functional central limit theorem (FCLT) holds:

$$\sup_{\theta \in \Theta} \left[n^{-1/2} \sum_{i=1}^n \{g_l(Z_i, \theta) - E[g_l(Z_i, \theta)]\} \right] = O_p(1), \quad (13)$$

for $l = q, k$;

(ii) $G_l(\theta)$ is continuously differentiable in some neighborhood of θ_0 for $l = q, k$;

(iii) $\partial G_l(\theta_0)/\partial \theta'$ has full column rank;

(iv) the penalty function $P_{\tau_n}(\cdot)$ satisfies $P_{\tau_n}(0) = 0$ and is continuously twice differentiable at $\beta_{0,j}$ for any $j \in \mathcal{S}_\beta$ with

$$\max_{j \in \mathcal{S}_\beta} |P_{\tau_n}''(\beta_{0,j})| = o_p(1). \quad (14)$$

Assumption 3(i) can be verified by applying Donsker's theorem in specific models. Assumption 3(ii) imposes a local differentiability condition on the expectation of the moment function $g_l(Z, \theta)$, $l = q, k$. Assumption 3(iii) is a local identification condition for θ_0 . If this assumption fails, the resulting estimator $\hat{\alpha}_n$ may not be \sqrt{n} -consistent. Assumption 3(iv) imposes some local smoothness conditions on the penalty function $P_{\tau_n}(\cdot)$. Intuitively, this condition implies that attaching a penalty function to the empirical likelihood criterion function does not cause any local identification problem for the unknown parameter (θ_0, β_0) . It can be verified that the bridge, adaptive Lasso, and SCAD penalty functions satisfy Assumption 3(iv).

Theorem 4. Under Assumptions 1, 2, and 3, as $n \rightarrow \infty$ and with probability tending to 1, the EL shrinkage estimator defined in (8) satisfies

- (a) $\hat{\alpha}_n \rightarrow \alpha_0$ and
- (b) $\|\hat{\alpha}_n - \alpha_0\| = O_p(\delta_n)$,

where $\delta_n = \max\{b_n, n^{-1/2}\}$ and $b_n = \max_{j \in \mathcal{S}_\beta} |P_{\tau_n}'(\beta_{0,j})|$.

It is clear from Theorem 4 that, by choosing a proper τ_n , there exists a consistent EL shrinkage estimator $\hat{\alpha}_n$ whose convergence rate is of the order δ_n . We now show that this estimator must possess the sparsity property $\hat{\beta}_{n,j}^{\mathcal{S}} = 0$ for all $j \in \mathcal{S}_\beta^c$, which is stated in Theorem 6.

Assumption 5. (i) The tuning parameter τ_n satisfies

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} |P_{\tau_n}'(\beta_{0,j})| = o_p(1); \quad (15)$$

(ii) for any $j \in \mathcal{S}_\beta^c$ and any random sequence $\{\beta_{j,n}\}_n$ with $\beta_{j,n} \neq 0$ a.e. for all n and $\beta_{j,n} = O_p(n^{-1/2})$, there is

$$\liminf_{n \rightarrow \infty} \left[\frac{|P_{\tau_n}'(\beta_{j,n})|}{r_n \tau_n} \right] > 0 \text{ a.e.}, \quad (16)$$

where r_n is some nonnegative sequence such that $n^{1/2} \tau_n r_n \rightarrow \infty$.

Assumption 5(i) indicates that the convergence rate of $|P_{\tau_n}'(\beta_{0,j})|$ for all $j \in \mathcal{S}_\beta$ is faster than \sqrt{n} . Under this assumption, Theorem 4 implies that

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) = O_p(1); \quad (17)$$

that is, the convergence rate of $\hat{\alpha}_n$ is \sqrt{n} . Assumption 5(ii) is a generalized version of condition (3.5) in Fan and Li [8]. Intuitively, Assumption 5(ii) implies that the shrinkage estimator $\hat{\beta}_{n,j}$ of $\beta_{0,j}$, $j \in \mathcal{S}_\beta^c$ is the minimizer of $P_{\tau_n}(\cdot)$ w.p.a.l. From Assumptions 2(iv) and 3(iv), we know that $P_{\tau_n}(\cdot)$ is locally minimized at 0. Hence, Assumption 5(ii) is the key condition needed for showing consistent moment selection. It can be verified that the bridge, adaptive Lasso, and SCAD penalty functions satisfy Assumption 5.

Theorem 6. Under Assumptions 1, 2, 3, and 5, one has

$$\lim_{n \rightarrow \infty} \Pr(\hat{\beta}_{n,j} = 0) = 1 \text{ for any } j \in \mathcal{S}_\beta^c. \quad (18)$$

From the consistency of $\hat{\beta}_n$ and Theorem 6, we can immediately get

$$\lim_{n \rightarrow \infty} \Pr(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}) = 1, \quad (19)$$

that is, the consistent moment selection. We next provide the conditions needed for deriving the limiting distribution of the EL shrinkage estimator.

Assumption 7. (i) There exists a symmetric, nonrandom, and positive definite matrix W_0 such that

$$W_n \xrightarrow{p} W_0; \quad (20)$$

(ii) the following central limit theorem (CLT) holds:

$$n^{-1/2} \sum_{i=1}^n \{g_{q+k}(Z_i, \theta_0) - E[g_{q+k}(Z_i, \theta_0)]\} \xrightarrow{d} \Psi(\theta_0), \quad (21)$$

where $\Psi(\theta_0)$ is some Gaussian random vector.

Assumption 7(i) is a regularity condition. Assumption 7(ii) can be verified by applying CLTs in models with specific moment functions and data structure.

Next we will consider the oracle properties of the EL shrinkage estimation. The oracle properties state that the EL shrinkage estimation can consistently identify all potentially valid moment conditions in set-2 and its estimator of θ_0 is as efficient as the oracle EL estimator based on all valid moment conditions. As the consistent moment selection is directly implied by Theorems 4 and 6, the oracle properties follow if we can show that the asymptotic variance-covariance matrix of $\hat{\beta}_n$ coincides with that of the oracle EL estimator.

Let $g_{d_{\beta_-}}(Z, \theta)$ and $g_{d_{\beta_+}}(Z, \theta)$ denote the potentially valid and misspecified moment functions in set-2, respectively. We define

$$g_{q+d_{\beta_-}}(Z, \theta) \equiv \begin{pmatrix} g_q(Z, \theta) \\ g_{d_{\beta_-}}(Z, \theta) \end{pmatrix}. \tag{22}$$

If we had prior information about the validity of the moment conditions in set-2, then there would be $q + k_0$ moment conditions to estimate θ_0 . We can stack these moment conditions as

$$m_e(\theta_0) = E \left[g_{q+d_{\beta_-}}(Z, \theta_0) \right] = 0. \tag{23}$$

From the moment conditions in (23), we can compute the asymptotic variance-covariance matrix of the optimally oracle EL estimator as

$$\Sigma^* = \left(\left[\frac{\partial m_e(\theta)}{\partial \theta'_0} \right]^T V_{e,o}^{-1} \left[\frac{\partial m_e(\theta)}{\partial \theta'_0} \right] \right)^{-1}, \tag{24}$$

where $V_{e,o}$ is the leading $(q + k_0) \times (q + k_0)$ submatrix of $E[\Psi(\theta_0)\Psi^T(\theta_0)]$.

In the EL shrinkage estimation, if we choose a weight matrix W_n^* such that

$$W_n^* \rightarrow_p W_0 = \{E[\Psi(\theta_0)\Psi^T(\theta_0)]\}^{-1}, \tag{25}$$

then an interesting question is whether the resulting empirical likelihood (EL) shrinkage estimator $\hat{\theta}_n$ of θ_0 could be as efficient as the optimally weighted oracle EL estimator. The answer to the above question is affirmative, as illustrated in the following theorem.

Theorem 8 (Oracle Property). *Under Assumptions 1–5, one has*

$$\lim_{n \rightarrow \infty} Pr(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}) = 1. \tag{26}$$

Furthermore, if the weight matrix W_n satisfies (25) and Assumption 7 holds, then one has

$$\sqrt{n}(\hat{\theta}_n^S - \theta_0) \rightarrow_d N(0, \Sigma^*), \tag{27}$$

where Σ^* is defined in (24).

The empirical likelihood method is capable of finding estimators, constructing confidence regions, and testing hypotheses. The following theorem, a generalization of the Wilks theorem, allows us to carry out inference for parameters included in the model without estimating their estimators' covariance for our proposed method.

Theorem 9. *Suppose Assumptions 1–7 hold. The empirical likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ is*

$$W_E(\theta_0) = 2\ell_E(\theta_0) - 2\ell_E(\hat{\theta}_n^S), \tag{28}$$

where $\ell_E(\theta)$ is given by $\ell_E(\theta) = \sum_{i=1}^n \log[1 + \lambda^T g_{q+k_0}(Z_i, \theta)]$. Under Assumptions 1–7, $W_E(\theta_0) \rightarrow \chi_p^2$ as $n \rightarrow \infty$, where H_0 is true.

Theorem 9 allows us to use the EL ratio statistic for testing or obtaining confidence limits for parameters in a completely analogous way to that for parametric likelihood. The asymptotic confidence region of level $1 - \delta$ for θ_0 is

$$\{\theta_0 \mid W_E(\theta_0) \leq \chi_p^2(\delta)\}, \tag{29}$$

where $\chi_p^2(\delta)$ is the $(1 - \delta)$ quantile of the chi-square distribution with p degrees of freedom.

