A Frisch-Waugh-Lovell Theorem for Empirical Likelihood

Adam Dearing* Yichun Song†

October 18, 2022

Abstract

We present a Frisch-Waugh-Lovell-type (FWL) theorem for empirical likelihood estimation with instrumental variables. Our theorem is similar to the standard FWL theorem for partitioning-out exogenous regressors (including dummy variables), but the partitioning-out procedure uses the empirical likelihood weights at the solution instead of the original empirical distribution. We show how to leverage this result to simplify computation via an iterative estimation algorithm, where exogenous variables are partitioned out using weighted least squares, and these weights are updated between iterations. Furthermore, we show that iterations converge locally to the full empirical likelihood estimate at a stochastically super-linear rate. We also provide a feasible iterative constrained optimization algorithm for calculating empirical-likelihood-based confidence intervals and discuss its properties. Monte Carlo simulations on both point estimation and confidence intervals demonstrate that our iterative algorithms are robust and generate results within numerical tolerance of the full empirical likelihood estimator in the finite sample, while offering substantial savings on computation time as the number of exogenous regressors grows.

Keywords: Partitioned regression, Empirical Likelihood, Confidence Interval.
JEL classification codes: C18, C26, C36.

*Cornell University and NBER, SC Johnson College of Business, Ithaca NY 14853-6201. Email: aed237@cornell.edu

†**Corresponding author.** Department of Economics, The Ohio State University, Columbus OH 43210-1120. Email: song.1399@osu.edu.

1 Introduction and Literature Review

In this study, we consider empirical likelihood (EL) estimation in a linear instrumental variables model. Suppose we have a sample at hand, containing $i = 1, \ldots, N$ i.i.d. observations of an outcome variable, $y_i \in \mathbb{R}$, endogenous regressors, $w_i \in \mathbb{R}^K$, exogenous regressors, $x_i \in \mathbb{R}^M$, and instruments, $z_i \in \mathbb{R}^L$, where $L \geq K$ for identification. The researcher specifies a regression of the form $y_i = w_i'\beta_w^* + x_i'\beta_x^* + \varepsilon_i$, with the exogeneity condition, $E[\varepsilon_i | z_i, x_i] = 0$. The researcher's task is then to construct an estimator of $\beta^* = (\beta_w^*, \beta_x^*)$, where β_w^* is the parameter of interest, and β_x^* is a set of nuisance parameters. The estimator is based on the unconditional moments $E[\varepsilon_i \otimes (z_i', x_i')'] = 0$.

There are several methods from which a researcher could choose in facing those moment conditions. For example, these moments can serve as the basis for two-stage least squares (2SLS), the generalized method of moments (GMM), two-step GMM, iterated GMM, continuously-updated GMM, or empirical likelihood estimation (and its generalizations).

In this paper, we focus on empirical likelihood. In our instrumental variables setting, empirical likelihood estimates are defined as follows:

$$\begin{aligned} \left(\hat{\beta}_w^{EL}, \hat{\beta}_x^{EL}, \hat{\pi}^{EL}\right) = \underset{\beta_w, \beta_x, \pi}{\arg\max} \quad & \frac{1}{N}\sum_{i=1}^{N} \ln \pi_i \\ s.t. \quad & \sum_{i=1}^{N} \pi_i = 1 \\ & \sum_{i=1}^{N} \pi_i \left(y_i - w_i'\beta_w - x_i'\beta_x\right) \otimes \begin{pmatrix} z_i \\ x_i \end{pmatrix} = 0. \end{aligned} \tag{1}$$

Pioneered by Owen (1988), Owen (1990), Owen (1991) and extended to general moment-based estimation by Qin and Lawless (1994), empirical likelihood methods can be used for estimation, hypothesis testing, and constructing confidence intervals. Chen and Keilegom (2009) provides a comprehensive review on EL methods on parametric, semi-parametric, and non-parametric regression models.

Empirical likelihood estimates and confidence intervals have important theoretical and practical advantages over other methods. First, they do not suffer from many-instruments bias and have

lower finite-sample bias than GMM-based procedures (Newey and Smith (2004)) [1]. Second, empirical likelihood ratio confidence intervals are correct up to order $O(N^{-1})$, as opposed to the usual $O(N^{-1/2})$ for confidence intervals based on normal approximations (Chen (1993)). Chen and Hall (1993), Chen (1993), and Chen and Cui (2007) successively showed that confidence region using EL for quantile regression, for linear regression, and GMM-type estimators, are all subject to a delicate empirical Bartlett correction. Bartlett correction further reduces the magnitude of the coverage error to $O(N^{-2})$. Third, the empirical likelihood ratio test is the most powerful semi-parametric test when the model is correctly specified; for instance, Dong and Giles (2007) develops such a test for testing normality. Fourth, the estimates have a straightforward interpretation under misspecification, similar to maximum likelihood (see, e.g., White (1982) and Schennach (2007)).