4. Adaptive Selection of Tuning Parameter

From the results of the previous sections, we see that the tuning parameter λ_n plays an important role in deriving the oracle properties of the EL shrinkage estimator. Assumptions 2(iv), 3(iv), and 5(i)-(ii) are sufficient conditions imposed on λ_n for the oracle properties to hold. However, these conditions do not provide a straightforward mechanism for choosing the tuning parameter λ_n in finite samples. For practical implementation of the shrinkage techniques, it is important to have some procedures of selecting λ_n such that the EL shrinkage estimator not only enjoys the oracle properties asymptotically but also has good finite-sample properties.

To choose the penalty parameter λ , some data-driven approaches for selecting tuning parameters need to be proposed. In the following, we will propose empirical likelihood based AIC-type criterion (EmAIC), BIC-type criterion (EmBIC), and Hannan-Quinn information criterion (HQIC-) type criterion (EmHQIC). They are defined, respectively, as

$$\text{EmAIC}(\lambda) = \text{GEL}_n(\lambda) - 2|\mathcal{S}_{\beta,\lambda}^c|, \tag{30}$$

$$\text{EmBIC}(\lambda) = \text{GEL}_n(\lambda) - \{\log(n)\}|\mathcal{S}_{\beta,\lambda}^c|, \tag{31}$$

$$\text{EmHQIC}(\lambda) = \text{GEL}_n(\lambda) - Q\{\log \log(n)\}|\mathcal{S}_{\beta,\lambda}^c|, \tag{32}$$

where $\mathcal{S}_{\beta,\lambda}$ is the index set of nonzero elements in $\hat{\beta}_{\lambda,n}$, $\mathcal{S}_{\beta,\lambda}^c$ is the complement set of $\mathcal{S}_{\beta,\lambda}$, and $|\mathcal{S}_{\beta,\lambda}^c|$ denotes the cardinality of the index set of $\mathcal{S}_{\beta,\lambda}^c$ and it stands for the number of moment conditions selected by the EL shrinkage method given λ , $Q > 2$. $\text{GEL}_n(\lambda)$ is the generalized empirical likelihood (GEL) statistic proposed in Hong et al. [9], which is defined as $\text{GEL}_n(\lambda) = -2\min_{\alpha_{\mathcal{S}_{\beta,\lambda}}} \max_{\pi} \sum_{i=1}^n \nu[\pi' \rho(Z_i, \alpha_{\mathcal{S}_{\beta,\lambda}})]$, where $\nu(\cdot)$ is some concave function and its domain contains 0, π is some $q + k$ dimensional vector, and $\alpha_{\mathcal{S}_{\beta,\lambda}} = (\theta, \beta_{\mathcal{S}_{\beta,\lambda}}, 0)$ is transferred from α by setting the elements of β whose index belongs to $\mathcal{S}_{\beta,\lambda}^c$ to be zero.

5. Numerical Studies

In this section, we first carry out simulations to demonstrate the performance of our method for finite data sets. We then apply our method to one real dataset. We compare our proposed method with the adaptive EL shrinkage method and the GMM shrinkage method. We find that both the adaptive EL and GMM shrinkage methods can consistently select the correct moment conditions in set-2 and automatically include them into the estimation to improve the efficiency of estimating θ_0 . However, the adaptive EL shrinkage is more efficient relative to the adaptive GMM estimators because of the advantages of EL relative to GMM.

5.1. Simulation Example

Example 1. In this simulation study, the data are generated from the following linear model:

$$Y_i = \theta_{10} + \theta_{20}X_i + u_i, \quad (33)$$

where

$$\begin{aligned} u_i &\sim N(0, \sigma_u^2), & X_i &\sim N(0, \sigma_x^2), \\ E[X_i u_i] &\neq 0, \end{aligned} \quad (34)$$

for all i . The available IVs are $(Z_{1,i}, Z_{2,i})$, where $Z_{1,i}$ is a scale random variable and $Z_{2,i} = (Z_{21,i}, Z_{22,i})$ is a random vector. There are two elements in $Z_{21,i}$ which denote the potentially valid IVs and there are eight elements in $Z_{22,i}$ which are misspecified IVs.

In (33), we take $(\theta_{10}, \theta_{20}) = (0.8, 0.8)$. The random variables X_i , $Z_{1,i}$, $Z_{21,i}$, $Z_{22,i}^*$, and u_i are generated from the following joint normal distribution:

$$(X_i, Z_{1,i}, Z_{21,i}, u_i, Z_{22,i}^*)' \sim N(0, \Sigma), \quad (35)$$

where the diagonal elements of Σ are 1, $E(X_i Z_{1,i}) = \sigma_{z_{1,x}}$, $E(X_i Z_{21,i}) = (\sigma_{z_{2,x}}, \sigma_{z_{2,x}})$, $E(X_i u_i) = 0.4$, and all other elements in Σ are zero. $Z_{22,i}$ is generated by the following equation:

$$Z_{22,i} = Z_{22,i}^* + 0.5u_i * l, \quad (36)$$

where l is a 1×8 vector of ones. The correlation $\sigma_{z_{j,x}}$ of X_i and $Z_{j,i}$ ($j = 1, 2$) measures the signal strength of the IV $Z_{j,i}$ about the endogenous variable X_i . There is one specification of $(\sigma_{z_{1,x}}, \sigma_{z_{2,x}})$ used in the simulation; that is, $(\sigma_{z_{1,x}}, \sigma_{z_{2,x}}) = (0.4, 0.4)$.

We assume the econometrician knows that $Z_{1,i}$ is a valid IV, while being unsure about validity of the IVs in $Z_{2,i}$. Hence, the moment conditions in set-1 are

$$\begin{aligned} E[(Y_i - \theta_{10} - \theta_{20}X_i)] &= 0, \\ E[(Y_i - \theta_{10} - \theta_{20}X_i)Z_{1,i}] &= 0, \end{aligned} \quad (37)$$

while the moment conditions in set-2 are

$$E[(Y_i - \theta_{10} - \theta_{20}X_i)Z_{2,i}'] \stackrel{?}{=} 0. \quad (38)$$

TABLE 1: The selection probabilities of adaptive EL shrinkage estimation.

The selection probabilities	Case	The correct	Underselected	Overselected
P-LA	$n = 100$	0.46	0.53	0.01
	$n = 500$	0.64	0.36	0.00
P-LB	$n = 100$	0.61	0.30	0.09
	$n = 500$	0.89	0.11	0.00

P-LA and P-LB contain the selection probabilities of the correct, underselected, and overselected sets of moment conditions in EL (GMM) shrinkage estimation using the tuning parameters from EmAIC (GMM-AIC) and EmBIC (GMM-BIC), respectively.

TABLE 2: The selection probabilities of adaptive GMM shrinkage estimation.

The selection probabilities	Case	The correct	Underselected	Overselected
P-LA	$n = 100$	0.44	0.55	0.01
	$n = 500$	0.62	0.38	0.00
P-LB	$n = 100$	0.58	0.31	0.11
	$n = 500$	0.86	0.14	0.00

P-LA and P-LB contain the selection probabilities of the correct, underselected, and overselected sets of moment conditions in EL (GMM) shrinkage estimation using the tuning parameters from EmAIC (GMM-AIC) and EmBIC (GMM-BIC), respectively.

The SCAD penalty is used in the empirical likelihood shrinkage estimation, where the first-step estimators of the moment selection coefficients are from the empirical likelihood estimation using the moment conditions in (37) and the reparametrized moment conditions in (38).

For the specification of $(\sigma_{z_{1,x}}, \sigma_{z_{2,x}})$, we use the simulated samples with sample sizes $n = 100$ and $n = 500$, respectively, in our simulation study, and for each sample size, 2000 simulated samples are drawn from the data generating mechanism. With each simulated sample, we calculate four different types of estimators, which include the oracle estimator, empirical likelihood estimator, empirical likelihood shrinkage estimator using λ_n selected by EmAIC, and EL shrinkage estimator using λ_n selected by EmBIC. The oracle estimator is an EL estimator based on the moment conditions in set-1 and all valid moment conditions in set-2. The EL estimator is an EL estimator based only on the moment conditions in the set-1. Given the specification of $(\sigma_{z_{1,x}}, \sigma_{z_{2,x}})$ and the sample size n , we can get 2000 estimators of $(\theta_{1,0}, \theta_{2,0})$ for each type of estimator using the 2000 simulated samples. Hence, we can estimate the finite sample marginal densities of different estimators for $(\theta_{1,0}, \theta_{2,0})$ and the simulation results are presented in Figures 1 and 2. Tables 1 and 2 contain the selection probabilities of the correct, underselected, and overselected sets of moment conditions in EL (GMM) shrinkage estimation using the tuning parameters from EmAIC (GMM-AIC) and EmBIC (GMM-BIC), respectively.