However, there is no free lunch. These advantages come at the cost of increased computational burden relative to GMM-based estimation methods, especially 2SLS, two-step GMM, and iterated GMM. While those other estimators have closed-form solutions (or iterate on them), computing the empirical likelihood estimates requires solving a non-concave maximization problem, which at first appears to be a very large-scale optimization problem over $K+M+N$ variables. Fortunately, duality results allow researchers to reduce the search over (β_w, β_x, π) to a search over $(\beta_w, \beta_x, \lambda)$, where λ are Lagrange multipliers on the moment constraints. This change decreases the number of search variables from $K+M+N$ to $K+2M+L$, where usually $M+L \ll N$.[2] While computation remains feasible even in large samples, the maximization problem is still non-concave and does not have a closed-form solution, with numerical search needed over the $K+2M+L$ unknown parameters. This can become particularly problematic when researchers want to control many exogenous covariates so that M is large. [3]

This large-M problem can also arise in other estimation methods – even for OLS and 2SLS. In those settings, it can be alleviated by partitioned regression methods. Perhaps the simplest example arises when we have $z_i \equiv w_i$, and we can consistently estimate the model with ordinary least squares (OLS) regression. In this case, OLS is equivalent to all of the IV methods previously mentioned. OLS has many favorable properties, and we are particularly interested in the Frisch-

1. However, as illustrated in Guggenberger and Hahn (2005) and Hausman et al. (2011), the Generalized EL-type (GEL) estimator may suffer from the no finite-sample moment problem, similar to LIML. Hausman et al. (2011) then propose a modified moment condition to alleviate this problem in GEL, which includes the CUE estimator of Hansen et al. (1996).

2. We provide a brief description of the duality result for both point estimation and confidence interval in Appendix A.

3. Although, we still assume that M is small relative to N.

Waugh-Lovell (FWL) theorem (Frisch and Waugh (1933) and Lovell (1963)). The FWL theorem establishes the validity of partitioned regression for OLS. That is, there are two ways to compute the OLS estimate of β_w:

> Method 1. Simultaneously estimate β_w^* and β_x^* using standard OLS.
>
> Method 2. Obtain the residuals, $\tilde{y}_i$ and $\tilde{w}_i$, from OLS regressions of y_i and w_i, respectively, on x_i. Then estimate β_w^* by regressing $\tilde{y}_i$ on $\tilde{w}_i$ via OLS (omitting x_i).

Both of these methods will produce the exact finite-sample estimate, $\hat{\beta}_w^{OLS}$. In the second method, x_i is "partitioned-out" of y_i and w_i, and we run OLS on the residuals of the partitioning-out process. The FWL theorem is particularly useful in practice when dealing with fixed effects, where it implies the finite-sample equivalence of the fixed effects estimator with full OLS when x_i includes dummy variables. The fixed effects estimator incurs less computational burden since it avoids the inversion of a large matrix.

Giles (1984) extends the Frisch-Waugh-Lovell theorem to 2SLS regression, where $z_i \neq w_i$, and shows that it admits several equivalent formulas for $\hat{\beta}_w^{2SLS}$. Similar results extend to GMM and two-step GMM, but the result does not extend to iterated nor continuously-updated GMM (Baum et al. (2007)).

One potential potential solution to the large-M problem in empirical likelihood is to partition-out the exogenous variables, similar to the FWL theorem with OLS, 2SLS, and the two-step GMM. That is, the researcher first obtains partitioned-out residuals $\tilde{y}_i$, $\tilde{w}_i$, and $\tilde{z}_i$ as they would for those other methods, and then replaces the full empirical likelihood problem in (1) with

$$\begin{aligned} \max_{\beta_w, \pi} \quad & \frac{1}{N} \sum_{i=1}^{N} \ln \pi_i \\ s.t. \quad & \sum_{i=1}^{N} \pi_i = 1 \\ & \sum_{i=1}^{N} \pi_i \left(\tilde{y}_i - \tilde{w}_i' \beta_w \right) \otimes \tilde{z}_i = 0. \end{aligned} \tag{2}$$

Unfortunately, the standard FWL equivalence result for OLS, 2SLS, and two-step GMM does not apply to empirical likelihood: in over-identified models ($L > K$), the solution to (2) will generally not be equal to $(\hat{\beta}_w^{EL}, \hat{\pi}^{EL})$ in the finite sample. But all is not lost; the process will still give a consistent estimate when the model is correctly specified, and that estimate of β_w will be asymp-

totically equivalent to the full empirical likelihood estimate up to $O(N^{-1/2})$. These properties have been discussed in several papers, especially those considering application in panel data, where the partitioning-out of fixed effects is needed to prevent the so-called incidental parameter problem. For instance, Zhang et al. (2011) and Eryuruk (2010) discuss the asymptotic properties of the one-step partitioning-out estimator in the context of dynamic panel data models.

Even so, this one-step partitioning-out estimator may lose some of the advantages of the full empirical likelihood solution in finite sample. Thus, an open question arises: is there a way to exploit FWL-type results to alleviate computational burden for full EL, while retaining equivalence in the finite sample?

Motivated by that question, this paper makes several contributions to the empirical likelihood literature. First, we provide a valid Frisch-Waugh-Lovell Theorem for full empirical likelihood estimation. The key insight of our theorem is that the partitioning-out process should be done with the empirical likelihood weights ($\pi_i = \pi_i^{EL}$), while standard FWL results partition-out using the sampling distribution ($\pi_i = 1/N$). Second, we provide an algorithm that leverages this result to simplify computation of the empirical likelihood estimator in the finite sample. Our algorithm iterates between a weighted partitioning-out procedure and a reduced-dimension empirical likelihood problem, with weights updated at each iteration. We show that this algorithm converges to the full EL estimate asymptotically at a stochastically super-linear rate. Third, we present a similar iterative algorithm to compute the bounds for empirical likelihood confidence intervals. In developing the iterative scheme for confidence intervals, we derive novel quasi-duality results for the corresponding constrained optimization problem and discuss some basic properties. We also show how Bartlett correction can be implemented with our algorithm.

We demonstrate the performance of our iterative algorithm through a series of Monte Carlo experiments. Our simulations indicate that our iterative algorithm is superior to both the original (full) empirical likelihood algorithm and to one-step partitioning-out. Our algorithm gives almost identical solutions (i.e., within numerical tolerance) to the original empirical likelihood algorithm, while one-step partitioning produces larger discrepancies. Additionally, our iterative algorithms exhibit remarkable speed advantages when the number of exogenous covariates is large.

We organize the rest of this paper as follows. Section 2 provides a discussion of the iterative algorithm for point empirical likelihood estimation. Section 3 provides a similar iterative algorithm to

the EL-based confidence intervals. Section 4 contains Monte Carlo simulations. Section 5 discusses further extensions, and Section 6 concludes.

2 Algorithm for Point Estimation

In this section, we describe the iterative algorithm and its properties. We begin with some assumptions that guarantee the partitioning-out process and the full empirical likelihood estimation are well-behaved.

Assumption 1. $\sum_{i=1}^{N} x_i x_i'$ *is non-singular.*

Assumption 2. *The finite-sample full empirical likelihood problem has a unique maximizer.*

The unique maximizer for full EL problem is denoted as $(\hat{\beta}_w^{EL}, \hat{\beta}_x^{EL}, \hat{\pi}^{EL})$. It is harmless to have Assumption 1 and 2 such that the full EL is well-defined and has a proper (unique) solution.

Then, we define our partitioned-out estimator $(\hat{\beta}_w^P, \hat{\beta}_x^P, \pi^P)$ as the fixed point with the largest primal objective function value among all other fixed point solutions $(\check{\beta}_w, \check{\beta}_x, \check{\pi})$. And a fixed point is calculated using the Algorithm 1 below,

Algorithm 1. *An iterative algorithm for empirical likelihood estimation that leverages FWL proceeds on in the following steps:*

1. *Obtain an initial feasible solution to the full empirical likelihood problem,* $(\beta_w^0, \beta_x^0, \pi^0)$.

2. *For* $k \geq 1$, *obtain* $(\tilde{y}^{k-1}, \tilde{w}^{k-1}, \tilde{z}^{k-1})$ *as the residuals from weighted least squares (WLS) regression of* y, w, *and* z *on* x *with* π^{k-1} *as weights.*

3. *Update* β_w *and* π*:*

$$\begin{aligned}\left(\beta_w^k, \pi^k\right) = \underset{\beta_w, \pi}{arg\ max} \quad & \frac{1}{N}\sum_{i=1}^{N} \ln \pi_i \\ s.t. \quad & \sum_{i=1}^{N} \pi_i = 1 \\ & \sum_{i=1}^{N} \pi_i \left(\tilde{y}_i^{k-1} - \left(\tilde{w}_i^{k-1}\right)' \beta_w\right) \otimes \tilde{z}_i = 0.\end{aligned}$$

4. *Repeat steps 2 and 3 until convergence, and the solution at convergence is* $(\check{\beta}_w, \check{\pi})$.

5. Calculate the corresponding $\check{\beta}_x$ from $(\check{\beta}_w, \check{\pi})$ by just-identification

$$\check{\beta}_x = \left[\sum_{i=1}^{N} \check{\pi}_i x_i x_i'\right]^{-1} \left[\sum_{i=1}^{N} \check{\pi}_i x_i (y_i - w_i' \check{\beta}_w)\right] \tag{3}$$

In practice, the primal problem illustrated in Algorithm 1 should be implemented in its dual form. Algorithm 1 is similar in spirit to other iterative methods for likelihood problems, such as the EM algorithm (Dempster et al. (1977)), in the sense that it replaces one large, difficult problem with a sequence of much simpler problems. Algorithm 1 replaces the full EL problem's search over $K+2M+L$ parameters with a sequence of smaller-scale searches over $K+L$ parameters. This can greatly simplify computation as M becomes relatively large, allowing for a large number of control variables in the model.

Some important properties of Algorithm 1 are discussed in the Theorem 1 below.

Theorem 1. *(Frisch-Waugh-Lovell for EL) Under Assumptions 1, 2 and Assumptions 4 and 5 in Appendix B, Algorithm 1 has the following properties,*

1. Suppose (β_w, β_x, π) is feasible to the full empirical likelihood problem. Then (β_w, β_x, π) is feasible for the partitioned-out primal empirical likelihood problem when partitioned with weights π.