There are several remarks we can make based on the simulation results presented in Figures 1 and 2. First, when

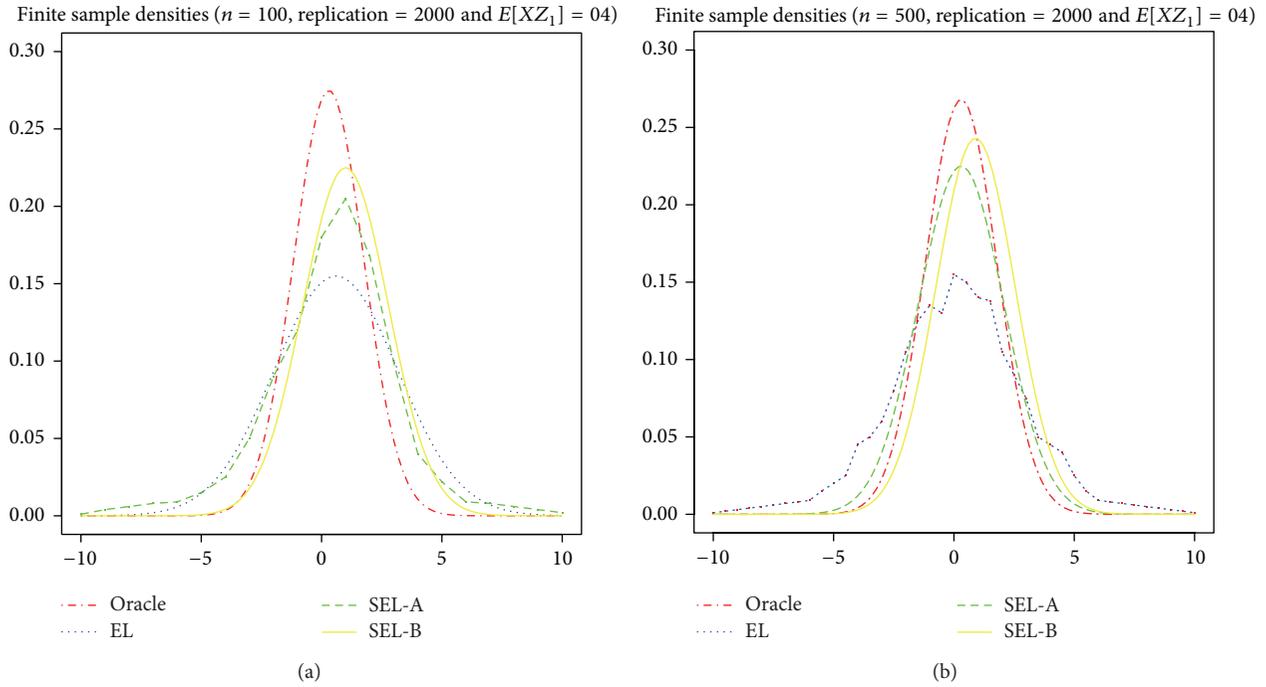


FIGURE 1: Finite sample densities of EL and EL shrinkage estimators for $\theta_{2,0}$ in the cases of example (1). The SCAD penalty is used in the EL shrinkage estimation and the first-step estimators of moment selection coefficients are the EL estimators; (2) oracle estimators are the EL estimators using the moment conditions in set-1 and all correct moment conditions in set-2; (3) EL estimators only use the moment conditions in set-1; (4) SEL-A refers to the EL shrinkage estimators using tuning parameters selected by minimizing EmAIC; (5) SEL-B refers to the EL shrinkage estimators using tuning parameters selected by minimizing EmbIC.

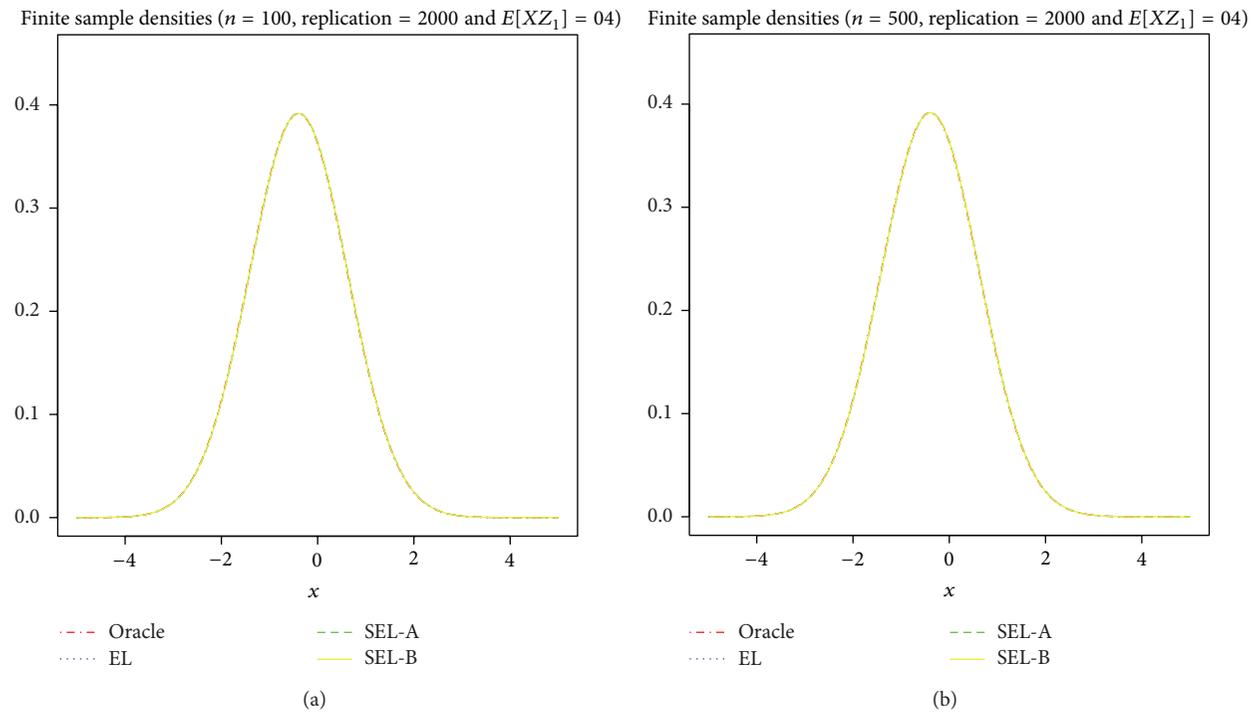


FIGURE 2: Finite sample densities of EL and EL shrinkage estimators for $\theta_{1,0}$ in the cases of example (1). The SCAD penalty is used in the EL shrinkage estimation and the first-step estimators of moment selection coefficients are the EL estimators; (2) oracle estimators are the EL estimators using the moment conditions in set-1 and all correct moment conditions in set-2; (3) EL estimators only use the moment conditions in set-1; (4) SEL-A refers to the EL shrinkage estimators using tuning parameters selected by minimizing EmAIC; (5) SEL-B refers to the EL shrinkage estimators using tuning parameters selected by minimizing EmbIC.

TABLE 3: EL and GMM shrinkage estimators of the moment selection coefficients¹.

Method	IV β_0	edu $\beta_{1,0}$	edu ² $\beta_{2,0}$	edu_ <i>f</i> $\beta_{3,0}$	Age $\beta_{4,0}$	edu * age $\beta_{5,0}$	$\omega_{i,t}^*$ $\beta_{6,0}$	$\omega_{i,t}$ $\beta_{7,0}$
ada-EL	EL ²	-0.0016 (0.0071)	-0.0145 (0.0518)	-0.0022 (0.0057)	0.0049 (0.1145)	-0.0123 (0.4698)	-0.0008 (0.0074)	-0.0441 (0.0996)
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0306
	SEL ^{2,4}	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0097)
	EL ³	-0.0030 (0.0094)	0.0251 (0.0553)	-0.0023 (0.0065)	0.0357 (0.2165)	0.0385 (0.6342)	-0.0015 (0.0091)	0.0496 (0.1143)
		0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	-0.0273 (0.0081)
ada-GMM	GMM ²	-0.0019 (0.0073)	-0.0145 (0.0518)	-0.0021 (0.0057)	0.0048 (0.1145)	-0.0123 (0.4696)	-0.0007 (0.0071)	-0.0442 (0.0995)
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0306
	SGMM ^{2,4}	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0097)
	GMM ³	-0.0034 (0.0095)	0.0253 (0.0558)	-0.0025 (0.0063)	0.0357 (0.2166)	0.0382 (0.6348)	-0.0018 (0.0094)	0.0496 (0.1142)
		0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	-0.0273 (0.0076)

For detailed instructions see Remark 3.

the signal strength of the moment conditions in set-1 is $E[X_i Z_{1,i}] = 0.4$, the EL shrinkage method selects all valid moment conditions in set-2 with high probability and selects the overselected sets of moment conditions with low probability. Second, when the sample size n is increased from 100 to 500, the probability of selecting the set of valid moment conditions in set-2 increases greatly and the probability of selecting the overselected or underselected sets of moment conditions decreases sharply. Third, if we compare the EL shrinkage estimators based on different data-driven procedures of selecting the tuning parameter, we see that the EL shrinkage estimation using the tuning parameters from EmAIC has lower probability of selecting inconsistent sets of moment conditions. But it has nontrivial probability of selecting underselected sets of moment conditions, even when the sample size is increased from 100 to 500. On the other hand, the EL shrinkage estimation using the tuning parameters from EmBIC has lower probability of selecting the overselected sets of moment conditions and higher probability of selecting the set of correct moment conditions, but its probability of selecting the overselected sets of moment conditions is higher. Fourth, the finite sample densities of the EL shrinkage estimators behave much better than those of the EL estimators in all scenarios of this simulation study. Comparing the EL shrinkage estimator with the EL estimator, the most obvious improvement is the reduction of the variance, as we can see from the finite sample densities depicted in Figures 1 and 2. Also note that when the sample size is increased, the finite sample densities of the EL shrinkage estimators are approaching those of the oracle EL estimators. Finally, when the moment conditions in (37) are $E[X_i Z_{1,i}] = 0.4$, the densities of the EL shrinkage estimators $\hat{\theta}_{1,n}$ of $\theta_{1,0}$ are almost the same as those of the oracle estimators and the EL estimators. This is because the

moment conditions in set-2 only contain the information about $\theta_{2,0}$. Hence, when $\theta_{2,0}$ could be reliably estimated using the set-1 moment conditions, the extra valid moment conditions in set-2 do not help to reduce the variances of the estimators of $\theta_{1,0}$.

5.2. Real Data Example

Example 2. We apply the EL shrinkage method to study the following labor supply equation in the life-cycle labor supply model [6, 10, 11]:

$$\Delta \log(h_{i,t}) = \alpha_t + \Delta \log(\omega_{i,t}) \delta_0 + \varepsilon_{i,t}, \tag{39}$$

where $h_{i,t}$ is the annual hours working for money, $\omega_{i,t}$ is the hourly wage rate of individual i at period t , α_t is a time varying constant, $\varepsilon_{i,t}$ is the time varying error term, and δ_0 measures the intertemporal substitution elasticity of labor supply with respect to the evolutionary wage changes and the theoretical prediction for its sign is positive.