2. For any fixed point solution $(\check{\beta}_w, \check{\pi})$ in Algorithm 1, the just-identified estimation of $\check{\beta}_x$ in Step 5, combined with $(\check{\beta}_w, \check{\pi})$, is a feasible solution to the full empirical likelihood problem.

3. The full EL estimate $(\hat{\beta}_w^{EL}, \hat{\beta}_x^{EL})$ satisfies the first order condition for the partitioned problem, when partitioned by $\hat{\pi}^{EL}$ in finite sample. And as $N \to \infty$, the full empirical likelihood estimation $(\hat{\beta}_w^{EL}, \hat{\beta}_x^{EL}, \hat{\pi}^{EL})$ becomes the fixed-point solution $(\hat{\beta}_w^{P}, \hat{\beta}_x^{P}, \hat{\pi}^{P})$ of the partitioned problem almost surely.

Proof. *See Appendix B.* □

Theorem 1.1 shows that the partitioned problem is primal feasible in the finite sample whenever the original full problem is feasible. Theorem 1.2 demonstrates that any fixed point of the partitioned-out primal problem has a corresponding primal solution to the full problem. Theorem 1.3 establishes the main result of this paper. Specifically, while $(\hat{\beta}_w^{EL}, \hat{\beta}_x^{EL}, \hat{\pi}^{EL})$ satisfies the first order necessary condition (FOC) in finite sample when partitioned with $\hat{\pi}^{EL}$, the second-order sufficient condition

(SOC) for $(\hat{\beta}_w^{EL}, \hat{\beta}_x^{EL}, \hat{\pi}^{EL})$ takes a complicated form that converges almost surely to a negative definite matrix as $N \to \infty$.[4]

Together, the results in Theorem 1 show that Algorithm 1 provides a convenient and reliable method for computing the empirical likelihood estimate and corresponding weights, $(\hat{\beta}_w^{EL}, \hat{\pi}^{EL})$, when it converges [5]. Theorem 2 shows that the local convergence is quite fast, occurring at a "stochastically super-linear" rate (Kasahara and Shimotsu (2008)) asymptotically.

Theorem 2. *Under Assumptions 1, 2, and Assumptions 4 and 5 in Appendix B, the iterations defined by Algorithm 1 satisfy*

$$\left\|\hat{\beta}_w^k - \hat{\beta}_w^{EL}\right\| = O_p\left(N^{-1/2}\left\|\hat{\beta}_w^{k-1} - \hat{\beta}_w^{EL}\right\| + \left\|\hat{\beta}_w^{k-1} - \hat{\beta}_w^{EL}\right\|^2\right)$$

Proof. *See Appendix B.* □

3 Algorithm for Confidence Intervals

One of the advantages of the empirical likelihood framework is that it leads to empirical likelihood ratio tests and confidence intervals (CIs). Like full maximum likelihood CIs, the empirical likelihood ratio CIs can be asymmetric and are amenable to Bartlett correction (Chen (1993), Chen and Cui (2007)). The CIs are constructed by inverting the empirical likelihood ratio tests, which can be a numerically cumbersome procedure also.

We focus on $(1-\alpha)$-CIs for a scalar-valued, continuous function of the vector of parameters of interest, $T(\beta_w) : \mathbb{R}^K \to \mathbb{R}$. This includes individual components of the vector by setting $T(\beta_w) = e_j'\beta_w$, where e_j is the jth standard basis vector, so that $e_j'\beta_w$ gives the j'th component of β_w. Reich and Judd (2020) recently proposed a constrained optimization approach to compute the upper and lower bounds of full maximum likelihood CIs. We adopt their approach for empirical likelihood

4. For more discussion regarding this asymptotic results on the SOC, readers may refer to Appendix B.

5. Hansen and Lee (2021) in their iterated GMM paper show the existence of fixed point by contraction mapping, which is different from ours. As in Theorem 1, we illustrate that the FOC holds in the finite sample, and asymptotically, SOC holds almost surely as $N \to \infty$ when partitioned out by $\hat{\pi}_{EL}$. Indeed, there is a fixed point, and our study then focuses on the algorithms to obtain the fixed point.

CIs. The upper-bound for this unidimensional CI, $\overline{C}^{EL}$, can then be computed as

$$\begin{aligned}
\overline{C}^{EL} = \max_{\beta_w, \beta_x, \pi} \quad & T(\beta_w) \\
s.t. \quad & \sum_{i=1}^{N} \pi_i = 1 \\
& \sum_{i=1}^{N} \pi_i \left(y_i - w_i'\beta_w - x_i'\beta_x\right) \otimes \begin{pmatrix} z_i \\ x_i \end{pmatrix} = 0 \\
& \sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} - \sum_{i=1}^{N} \ln \pi_i \leq \chi_1^2(1-\alpha)/2.
\end{aligned} \tag{4}$$

where $\chi_1^2(1-\alpha)$ indicates the $(1-\alpha)$ quantile of the χ_1^2 distribution. The lower-bound, $\underline{C}^{EL}$, is obtained by simply replacing max with min. Therefore, we limit our attention to the details of computing the upper bound for parsimony.