Due to the measurement errors in $\omega_{i,t}$, the OLS estimator of (39) may be inconsistent. MaCurdy [10] proposes to use a set of family background variables to construct the set-1 moment conditions; we only use the parents' economic status as the credibly valid IV and include the rest of them into set-2. We also include the alternative measure of wage $\omega_{i,t}^*$ and the wage $\omega_{i,t}$ itself into set-2. Our sample is constructed from the Michigan Panel Study of Income Dynamics (PSID) dataset from year 1970 to year 1981.

We next apply the EL shrinkage estimation to the labor supply equation (39). The estimators of the moment selection coefficients are included in Table 3. As a comparison, we also include the EL estimators of the moment selection coefficients in different specifications of α_t in Table 3. In the

TABLE 4: EL and GMM shrinkage estimation of the labor supply equation¹.

Method	IV	EL ³ (1)	EL ³ (2)	SEL ⁴ (3)	SEL ⁴ (4)	P-EL ⁵ (5)	P-EL ⁵ (6)
Adaptive-EL	a	-0.0157 (0.0253)	-0.0226 (0.0238)	-0.0117 (0.0052)	-0.0226 (0.0196)	-0.0165 (0.0195)	-0.0187 (0.0191)
	δ_0	0.3243 (1.227)	0.2685 (1.145)	0.0697 (0.1469)	0.1832 (0.1923)	0.1395 (0.1455)	0.1726 (0.1397)
	d_t ? ²	No	Yes	No	Yes	No	Yes
Adaptive-GMM	a	-0.0157 (0.025)	-0.0223 (0.0238)	-0.0116 (0.0049)	-0.0228 (0.0199)	-0.0165 (0.0196)	-0.0192 (0.0191)
	δ_0	0.3240 (1.223)	0.2679 (1.141)	0.0697 (0.1469)	0.1827 (0.1918)	0.1398 (0.1455)	0.1721 (0.1393)
	d_t ? ²	No	Yes	No	Yes	No	Yes

For detailed instructions see Remark 4.

first two rows of Table 3, the constant term α_t in (39) is treated to be time variant, while in its last two rows, α_t is taken to be a time invariant constant. From Table 3, we see that the EL estimators of the moment selection coefficients are nonzero and it is hard to determine which moment conditions are valid (misspecified) based on these estimators. On the other hand, the EL shrinkage estimation gives the same moment selection result in the different specifications of α_t . The moment conditions constructed from the IVs by MaCurdy [10] and Altonji [11] are picked up by our shrinkage method, while the moment condition constructed using the imputed wage $\omega_{i,t}$ is not selected, which implies that $\Delta \log(\omega_{i,t})$ is an endogenous variable in the labor supply equation (39).

The results of the EL shrinkage estimation of the labor supply equation (39) are contained in Table 4. As a comparison, we also include the EL estimators of δ_0 based on the moment condition in set-1 and the postmoment selection EL (PEL) estimators of δ_0 in Table 4. Columns (1)-(2) of Table 4 present the EL estimators of δ_0 based on the following IV: parent's economic status when individual was young, which provides the moment condition in set-1. Compared with other estimators in Table 4, the EL estimators in columns (1)-(2) are larger in magnitude and have larger standard errors. On the other hand, the EL shrinkage estimators in columns (3)-(4) have much smaller standard errors, because some moment conditions in set-2 are selected and automatically included into estimation by the EL shrinkage method.

From Table 4, we see that, compared with the EL estimators, the EL shrinkage estimators of $\beta_{7,0}$ are closer to zero, which implies that part of the information in the moment condition constructed by $\omega_{i,t}$ is indeed used in the EL shrinkage estimation. Based on the above reasoning, we can deduce that the shrinkage effect of the penalty function on the estimators of $\beta_{7,0}$ may introduce some bias to the estimator of δ_0 . To get rid of this bias, we conduct another EL estimation based on the moment condition in set-1 and the moment conditions in set-2 selected by our method. These PEL estimators are included in columns (5)-(6) of Table 4. We can see that the PEL estimators are slightly larger in magnitude than the EL shrinkage estimators and their standard errors are almost the same.

Remark 3. We now give detailed instructions about some marks in Table 3.

- (1) Standard errors are in parentheses and $n = 3487$.
- (2) EL (GMM) estimation with the time dummy variables.
- (3) EL (GMM) estimation without time dummy variables.
- (4) EL (GMM) shrinkage estimation with time dummy variables, where the penalty function is the SCAD and the tuning parameter equals 0.000374 (selected by EmAIC (GMM-AIC), EmBIC (GMM-BIC), and EmHQIC (GMM-HQ)).
- (5) EL (GMM) shrinkage estimation without time dummy variables, where the penalty function is the SCAD and the tuning parameters equals 0.000948 (selected by EmAIC (GMM-AIC), EmBIC (GMM-BIC), and EmHQIC (GMM-HQ)).

Remark 4. We now give detailed instructions about some marks in Table 4.

- (1) Standard errors are in parentheses and sample size $n = 3487$.
- (2) d_t refers to the set of time dummy variables for the years from 1971 to 1981.
- (3) EL (GMM) is the EL (GMM) estimation only using the moment conditions in set-1.
- (4) SEL (SGMM) denotes the EL (GMM) shrinkage estimation based on the SCAD penalty. In column (3) the tuning parameter equals 0.000949 and in column (4) the tuning parameter equals 0.000374. EmAIC (GMM-AIC), EmBIC (GMM-BIC), and EmHQIC (GMM-HQ) produce the same number of the tuning parameter in each case.
- (5) PEL (P-GMM) denotes the EL (GMM) estimation based on the moment conditions selected by the EL (GMM) shrinkage estimation. The results in columns (5) and (6) are based on the moment conditions selected in (3) and (4), respectively.

Appendix

Proofs. We first introduce some notations and definitions. Let $\ell(\theta, \beta, \lambda) = n^{-1} \sum_{i=1}^n \log\{1 + \lambda^T \rho(Z_i, \theta, \beta)\}$ and $\bar{\rho}(\theta) = n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta)$. Let $v_n(g) = (1/\sqrt{n}) \sum_{i=1}^n [g(Z_i) - E[g(Z_i)]]$ denote the empirical process indexed by some function g . Suppose that $\{X_n\}$ is a sequence of random vectors; then for a given sequence of nonnegative constants δ_n , we write $X_n = O_p(\delta_n)$ to mean that, for any constant $\varepsilon > 0$, there is a finite constant C_ε such that $\Pr(\|X_n\| > C_\varepsilon \delta_n) < \varepsilon$ eventually; we write $X_n = o_p(\delta_n)$ to mean that, for any constants $\varepsilon_1, \varepsilon_2 > 0$, there is $\Pr(\|X_n\| > \varepsilon_1 \delta_n) < \varepsilon_2$ eventually.

In this appendix, we prove two lemmas which are useful for deriving the asymptotic properties of the EL shrinkage estimator. Define

$$V_n(\theta, \beta) \equiv \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right]^T W_n \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right], \tag{A.1}$$

$$V_{0,n}(\theta, \beta) \equiv \{E[\rho(Z, \theta, \beta)]\}^T W_n \{E[\rho(Z, \theta, \beta)]\}, \tag{A.2}$$

where $W_n = (1/2n) \sum_{i=1}^n \rho(Z_i, \theta, \beta) \rho^T(Z_i, \theta, \beta)$.

Lemma 5. *Under Assumption 1(iii), one has*

$$\frac{1}{2} V_{0,n}(\theta, \beta) - R_n \leq V_n(\theta, \beta) \leq 2V_{0,n}(\theta, \beta) + 2R_n \tag{A.3}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$, where

$$R_n \equiv \sup_{(\theta, \beta) \in \Theta \times \mathcal{B}} n^{-1} \{v_n[\rho(Z, \theta, \beta)]\}^T W_n \{v_n[\rho(Z, \theta, \beta)]\}. \tag{A.4}$$

Proof. By Assumption 2(iii), we deduce that

$$\begin{aligned} & \left[\frac{2 \sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - E[\rho(Z, \theta, \beta)] \right]^T \\ & \times W_n \left[\frac{2 \sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - E[\rho(Z, \theta, \beta)] \right] \geq 0, \end{aligned} \tag{A.5}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$, which implies that

$$\begin{aligned} & \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right]^T W_n \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right] \\ & \geq \frac{1}{2} V_{0,n}(\theta, \beta) - R_n. \end{aligned} \tag{A.6}$$

Note that Assumption 2(iii) also implies

$$\begin{aligned} & \left[\frac{\sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - 2E[\rho(Z, \theta, \beta)] \right]^T \\ & \times W_n \left[\frac{\sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - 2E[\rho(Z, \theta, \beta)] \right] \geq 0 \end{aligned} \tag{A.7}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$, which implies that

$$\begin{aligned} & \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right]^T W_n \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right] \\ & \geq 2V_{0,n}(\theta, \beta) + 2R_n. \end{aligned} \tag{A.8}$$

From the inequalities in (A.6) and (A.8), we immediately get the claimed results in (A.3). \square

Lemma 6. *Under Assumptions 2(iii) and 3(ii)-(iii), one has*

$$\begin{aligned} [c_1 + o(1)] \|\alpha - \alpha_0\|^2 & \leq V_{0,n}(\theta, \beta) \\ & \leq [c_2 + o(1)] \|\alpha - \alpha_0\|^2, \end{aligned} \tag{A.9}$$

for all α in shrinking neighborhoods of α_0 , where c_1, c_2 are generic positive finite constants.

Proof. Denote

$$\begin{aligned} g_q(Z, \theta) & = [g_{q,1}(Z, \theta), \dots, g_{q,q}(Z, \theta)], \\ g_k(Z, \theta) & = [g_{k,1}(Z, \theta), \dots, g_{k,k}(Z, \theta)]. \end{aligned} \tag{A.10}$$

First note that by Assumption 3(ii)

$$\begin{aligned} m(\alpha) & = \begin{pmatrix} G_q(\theta) \\ G_k(\theta) - \beta \end{pmatrix} = \begin{pmatrix} \frac{\partial G_q(\bar{\theta})}{\partial \theta^T} & 0 \\ \frac{\partial G_k(\bar{\theta})}{\partial \theta^T} & -I_k \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \beta - \beta_0 \end{pmatrix} \\ & \equiv \frac{\partial m_e(\bar{\theta})}{\partial \alpha^T} \begin{pmatrix} \theta - \theta_0 \\ \beta - \beta_0 \end{pmatrix}, \end{aligned} \tag{A.11}$$

where

$$\begin{aligned} \frac{\partial G_q(\bar{\theta})}{\partial \theta^T} & = \left[\left(\frac{\partial E[g_{q,1}(Z, \bar{\theta}_1)]}{\partial \theta^T} \right)^T, \dots, \right. \\ & \quad \left. \left(\frac{\partial E[g_{q,q}(Z, \bar{\theta}_1)]}{\partial \theta^T} \right)^T \right], \\ \frac{\partial G_k(\bar{\theta})}{\partial \theta^T} & = \left[\left(\frac{\partial E[g_{k,1}(Z, \bar{\theta}_1)]}{\partial \theta^T} \right)^T, \dots, \right. \\ & \quad \left. \left(\frac{\partial E[g_{k,k}(Z, \bar{\theta}_1)]}{\partial \theta^T} \right)^T \right], \end{aligned} \tag{A.12}$$

$\bar{\theta}$ ($j = 1, \dots, q+k$) lies between θ and θ_0 and I_k is $k \times k$ identity matrix. As θ is in the shrinking neighborhood of θ_0 and $\partial G_l(\theta)/\partial \theta^T$, ($l = q, k$) is continuous in θ ; we deduce that

$$\frac{\partial G_l(\bar{\theta})}{\partial \theta^T} = \frac{\partial G_l(\theta_0)}{\partial \theta^T} + o(1) \quad \text{for } l = q, k. \tag{A.13}$$

By (A.11), (A.13), and Cauchy-Schwarz inequality, we have

$$m(\alpha) = \frac{m(\alpha_0)}{\partial\alpha^T}(\alpha - \alpha_0) + o(\|\alpha - \alpha_0\|). \quad (\text{A.14})$$

Using Assumption 2(iii), the result in (A.14), and Cauchy-Schwarz inequality, we get

$$\begin{aligned} V_{0,n}(\theta, \beta) &= (\alpha - \alpha_0)^T \left[\frac{m(\alpha_0)}{\partial\alpha^T} \right]^T W_n \left[\frac{m(\alpha_0)}{\partial\alpha^T} \right] (\alpha - \alpha_0) \\ &\quad + o(\|\alpha - \alpha_0\|^2). \end{aligned} \quad (\text{A.15})$$

As $\partial G_q(\theta_0)/\partial\theta^T$ has full column rank and is strictly positive definite, $m(\alpha_0)/\partial\alpha^T$ has full rank and $[m(\alpha_0)/\partial\alpha^T]^T W_n [m(\alpha_0)/\partial\alpha^T]$ is strictly positive definite. Let $\gamma_{1,n}$ and $\gamma_{2,n}$ ($\gamma_{1,n}, \gamma_{2,n} > 0$) denote the smallest and largest eigenvalues of $[m(\alpha_0)/\partial\alpha^T]^T W_n [m(\alpha_0)/\partial\alpha^T]$; then by Assumptions 2(iii) and 3(iii), we have

$$0 < c_1 \leq \gamma_{1,n} \leq \gamma_{2,n} \leq c_2 < \infty \quad (\text{A.16})$$

which together with (A.15) implies that

$$\begin{aligned} &(\alpha - \alpha_0)^T \left[\frac{m(\alpha_0)}{\partial\alpha^T} \right]^T W_n \left[\frac{m(\alpha_0)}{\partial\alpha^T} \right] \\ &\quad \times (\alpha - \alpha_0) + o(\|\alpha - \alpha_0\|^2) \\ &\leq \gamma_{2,n} \|\alpha - \alpha_0\|^2 + o(\|\alpha - \alpha_0\|^2) \\ &\leq [c_2 + o(1)] \|\alpha - \alpha_0\|^2. \end{aligned} \quad (\text{A.17})$$

The right inequality in (A.9) is implied by (A.17). The left inequality in (A.9) can be similarly derived. This finishes the proof. \square

Lemma 7. Under Assumption 1, for any ξ with $(1/\kappa + 1/10) \leq \xi \leq 2/5$ and as $n \rightarrow \infty$, $\max_{1 \leq i \leq n} \sup_{(\theta, \beta) \in \Theta \times \mathcal{B}} |\lambda^T \rho(Z_i, \theta, \beta)| = o_p(1)$ for all $\lambda \in \Lambda_n = \{\lambda : \|\lambda\| \leq n^{-\xi}\}$, and $\Lambda_n \subseteq \widehat{\Lambda}_n(\theta) = \{\lambda : \lambda^T \rho(Z_i, \theta, \beta) > -1, i = 1, \dots, n\}$ for all $(\theta, \beta) \in \Theta \times \mathcal{B}$.

Proof. Following Owen [12], Assumption 2(vi) implies that

$$\begin{aligned} \max_{1 \leq i \leq n} \sup_{(\theta, \beta) \in \Theta \times \mathcal{B}} \|\rho(Z_i, \theta, \beta)\| &= o_p(n^{1/5}), \\ \frac{1}{\kappa} + \frac{1}{10} &\leq \xi, \\ \max_{1 \leq i \leq n} \sup_{(\theta, \beta) \in \Theta \times \mathcal{B}} |\lambda^T \rho(Z_i, \theta, \beta)| & \\ \leq n^{-\xi} \max_{1 \leq i \leq n} \sup_{(\theta, \beta) \in \Theta \times \mathcal{B}} |\rho(Z_i, \theta, \beta)| & \\ = o_p(n^{-\xi+1/\kappa}(q+k)^{1/2}). & \end{aligned} \quad (\text{A.18})$$

\square

Lemma 8. Let $\ell(\theta, \beta, \lambda) = n^{-1} \sum_{i=1}^n \log\{1 + \lambda^T \rho(Z_i, \theta, \beta)\}$; one has

$$\begin{aligned} \ell(\theta, \beta, \lambda) &= \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right]^T \\ &\quad \times W_n \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right] + o_p\left(\frac{1}{n}\right), \end{aligned} \quad (\text{A.19})$$

where $W_n = (1/2n) \sum_{i=1}^n \rho(Z_i, \theta, \beta) \rho^T(Z_i, \theta, \beta)$.

The detailed theoretical reasons of this lemma can be found in Following Hjort et al. [13], so we neglect it here.

Proof of Theorem 4. (a) By the definition of $\widehat{\alpha}_n$, one has

$$\ell(\widehat{\theta}_n, \widehat{\beta}_n, \lambda) + \sum_{j=1}^k P_{\tau_n}(\widehat{\beta}_n) \leq \ell(\theta_0, \beta_0, \lambda) + \sum_{j=1}^k P_{\tau_n}(\beta_{0,j}). \quad (\text{A.20})$$

Applying Lemma 7, one can deduce that

$$\begin{aligned} &V_n(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^k P_{\tau_n}(\widehat{\beta}_n) + o_p\left(\frac{1}{n}\right) \\ &\leq V_n(\theta_0, \beta_0) + \sum_{j=1}^k P_{\tau_n}(\beta_{0,j}) + o_p\left(\frac{1}{n}\right); \end{aligned} \quad (\text{A.21})$$

that is,

$$V_n(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^k P_{\tau_n}(\widehat{\beta}_n) \leq V_n(\theta_0, \beta_0) + \sum_{j=1}^k P_{\tau_n}(\beta_{0,j}). \quad (\text{A.22})$$

Applying Lemma 5 and Assumption 2(iv), one can deduce from (A.1) that

$$V_{0,n}(\widehat{\theta}_n, \widehat{\beta}_n) \leq 2 \sum_{j=1}^k P_{\tau_n}(\beta_{0,j}) + 4R_n, \quad (\text{A.23})$$

where R_n is defined in Lemma 5.

From Assumption 2(ii) and the definition of $\rho(Z, \theta, \beta)$, one gets

$$\begin{aligned} &\sup_{(\theta, \beta) \in \Theta \times \mathcal{B}} \frac{v_n[\rho(Z, \theta, \beta)]}{\sqrt{n}} \\ &= \sup_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \{g_{q+k}(Z_i, \theta) - E[g_{q+k}(Z_i, \theta)]\} = o_p(1). \end{aligned} \quad (\text{A.24})$$

By the triangle inequality, ULLN in (A.24), and Assumptions 2(iii)-(iv), one has

$$R_n = o_p(1), \quad \sum_{j=1}^k P_{\tau_n}(\beta_{0,j}) = o_p(1). \quad (\text{A.25})$$

From Assumption 2(iii) and results in (A.23) and (A.25), one can deduce that

$$\|G_q(\widehat{\theta}_n)\| = o_p(1), \quad \|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\| = o_p(1). \quad (\text{A.26})$$

Now, the first result in (A.26) and Assumption 1(i) imply that $\widehat{\theta}_n \rightarrow_p \theta_0$. From the second result in (A.26), the triangle inequality, consistency of $\widehat{\theta}_n$, and Assumption 2(i), one has

$$\begin{aligned} o_p(1) &= \|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\| \\ &\geq \|G_k(\widehat{\theta}_n) - G_k(\widehat{\theta}_0)\| - \|\widehat{\beta}_n - \beta_0\| \\ &= \|\widehat{\beta}_n - \beta_0\| + o_p(1), \end{aligned} \quad (\text{A.27})$$

which implies that $\widehat{\beta}_n \rightarrow_p \beta_0$. So one gets the result that as $n \rightarrow \infty$ and with probability tending to 1, the EL shrinkage estimator defined in (6) satisfies $\widehat{\alpha}_n \rightarrow \alpha_0$.