One issue with the above optimization problem (4) is that, like the estimation problem in Section 2, the number of search parameters grows with the sample size. To remedy this issue, we provide an alternative "quasi-dual" problem in Appendix A. It is called quasi-dual because of its non-standard dual formation but still successfully transforms the high-dimension $(K + M + N)$ primal problem into an equivalent low-dimension problem $(K + 2M + L)$.

Even when using the alternative quasi-dual representation of this problem for dimension reduction, the same computational difficulties arise here as in the point estimation when there is a large number of exogenous covariates, M. Nevertheless, similar to estimation, we can use a simple iterative algorithm to alleviate the computational burden. Here, we present the algorithm for computing $\overline{C}^{EL}$. Simply replacing the arg max in Step 2 with arg min would give the algorithm for $\underline{C}^{EL}$.

Algorithm 2. *An iterative algorithm for EL-CIs that leverages FWL proceeds on in the following steps:*

1. *Set* $(\beta_w^0, \pi^0, \overline{C}^0) = (\hat{\beta}_w^{EL}, \hat{\pi}^{EL}, T(\hat{\beta}_w^{EL}))$.

2. *For* $k \geq 1$, *obtain* $(\tilde{y}^{k-1}, \tilde{w}^{k-1}, \tilde{z}^{k-1})$ *as the residuals from weighted least squares regression of* y, w, *and* z *on* x *using* π^{k-1} *as weights.*

Update β_w, π *and* $\overline{C}^k$*:*

$$\begin{aligned}\left(\beta_w^k, \pi^k, \overline{C}^k\right) = & \underset{\beta_w, \pi}{arg\ max} \quad T(\beta_w) \\ & s.t. \quad \sum_{i=1}^{N} \pi_i = 1 \\ & \sum_{i=1}^{N} \pi_i \left(\tilde{y}_i^{k-1} - \left(\tilde{w}_i^{k-1}\right)' \beta_w\right) \otimes \tilde{z}_i = 0 \\ & \sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} - \sum_{i=1}^{N} \ln \pi_i \leq \chi_1^2(1-\alpha)/2.\end{aligned}$$

4. Repeat steps 2-3 until convergence.

The primal problem illustrated in Algorithm 2 is carried on by the quasi-dual formulation. Our algorithm once again replaces one higher-dimension $(K + 2M + L)$ optimization problem with a sequence of lower-dimension $(K + L)$ optimization problems, eliminating $2M$ search parameters from the quasi-dual problem. Theorem 3 gives a formal statement of some properties for Algorithm 2 with an extra Assumption 3 on the existence of the full solution.

Assumption 3. *The finite-sample full EL problem for confidence intervals has a unique maximizer/minimizer.*

Theorem 3. *Under Assumptions 1, 2 and 3, Algorithm 2 has the following properties,*

1. In an optimal solution, the extra inequality constraint for (4) would be binding.

2. The full empirical likelihood CI estimate $(\hat{\beta}_w^{CI}, \hat{\beta}_x^{CI}, \hat{\pi}^{EL})$ *satisfies the FOC of the primal partitioned-out problem when partitioned out by* $\hat{\pi}^{EL}$*.*

Proof. See Appendix B. □

4 Monte Carlo Simulations

We now conduct several Monte Carlo simulations to demonstrate the improvement of the iterative algorithms for both point estimation and confidence intervals in finite samples.

4.1 Simulations for Point Estimation

Consider point estimation first. Suppose the successive structure generates the data – the dependent variable, y, is a linear function of two sets of regressors: exogenous regressors, x, and an endogenous

regressor, w, such that $y_i = w_i\beta_w + x_i'\beta_x + \varepsilon_i$, $E[\varepsilon_i|w_i] \neq 0$, $E[\varepsilon_i|x_i] = 0$. Moreover, there is a vector of instruments, z with $w_i = z_i'\theta_z + x_i'\theta_x + u_i$, $E[\varepsilon_i|z_i] = 0$, $E[u_i|x_i, z_i] = 0$.

We run a variety of different estimators and evaluate their behaviors. Here we give a brief description of each estimator. We carry out 2SLS in the usual way, and this analytical solution serves as the starting value for all EL-related methods. Full-EL solves the full problem, and the computation comes after the inner-outer loop iteration with Matlab codes provided by Evdokomiv and Kitamura (2011). One-step EL solves the partitioned problem as in Algorithm 1 but with the maximum number of iterations k hold at $k_{max} = 1$. One-step EL is asymptotically valid but needs not end up at a fixed point. Therefore, the just-identification β_x from equation (3) may not be well-defined in the finite sample. Iter-EL refers to the estimator provided in Algorithm 1 and iterates until convergence. To improve the performance of fixed-point iteration, we adopt the spectral algorithm proposed in La Cruz et al. (2006), with further details available in Appendix C.