(b) Using the inequalities in (A.3) and (A.22) and Lemma 7, one gets

$$\frac{1}{2}V_{0,n}(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^k p_{\tau_n}(\widehat{\beta}_{n,j}) \leq \sum_{j=1}^k p_{\tau_n}(\widehat{\beta}_{0,j}) + 2R_n, \quad (\text{A.28})$$

where R_n is defined in Lemma 5. By Assumptions 2(iv) and 3(iv) and the inequality in (A.28), one has

$$V_{0,n}(\widehat{\theta}_n, \widehat{\beta}_n) + 2 \sum_{j \in \mathcal{S}_\beta} [p_{\tau_n}(\widehat{\beta}_{n,j}) - p_{\tau_n}(\widehat{\beta}_{0,j})] \leq 4R_n. \quad (\text{A.29})$$

Next, by Assumption 3(iv), Taylor expansion, the triangle inequality, and Cauchy-Schwarz inequality, one gets

$$\begin{aligned} &|p_{\tau_n}(\widehat{\beta}_{0,j}) - p_{\tau_n}(\widehat{\beta}_{n,j})| \\ &= \left| \sum_{j \in \mathcal{S}_\beta} \left[p'_{\tau_n}(\widehat{\beta}_{0,j})(\widehat{\beta}_{n,j} - \widehat{\beta}_{0,j}) + \frac{1}{2} p''_{\tau_n}(\widetilde{\beta}_j)(\widehat{\beta}_{n,j} - \widehat{\beta}_{0,j})^2 \right] \right| \\ &\leq \max_{j \in \mathcal{S}_\beta} |p'_{\tau_n}(\widehat{\beta}_{0,j})| \|\widehat{\alpha}_n - \alpha_0\| \\ &\quad + \max_{j \in \mathcal{S}_\beta} \left| \frac{p''_{\tau_n}(\beta_{0,j})}{2} + o_p(1) \right| \|\widehat{\alpha}_n - \alpha_0\|^2, \end{aligned} \quad (\text{A.30})$$

where $\widetilde{\beta}_j$ lies between $\beta_{0,j}$ and $\widehat{\beta}_{n,j}$ for $j \in \mathcal{S}_\beta$. From Lemma 6, one obtains

$$V_{0,n}(\widehat{\theta}_n, \widehat{\beta}_n) \geq [c_1 + o_p(1)] \|\widehat{\alpha}_n - \alpha_0\|^2, \quad (\text{A.31})$$

where $c_1 > 0$ is a finite constant. The inequality in (A.31), together with Assumption 3(iv), and the inequalities in (A.29) and (A.30) imply that

$$\begin{aligned} &[c_1 + o_p(1)] \|\widehat{\alpha}_n - \alpha_0\|^2 \\ &\quad - 2 \max_{j \in \mathcal{S}_\beta} |p'_{\tau_n}(\widehat{\beta}_{0,j})| \|\widehat{\alpha}_n - \alpha_0\| \leq 4R_n. \end{aligned} \quad (\text{A.32})$$

By Assumption 3(i), one has $R_n = O_p(n^{-1})$. When the sample size n is large enough, by definition, the probability that $c_1 + o_p(1) \leq c_1/2$ is strictly smaller than any given small number $\omega/2$, together with the inequality in (A.32), implies that

$$\begin{aligned} &\Pr\left(\frac{\|\widehat{\alpha}_n - \alpha_0\|}{\delta_n} > M\right) \\ &\leq \inf_{\alpha \in \mathcal{A}: \|\alpha - \alpha_0\| > M\delta_n} \|\alpha - \alpha_0\| \\ &\leq 2c_1 \left(b_n + \sqrt{b_n^2 + \frac{2R_n}{c_1}} \right) + \frac{\omega}{2} \\ &\leq \Pr\left(M < \frac{2c_1 \left(b_n + \sqrt{b_n^2 + 2R_n/c_1} \right)}{\delta_n} \right) + \frac{\omega}{2}, \end{aligned} \quad (\text{A.33})$$

where $\delta_n = \max\{b_n, n^{-1/2}\}$. By definition, $2c_1(b_n + \sqrt{b_n^2 + 2R_n/c_1})/\delta_n = O_p(1)$. Hence, one can choose some large enough number M_ω such that

$$\Pr\left(M < \frac{2c_1 \left(b_n + \sqrt{b_n^2 + 2R_n/c_1} \right)}{\delta_n} \right) < \frac{\omega}{2}. \quad (\text{A.34})$$

This and the results in (A.33) immediately imply that

$$\Pr\left(\frac{\|\widehat{\alpha}_n - \alpha_0\|}{\delta_n} > M\right) < \omega \quad (\text{A.35})$$

eventually, which gives us $\|\widehat{\alpha}_n - \alpha_0\| = O_p(\delta_n)$. \square

Proof of Theorem 6. We know that

$$\begin{aligned} \ell(\theta, \beta, \lambda) &= \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right]^T \\ &\quad \times W_n \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right] + o_p\left(\frac{1}{n}\right). \end{aligned} \quad (\text{A.36})$$

On the event $\{\widehat{\beta}_{n,j} \neq 0\}$ for some $j \in \mathcal{S}_\beta^c$, we have the following KKT optimality condition:

$$\begin{aligned} &2 \left[n^{-1/2} \sum_{i=1}^n \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n) \right]^T W_n \left[n^{-1/2} \sum_{i=1}^n \frac{\partial \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n)}{\partial \beta_j} \right], \\ &\quad + n p'_{\tau_n}(\widehat{\beta}_{n,j}) = 0 \end{aligned} \quad (\text{A.37})$$

which implies

$$\left| W_n(q+j) \left[n^{-1/2} \sum_{i=1}^n \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n) \right] \right| = \frac{\sqrt{n} |p'_{\tau_n}(\widehat{\beta}_{n,j})|}{2}, \quad (\text{A.38})$$

where $W_n(q+j)$ denotes the $(q+j)$ th row of the weight matrix W_n .

By Assumption 3(ii) and the consistency of $\widehat{\theta}_n$, there is

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n) &= v_n [\rho(Z, \widehat{\alpha}_n)] + n^{1/2} m(\widehat{\alpha}_n) \\ &= v_n [\rho(Z, \widehat{\alpha}_n)] + \frac{\partial m(\widehat{\theta}_n)}{\partial \alpha^T} [n^{1/2}(\widehat{\alpha}_n - \alpha_0)], \end{aligned} \quad (\text{A.39})$$

where $\partial m(\theta)/\partial \alpha^T$ is defined in (A.11) and $\widehat{\theta}_n = (\widehat{\theta}_{1,n}, \dots, \widehat{\theta}_{q+k,n})$ and $\widehat{\theta}_{j,n}$ ($j = 1, \dots, q+k$) lie between θ_0 and $\widehat{\theta}_n$. From Assumption 3(i), we have $v_n[\rho(Z, \widehat{\alpha}_n)] = O_p(1)$. By Theorem 4 and Assumption 5(i), we have $n^{1/2}(\widehat{\alpha}_n - \alpha_0) = O_p(1)$. By the triangle inequality, Assumption 3(ii), and the consistency of $\widehat{\theta}_n$, we deduce that

$$\left\| \frac{\partial m(\widehat{\theta}_n)}{\partial \alpha^T} \right\| \leq \left\| \frac{\partial m(\widehat{\theta}_n)}{\partial \alpha^T} - \frac{\partial m(\theta_0)}{\partial \alpha^T} \right\| + \left\| \frac{\partial m(\theta_0)}{\partial \alpha^T} \right\| = O_p(1). \quad (\text{A.40})$$

Hence, we have $n^{-1/2} \sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n) = O_p(1)$ which combined with Assumption 2(iii) implies that

$$\left| W_n(q+j) \left[n^{-1/2} \sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n) \right] \right| = O_p(1). \quad (\text{A.41})$$

By Theorem 6 and Assumption 5(i), we have $n^{1/2} \widehat{\beta}_{n,j} = O_p(1)$ for all $j \in \mathcal{S}_\beta^c$. Hence, conditional on the event $\{\widehat{\beta}_{n,j} \neq 0\}$ for some $j \in \mathcal{S}_\beta^c$, we can invoke Assumption 5(ii) to deduce that

$$\frac{n^{1/2} |P'_{\tau_n}(\widehat{\beta}_{n,j})|}{2} = \frac{n^{1/2} r_n \tau_n |P'_{\tau_n}(\widehat{\beta}_{n,j})|}{r_n \lambda_n} \xrightarrow{p} \infty, \quad (\text{A.42})$$

for $j \in \mathcal{S}_\beta^c$. Now, using (A.38), (A.41), and (A.42), we deduce that $\Pr(\widehat{\beta}_{n,j} = 0) \rightarrow 1$ as $n \rightarrow \infty$ for any $j \in \mathcal{S}_\beta^c$. \square

Proof of Theorem 8. Let $g_{d_{\beta_-}}(Z, \theta)$ and $g_{d_{\beta_+}}(Z, \theta)$ denote the potentially valid and misspecified moment functions in set-2, respectively. We define

$$\begin{aligned} g_{q+d_{\beta_-}}(Z, \theta) &\equiv \begin{pmatrix} g_q(Z, \theta) \\ g_{d_{\beta_-}}(Z, \theta) \end{pmatrix}, \\ \frac{\partial m(\theta_0)}{\partial \alpha_\mathcal{S}^T} &= \begin{pmatrix} \frac{\partial E[g_{q+d_{\beta_-}}(Z, \theta_0)]}{\partial \theta^T} & 0 \\ \frac{\partial E[g_{d_{\beta_+}}(Z, \theta_0)]}{\partial \theta^T} & -I_{k-k_0} \end{pmatrix}, \end{aligned} \quad (\text{A.43})$$

where I_{k-k_0} denotes a $(k-k_0) \times (k-k_0)$ identity matrix. If we define

$$M_\mathcal{S} = \left[\frac{\partial m(\theta_0)}{\partial \alpha_\mathcal{S}^T} \right]^T W_0 \left[\frac{\partial m(\theta_0)}{\partial \alpha_\mathcal{S}^T} \right], \quad (\text{A.44})$$

then under Assumptions 2(iii) and 3(iii), we know that $M_\mathcal{S}$ is nonsingular matrix.