Besides, we also try the estimator provided in Guggenberger and Hahn (2005), a k-step iterative method (mainly with k = 2), targeting a decent approximation to the full EL and recursively updating the unknown parameters by Newton's method, such that,

$$m(\beta) = \begin{pmatrix} \frac{1}{1+\lambda' g_i(\beta)} g_i(\beta) \\ \\ \frac{1}{1+\lambda' g_i(\beta)} \frac{\partial\left[\lambda' g_i(\beta)\right]}{\partial \beta} \end{pmatrix}, \quad m_n(\beta) = \sum_{i=1}^{N} m(\beta), \quad \beta^{k+1} = \beta^k - \left[\bigtriangledown m_n(\beta^k) \right]^{-1} m_n(\beta)$$

Then repeat this process until convergence or the number of iterations exceeds its maximum value.

Tables 1 and 2 show the simulated results for 1000 repetitions, using small sample sizes, $N = 50$ and $N = 100$, and with various unknowns and moment conditions. As expected, with more moment conditions, 2SLS may suffer from some finite-sample bias. The Newton-step approximation of EL estimator ends up in similar behavior as 2SLS, which confirms the conclusion in Guggenberger and Hahn (2005) – the higher-order asymptotic is not a good approximation for the finite-sample behavior for the two-step EL. As for the EL-type estimators, while convergence is usually not a problem, Iter-EL may give a slightly better convergence result. One-step EL, as a consistent estimator, perform good in all simulations except for the case where $N = 50$, with $M = 2, L = 5, K = 1$. Above all, Iter-EL is the one that gives the closest results towards Full-EL in all simulation settings.

Tables 1 and 2 may not clearly illustrate the improvement of Iter-EL over the One-step EL in

approximating Full-EL in a finite sample. Furthermore, Table 3 shows that, from the simulation results, when both converge and the optimality condition is assured at β^{EL}, Iter-EL gives an estimation of β_w that is always much closer to the Full-EL estimates, compared with the One-step EL, in both mean and max absolute difference. The difference between $\beta_w^{Iter-EL}$ and $\beta_w^{Full-EL}$ is negligible, and is mostly due to the error in numerical optimization, while is not between $\beta_w^{One-step\,EL}$ and $\beta_w^{Full-EL}$.[6] Therefore, Iter-EL would better approximate Full-EL in the finite sample, especially when we have a small sample size.

Moreover, our iterated method would greatly increase the computational speed when M and L are both large. Intuitively, our Algorithm 1 breaks a large nonlinear optimization problem into several low-dimensional numerical optimizations solved by a few iterations. Whether or not there is positive computational gain depends on the trade-off between solving one large problem and solving many small problems. Tables 4 and 5 illustrate two different situations.[7]

In Table 4, with M and L both small compared with N, solving the entire problem does not incur any computational difficulty, leading to a higher speed in Full-EL. However, the relation reverses when more exogenous variables and moment conditions join in. Table 5 illustrates the elapsed time when we have $M = 50$ and $L = 100$. While Iter-EL and Full-EL give almost identical results, the elapsed time becomes different. On average, Iter-EL finishes with 1.5941 seconds, accounting for around 15% of the elapsed time by Full-EL [8]. Also, Iter-EL has its maximum elapsed time lower than the minimum of Full-EL. Thus, as demonstrated in the simulation, Iter-EL could improve the computation time, which contributes to the literature of numerical computation on EL, summarized in, e.g., Priam (2021).

6. Table 3 is restricted to the case where both Full-EL and Iter-EL converge and full-EL's optimality, but in Table 1 or 2, the convergence is restricted to itself. So they are comparing different simulation samples.

7. The optimization is run by Apple M1 Max 2021, Ram 64GB, Matlab R2021b. Evdokomiv and Kitamura (2011) provide three optimization paths: fminunc with analytical gradient, fminsearch, and user-supplied fminsims. fminsearch is more stable, but usually takes a longer time. For a large sample ($N = 1000$) comparison, we stick to fminunc for every estimator. Besides, we do not compare the computational time for $N = 50$ and $N = 100$ here. When we have such a small sample size, we usually need to begin with different starting values and alternate among various paths, which would mask the actual underlying computational gains.

8. A reduction of several seconds alone may not be a significant achievement at first sight. However, consider the case where one needs to solve the EL problem many times, like, in bootstrap or EL serving as an inner loop (e.g., Conlon (2013)). A speed increase like this would be favorable.

	2SLS	Full-EL	One-step EL	Iter-EL	Newton Step
			$M = 2,\ L = 2,\ K = 1$		
Mean $\hat{\beta}_w$	-3.7351	-3.7412	-3.7410	-3.7412	-3.7342
Median $\hat{\beta}_w$	-3.7246	-3.7412	-3.7410	-3.7412	-3.7342
Std $\hat{\beta}_w$	0.1000	0.1058	0.1059	0.1058	0.1015
Mean Bias $\hat{\beta}_x$	0.0050	0.0087	0.0086	0.0087	0.0043
% of Converge [1]	-	100%	100%	100%	100%
% of Optimization [2]	-	99.9%	99.9 %	99.9 %	-
			$M = 2,\ L = 5,\ K = 1$		
Mean $\hat{\beta}_w$ if converge	-3.6398	-3.7341	-3.7459	-3.7325	-3.6395
Median $\hat{\beta}_w$ if converge	-3.6251	-3.7026	-3.7034	-3.7026	-3.6246
Std $\hat{\beta}_w$ if converge	0.2129	0.8061	0.9288	0.8005	0.2250
Mean Bias $\hat{\beta}_x$ if converge	0.0669	0.0085	0.0109	0.0068	0.0667
% of Converge	-	97.8%	98.9%	97.8%	99.1%
% of Optimization	-	98.3%	97.2%	96.7%	-

[1] Convergence in this case refer to (1) the algorithm stops at a reasonable solution; (2) it at least achieves the observed global minimum. As pointed out by Chen et al. (2008), in small sample, Empirical Likelihood may not have a solution, and if this happens, the $\hat{\beta}_w$ would blow up to an unreasonable level. Here with true $\beta_w = -3.7379$, we treat any estimated $|\hat{\beta}_w| > 12$ as not converge. Also, sometimes the primal function value $(1/N)\sum_{i=1}^{N} \ln \pi_i$ for full EL can be smaller than the fixed point iterative, and vise versa. If this happens, we then treat the method as if it does not converge to the global optimization. Table 2 follows this same notes for convergence.

[2] Optimization in this case refer to the satisfaction of both the FOC and SOC for both β_w and β_x, as illustrated in Appendix B in the finite sample. Note that for the partitioned algorithms, we still calculate the SOC back to the full problem.

Table 1: Point Estimation Simulations for N = 50.

	2SLS	Full-EL	One-step EL	Iter-EL	Newton Step
			$M = 5$, $L = 5$, $K = 1$		
Mean $\hat{\beta}_w$	-3.7091	-3.7543	-3.7544	-3.7543	-3.7102
Median $\hat{\beta}_w$	-3.7027	-3.7436	-3.7437	-3.7436	-3.7038
Std $\hat{\beta}_w$	0.1189	0.1341	0.1340	0.1341	0.1224
Mean Bias $\hat{\beta}_x$	0.0117	0.0073	0.0073	0.0073	0.0113
% of Converge	-	100%	100%	100%	100%
% of Optimization	-	98.9%	98.4%	98.7%	-
			$M = 5$, $L = 10$, $K = 1$		
Mean $\hat{\beta}_w$ if converge	-3.5812	-3.7675	-3.7668	-3.7677	-3.5845
Median $\hat{\beta}_w$ if converge	-3.5738	-3.7316	-3.7292	-3.7316	-3.5777
Std $\hat{\beta}_w$ if converge	0.1452	0.2361	0.2389	0.2365	0.1557
Mean Bias $\hat{\beta}_x$ if converge	0.0649	0.0128	0.0127	0.0128	0.0635
% of Converge	-	99.9 %	100%	99.9%	100%
% of Optimization	-	93.6%	87.9%	92.9%	-
			$M = 5$, $L = 15$, $K = 1$		
Mean $\hat{\beta}_w$ if converge	-3.4949	-3.8012	-3.8106	-3.8155	-3.4960
Median $\hat{\beta}_w$ if converge	-3.4910	-3.7589	-3.7619	-3.7619	-3.4910
Std $\hat{\beta}_w$ if converge	0.1372	0.3147	0.3429	0.4049	0.1583
Mean Bias $\hat{\beta}_x$ if converge	0.1014	0.0257	0.0296	0.0320	0.1008
% of Converge	-	98.9%	100%	99.9%	99%
% of Optimization	-	70.8%	69.7%	55.8%	-

Table 2: Point Estimation Simulations for N = 100.

	$N = 50$, $M = 2$, $K = 1$		$N = 100$, $M = 5$, $K = 1$		
	$L = 2$	$L = 5$	$L = 5$	$L = 10$	$L = 15$
$max\lvert\beta_w^{Iter-EL} - \beta_w^{Full-EL}\rvert$	0.000021	0.000084	0.000020	0.000050	0.029628
$mean\lvert\beta_w^{Iter-EL} - \beta_w^{Full-EL}\rvert$	0.000001	0.000005	0.000002	0.000002	0.000045
$max\lvert\beta_w^{One-step\,EL} - \beta_w^{Full-EL}\rvert$	0.076693	6.921621	0.082203	0.248717	0.472581
$mean\lvert\beta_w^{One-step\,EL} - \beta_w^{Full-EL}\rvert$	0.001353	0.044905	0.004032	0.015091	0.029959

Table 3: Comparison with Full-EL, if both converge, and necessary optimality conditions for full EL are hold.

		2SLS	Full-EL	One-step EL	Iter-EL
Mean $\hat{\beta}_w$		-3.7348	-3.7385	-3.7385	-3.7385
Median $\hat{\beta}_w$		-3.7346	-3.7388	-3.7388	-3.7388
Std $\hat{\beta}_w$		0.0369	0.0374	0.0374	0.0374
% of Converge		-	100%	100%	100%
% of Optimization		-	100%	100%	100%
Elapsed Time (s)	Mean	-	0.0236	0.0059	0.0683
	Max	-	0.2898	0.0259	0.2185
	Min	-	0.0119	0.0038	0.0319
Distance to $\beta_w^{Full-EL}$	Mean	-	-	0.000036	0.000002
	Max	-	-	0.000367	0.000113

Table 4: Point Estimation Simulations for N = 1000, M = 5, L = 5, K = 1.