Recall that $\alpha_{0,\mathcal{S}}^T = (\theta_0^T, \beta_{0,+}^T)$ and accordingly $\widehat{\alpha}_{n,\mathcal{S}}^T = (\widehat{\theta}_n^T, \widehat{\beta}_{n,+}^T)$. For any compact subset K in $R^{d_\theta+d_{k-k_0}}$, we denote any element $u_\mathcal{S} \in K$ as $u_\mathcal{S}^T = (u_\theta^T, u_{\beta,+}^T)$, where u_θ are the first d_θ elements in $u_\mathcal{S}$ and $u_{\beta,+}$ are the last d_{k-k_0} elements in $u_\mathcal{S}$. Denote

$$\begin{aligned} V_{2,n}(u_\mathcal{S}) &= \left[n^{-1/2} \sum_{i=1}^n \rho^\mathcal{S} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) \right]^T \\ &\quad \times W_n \left[n^{-1/2} \sum_{i=1}^n \rho^\mathcal{S} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) \right] \\ &\quad - \left[n^{-1/2} \sum_{i=1}^n \rho(Z, \alpha_0) \right]^T \\ &\quad \times W_n \left[n^{-1/2} \sum_{i=1}^n \rho(Z, \alpha_0) \right] \\ &\quad + n \sum_{j \in \mathcal{S}} \left[P_{\tau_n} \left(\beta_{0,j} + \frac{u_{\beta_+,j}}{\sqrt{n}} \right) - P_{\tau_n}(\beta_{0,j}) \right] \\ &\equiv V_{2,n}^*(u_\mathcal{S}) \\ &\quad + n \sum_{j \in \mathcal{S}} \left[P_{\tau_n} \left(\beta_{0,j} + \frac{u_{\beta_+,j}}{\sqrt{n}} \right) - P_{\tau_n}(\beta_{0,j}) \right], \end{aligned} \quad (\text{A.45})$$

where $\rho^\mathcal{S}(Z, \alpha_{0,\mathcal{S}} + u_\mathcal{S}/\sqrt{n}) \equiv \rho(Z_i, \theta_0 + u_\theta/\sqrt{n}, \beta_{0,+} + u_{\theta_0,+}/\sqrt{n}, \beta_{0,-})$. From Theorem 6, we know that $\widehat{\beta}_{0,-} = 0$ w.p.a.1. Thus, $\sqrt{n}(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{0,\mathcal{S}})$ is the minimizer of $V_{2,n}(u_\mathcal{S})$ w.p.a.1.

If we define

$$\mathcal{F}_n \equiv \left(f_{u_\mathcal{S}}^n = \rho^\mathcal{S} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) - \rho(Z, \alpha_0) : u_\mathcal{S} \in K \right), \quad (\text{A.46})$$

then by Assumptions 3(i) we know that \mathcal{F}_n is a Donsker class. As K is compact, so there exists some constant C_k , such that $\sup_{u_\mathcal{S} \in K} \|n^{-1/2} u_\mathcal{S}\| \leq n^{-1/2} C_k = o(1)$. Now we can use Lemma 2.17 in Pakes and Pollard [14] to deduce that

$$v_n \left(\rho^\mathcal{S} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) - \rho(Z, \alpha_0) \right) = o_p(1), \quad (\text{A.47})$$

uniformly over $u_\mathcal{S} \in K$.

By Assumption 3(ii) and the compactness of K , we have

$$\begin{aligned} \sqrt{n} \left(E \left[\rho^\mathcal{S} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) \right] - E[\rho(Z, \alpha_0)] \right) \\ = \frac{\partial m(\theta_0)}{\partial \alpha_\mathcal{S}^T} u_\mathcal{S} + o(1), \end{aligned} \quad (\text{A.48})$$

uniformly over $u_{\mathcal{S}} \in K$. Thus, (A.47) and (A.48) imply that, uniformly over $u_{\mathcal{S}} \in K$, there is

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^n \rho^{\mathcal{S}} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}} \right) \\ &= v_n \left(\rho^{\mathcal{S}} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}} \right) - \rho(Z, \alpha_0) \right) \\ &+ v_n [\rho(Z, \alpha_0)] \\ &+ \sqrt{n} \left(E \left[\rho^{\mathcal{S}} \left(Z, \alpha_{0,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}} \right) \right] - E[\rho(Z, \alpha_0)] \right) \\ &= v_n [\rho(Z, \alpha_0)] + \left[\frac{\partial m(\theta_0)}{\partial \alpha'_{\mathcal{S}}} \right] u_{\mathcal{S}} + o_p(1). \end{aligned} \quad (\text{A.49})$$

Now, we can use the result in (A.49), Assumptions 3(i) and 7(i), and the compactness of K to deduce that

$$\begin{aligned} V_{2,n}^*(u_{\mathcal{S}}) &= u_{\mathcal{S}}^T \left[\frac{\partial m(\theta_0)}{\partial \alpha'_{\mathcal{S}}} \right]^T \\ &\times W_0 \left[\frac{\partial m(\theta_0)}{\partial \alpha'_{\mathcal{S}}} \right] u_{\mathcal{S}} \\ &+ 2u_{\mathcal{S}}^T \left[\frac{\partial m(\theta_0)}{\partial \alpha'_{\mathcal{S}}} \right]^T \\ &\times W_0 \{v_n[\rho(Z, \alpha_0)]\} + o_p(1), \end{aligned} \quad (\text{A.50})$$

uniformly over $u_{\mathcal{S}} \in K$. If $j \in \mathcal{S}$, then by Assumptions 3(iv) and 5(i) and the compactness of K we have

$$\begin{aligned} & n \left[p_{\tau_n} \left(\beta_{0,j} + \frac{u_{\beta_{+,j}}}{\sqrt{n}} \right) - p_{\tau_n}(\beta_{0,j}) \right] \\ &= \sqrt{n} p'_{\tau_n}(\hat{\beta}_{0,j}) u_{\beta_{+,j}} + [p''_{\tau_n}(\hat{\beta}_{0,j}) + o_p(1)] u_{\beta_{+,j}}^2 \rightarrow 0 \end{aligned} \quad (\text{A.51})$$

uniformly over $u_{\beta_{+,j}}$.

Using Assumption 7(ii), the results in (A.50), and (A.51), we get

$$\begin{aligned} V_{2,n}(u_{\mathcal{S}}) &\rightarrow_d V_2(u_{\mathcal{S}}) = u_{\mathcal{S}}^T M_{\mathcal{S}} u_{\mathcal{S}} \\ &+ 2u_{\mathcal{S}}^T \left[\frac{\partial m(\theta_0)}{\partial \alpha'_{\mathcal{S}}} \right]^T W_0 \Psi(\theta_0), \end{aligned} \quad (\text{A.52})$$

uniformly over $u_{\mathcal{S}} \in K$. It is clear that $V_2(u_{\mathcal{S}})$ is uniquely minimized at

$$u_{\mathcal{S}}^* = -M_{\mathcal{S}}^{-1} \left[\frac{\partial m(\theta_0)}{\partial \alpha'_{\mathcal{S}}} \right]^T W_0 \Psi(\theta_0). \quad (\text{A.53})$$

By Theorem 4 and Assumption 5(i), there is

$$\sqrt{n}(\hat{\alpha}_{n,\mathcal{S}} - \alpha_{0,\mathcal{S}}) = O_p(1). \quad (\text{A.54})$$

Now, the uniform weak convergence in (A.52), the unique minimization in (A.53), and the asymptotic tightness of $\hat{\alpha}_{n,\mathcal{S}}$ in (A.54) enable us to invoke the ACMT to deduce that

$$\sqrt{n}(\hat{\alpha}_{n,\mathcal{S}} - \alpha_{0,\mathcal{S}}) \rightarrow_d N(0, M_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S}} M_{\mathcal{S}}^{-1}). \quad (\text{A.55})$$

The first result is implied by Theorems 4 and 6, so we only need to show the second claim. First note that if

$$W_n \rightarrow_p W_0 = \{E[\Psi(\theta_0) \Psi^T(\theta_0)]\}^{-1}, \quad (\text{A.56})$$

then the centered limiting distribution in (A.55) will be simplified to

$$\sqrt{n}(\hat{\alpha}_{n,\mathcal{S}} - \alpha_{0,\mathcal{S}}) \rightarrow_d N(0, M_{11}^{-1}). \quad (\text{A.57})$$

Denote Ω_{θ_0} to be the first $d_{\theta_0} \times d_{\theta_0}$ submatrix of $M_{\mathcal{S}}^{-1}$ and $\partial G_{\beta_{+}}(\theta_0)/\partial \theta^T = \partial E[g_{d_{\beta_{+}}}(Z, \theta_0)]/\partial \theta^T$. Note that

$$\begin{aligned} M_{\mathcal{S}} &= \begin{pmatrix} \frac{\partial E[g_{q+d_{\beta_{-}}}(Z, \theta_0)]}{\partial \theta^T} & 0 \\ \frac{\partial E[g_{d_{\beta_{+}}}(Z, \theta_0)]}{\partial \theta^T} & -I_{d_{\beta_{0}^{+}} \times d_{\beta_{0}^{+}}} \end{pmatrix}^T \\ &\times W_0 \begin{pmatrix} \frac{\partial E[g_{q+d_{\beta_{-}}}(Z, \theta_0)]}{\partial \theta^T} & 0 \\ \frac{\partial E[g_{d_{\beta_{+}}}(Z, \theta_0)]}{\partial \theta^T} & -I_{d_{\beta_{0}^{+}} \times d_{\beta_{0}^{+}}} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{\mathcal{S}}^{11} & \Sigma_{\mathcal{S}}^{12} \\ \Sigma_{\mathcal{S}}^{21} & W_{22} \end{pmatrix}, \end{aligned} \quad (\text{A.58})$$

where

$$\begin{aligned} \Sigma_{\mathcal{S}}^{11} &= \left[\frac{\partial m_e(\theta_0)}{\partial \theta^T} \right]^T W_{11} \left[\frac{\partial m_e(\theta_0)}{\partial \theta^T} \right] \\ &+ \left[\frac{\partial m_e(\theta_0)}{\partial \theta^T} \right]^T W_{12} \left[\frac{\partial G_{\beta_{+}}(\theta_0)}{\partial \theta^T} \right] \\ &+ \left[\frac{\partial G_{\beta_{+}}(\theta_0)}{\partial \theta^T} \right]^T W_{21} \left[\frac{\partial m_e(\theta_0)}{\partial \theta^T} \right] \\ &+ \left[\frac{\partial G_{\beta_{+}}(\theta_0)}{\partial \theta^T} \right]^T W_{22} \left[\frac{\partial G_{\beta_{+}}(\theta_0)}{\partial \theta^T} \right], \\ \Sigma_{\mathcal{S}}^{12} &= - \left[\frac{\partial m_e(\theta_0)}{\partial \theta^T} \right]^T W_{12} - \left[\frac{\partial G_{\beta_{+}}(\theta_0)}{\partial \theta^T} \right]^T W_{22} = (\Sigma_{\mathcal{S}}^{21})^T. \end{aligned} \quad (\text{A.59})$$

From (A.37), it is easy to get

$$\begin{aligned} \Omega_{\theta_0}^{-1} &= \Sigma_{\mathcal{S}}^{11} W_{22}^{-1} \Sigma_{\mathcal{S}}^{21} \\ &= \left[\frac{\partial m_e(\theta_0)}{\partial \theta^T} \right]^T V_{e,0}^{-1} \left[\frac{\partial m_e(\theta_0)}{\partial \theta^T} \right] = (\Sigma^*)^{-1}, \end{aligned} \quad (\text{A.60})$$

where the last equality is due to the fact that $(W_{11} - W_{12}W_{22}^{-1}W_{21})^{-1} = V_{e,0}$. Now, using results in (A.34) and (A.38) and the Continuous Mapping Theorem (CMT), we can deduce that

$$\sqrt{n}(\hat{\theta}_n^s - \theta_0) \longrightarrow_d N(0, \Sigma^*), \tag{A.61}$$

which establishes the semiparametric efficiency of the GMM shrinkage estimator $\hat{\theta}_n^s$. \square

Proof of Theorem 9. We denote $S_{22.1}^{-1} = (E(\partial g_{q+k_0}/\partial \theta)^T E(gg^T)E(\partial g_{q+k_0}/\partial \theta))^{-1}$, $S_{11} = -E(g_{q+k_0}g_{q+k_0}^T)$, $S_{22} = E(\partial g_{q+k_0}/\partial \theta)$, $S_{21} = E(\partial g_{q+k_0}/\partial \theta)^T$, and $\ell_E(\theta) = \sum_{i=1}^n \log[1 + \lambda^T g_{q+k_0}(Z_i, \theta)]$. The log-empirical likelihood ratio test statistic is

$$W_E(\theta_0) = 2 \left\{ \sum_i \log [1 + \lambda^T g_{q+k_0}(Z_i, \theta_0)] - \sum_i \log [1 + \lambda^T g_{q+k_0}(Z_i, \hat{\theta}_n^s)] \right\}. \tag{A.62}$$

Note that

$$\begin{aligned} \ell_E(\hat{\theta}_n^s) &= \sum_{i=1}^n \log \{1 + \lambda^T g_{q+k_0}(Z_i, \hat{\theta}_n^s)\} \\ &= -\frac{n}{2} Q_{1n}^T(\theta_0, 0) A Q_{1n}(\theta_0, 0) + o_p(1), \end{aligned} \tag{A.63}$$

where

$$\begin{aligned} Q_{1n}(\theta) &= \frac{1}{n} \sum_i \frac{1}{1 + \lambda^T g_{q+k_0}(Z_i, \theta)} g_{q+k_0}(Z_i, \theta), \\ Q_{2n}(\theta) &= \frac{1}{n} \sum_i \frac{1}{1 + \lambda^T g_{q+k_0}(Z_i, \theta)} \left(\frac{\partial g_{q+k_0}(Z_i, \theta)}{\partial \theta} \right)^T \lambda, \\ A &= S_{11}^{-1} \{I + S_{12} S_{22.1}^{-1} S_{21} S_{11}^{-1}\}. \end{aligned} \tag{A.64}$$

Also under H_0 ,

$$\begin{aligned} &\frac{1}{n} \sum_i \frac{1}{1 + \lambda^T g_{q+k_0}(Z_i, \theta_0)} g_{q+k_0}(Z_i, \theta_0) \\ &= 0 \implies \lambda_0 = -S_{11}^{-1} Q_{1n}(\theta_0, 0) + o_p(1), \\ &\sum_i \log [1 + \lambda_0^T g_{q+k_0}(Z_i, \theta_0)] \\ &= -\frac{n}{2} Q_{1n}^T(\theta_0, 0) S_{11}^{-1} Q_{1n}(\theta_0, 0) + O_p(1). \end{aligned} \tag{A.65}$$

Thus,

$$\begin{aligned} W_E(\theta_0) &= n Q_{1n}^T(\theta_0, 0) (A - S_{11}^{-1}) Q_{1n}(\theta_0, 0) + o_p(1) \\ &= [(-S_{11}^{-1/2}) \sqrt{n} Q_{1n}(\theta_0, 0)]^T \\ &\quad \times [(-S_{11}^{-1/2}) S_{12} S_{22.1}^{-1} S_{21} (-S_{11}^{-1/2})] \\ &\quad \times [(-S_{11}^{-1/2}) \sqrt{n} Q_{1n}(\theta_0, 0)] + o_p(1). \end{aligned} \tag{A.66}$$

Note that $(-S_{11}^{-1/2}) \sqrt{n} Q_{1n}(\theta_0, 0)$ converges to a standard multivariate normal distribution and that $(-S_{11}^{-1/2}) S_{12} S_{22.1}^{-1} S_{21} (-S_{11}^{-1/2})$ is symmetric and idempotent, with trace equal to p . Hence, the empirical likelihood ratio statistic $W_E(\theta_0)$ converges to χ_p^2 . \square

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the NSF Project (ZR2011AQ007) of Shandong Province of China, NNSF Projects (11201499 and 11301309) of China, and the Fundamental Research Funds for the Central Universities (27R1310008A).

References

- [1] F. Xie, S. Fan, J. Wang, H. Lu, and C. Li, "Special issue on "Artificial intelligence and data mining," *Abstract and Applied Analysis*, 2014.
- [2] L. P. Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica*, vol. 50, no. 4, pp. 1029–1054, 1982.
- [3] V. P. Godambe and C. C. Heyde, "Quasi-likelihood and optimal estimation," *International Statistical Review*, vol. 55, no. 3, pp. 231–244, 1987.
- [4] W. K. Newey and R. J. Smith, "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, vol. 72, no. 1, pp. 219–255, 2004.
- [5] S. M. Schennach, "Bayesian exponentially tilted empirical likelihood," *Biometrika*, vol. 92, no. 1, pp. 31–46, 2005.
- [6] Z. Liao, "Adaptive GMM shrinkage estimation with consistent moment selection," *Econometric Theory*, vol. 29, no. 5, pp. 857–904, 2013.
- [7] J. Qin and J. Lawless, "Empirical likelihood and general estimating equations," *The Annals of Statistics*, vol. 22, no. 1, pp. 300–325, 1994.
- [8] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [9] H. Hong, B. Preston, and M. Shum, "Generalized empirical likelihood-based model selection criteria for moment condition models," *Econometric Theory*, vol. 19, no. 6, pp. 923–943, 2003.

- [10] T. MaCurdy, "An empirical model of labor supply in a life-cycle setting," *Journal of Political Economy*, vol. 89, no. 6, pp. 1059–1085, 1981.
- [11] J. Altonji, "Intertemporal substitution in labor supply: evidence from micro data," *Journal of Political Economy*, vol. 94, no. 3, pp. 176–215, 1986.
- [12] A. B. Owen, *Empirical Likelihood*, Chapman and Hall-CCRC, New York, NY, USA, 2001.
- [13] N. L. Hjort, I. W. McKeague, and I. van Keilegom, "Extending the scope of empirical likelihood," *The Annals of Statistics*, vol. 37, no. 3, pp. 1079–1111, 2009.
- [14] A. Pakes and D. Pollard, "Simulation and the asymptotics of optimization estimators," *Econometrica*, vol. 57, no. 5, pp. 1027–1057, 1989.