		2SLS	Full-EL	One-step EL	Iter-EL
Mean $\hat{\beta}_w$		-3.3827	-3.7490	-3.7492	-3.7490
Median $\hat{\beta}_w$		-3.3804	-3.7423	-3.7429	-3.7422
Std $\hat{\beta}_w$		0.0508	0.0962	0.0972	0.0962
% of Converge		-	100%	100%	100%
% of Optimization		-	100%	100%	100%
Elapsed Time (s)	Mean	-	11.0658	0.1089	1.5941
	Max	-	17.0323	0.3242	3.7514
	Min	-	4.2369	0.0649	0.8450
Distance to $\beta_w^{Full-EL}$	Mean	-	-	0.008215	0.000010
	Max	-	-	0.039592	0.000109

Table 5: Point Estimation Simulations for N = 1000, M = 50, L = 100, K = 1.

4.2 Simulations for Confidence Interval

We follow the DGP in the point estimation for the confidence interval but with two different sets of error terms ε. We have $\varepsilon_i \sim N(0,1)$ as setting 1, and ε_i follows a demeaned Exponential distribution in setting 2. We calculate the confidence interval of β_w through several different methods. Table 6 summarizes the methods we adopted for comparison. We compare the EL-based confidence interval with the analytical CI constructed from asymptotic distribution and a simple bootstrap procedure.

Method for CI	Descriptions
Full-EL	Solve the constrained optimization as Eq. (4)
Iter-EL	Solve the constrained optimization following Algorithm 2.
One-step EL	Start with Algorithm 2, but stops after one iteration.
Analytical	After obtain the full-EL point estimation, calculate following $V_{\beta_w} = \frac{1}{N}\{(\sum_{i=1}^N \pi_i \frac{\partial g_i}{\partial \beta})'(\sum_{i=1}^N \pi_i g_i g_i')^{-1}(\sum_{i=1}^N \pi_i \frac{\partial g_i}{\partial \beta})\}^{-1}_{\beta_w}$ and CI is calculated by $[\beta_w^{EL} - t_{\alpha/2}V_{\beta_w}, \beta_w^{EL} + t_{\alpha/2}V_{\beta_w}]$.
Bootstrap	Draw a bootstrap sample and then use Full-EL to solve for the point estimation, and then construct CI as values between $\alpha/2$ and $1-\alpha/2$ quantile. # of bootstrap samples = 200.

Table 6: Summary of the method used in comparison for CIs.

Besides, empirical likelihood CI is also widely recommended because its Bartlett correctable property improves coverage even with limited sample sizes. Following the complicated derivation from Chen and Cui (2007), we calculate a bootstrap Bartlett correction for the partitioned confidence interval estimation to demonstrate its compatibility with our algorithm. In this simulation, we have β_w as a scalar, and hence, the bootstrap methods proposed in Chen and Cui (2007) is ready to use.[9] Specifically, now in Algorithm 2, we replace the original inequality by

$$\sum_{i=1}^N \ln \pi_i^{EL} - \sum_{i=1}^N \ln \pi_i \leq \chi_1^2(1-\alpha)\hat{\beta}_c/2$$

9. Chen and Cui (2007) also propose an exact theoretical derivation of the Bartlett correction with GMM-based estimators, but the theoretical formula is very cumbersome. Instead, they recommend to use an empirical Bartlett correction (EBC) with bootstrap estimation for $E[r(\beta^*)]$, which, nevertheless, decreases the convergence rate from n^{-2} to $n^{-3/2}$.

where $\hat{\beta}_c$ is calculated from the B-time bootstrap, with each procedure proceeding as

1. Generate the bootstrap sample $\{Y^b, X^b, W^b, Z^b\}$.

2. For this bootstrap sample, calculate the EL solution for the scalar $\hat{\beta}_w^b$, and $\sum_{i=1}^{n^b} \ln(\pi_i^b)$, from the partitioned problem.

3. Calculate the $\sum_{i=1}^{n^b} \ln(\pi_i^{b,EL})$ corresponding to $\hat{\beta}_w^{EL}$, and obtain $\hat{r}_b = 2 \times (\sum_{i=1}^{n^b} \ln(\pi_i^b) - \sum_{i=1}^{n^b} \ln(\pi_i^{b,EL}))$ [10].

Then $\hat{\beta}_c$ is inserted using $\hat{\beta}_c = B^{-1} \sum_{b=1}^{B} \hat{r}_b$. In this case, we have B = 250, as in the original Chen and Cui (2007).

We run the simulations for CI with M = 2, L = 2. Monte Carlo simulations in Tables 7, 8, 9, and 10 exhibit that, firstly, especially for setting 1, Iter-EL almost always generates identical results. The maximum difference between the $\overline{C}$ and $\underline{C}$ calculated is almost negligible, while discrepancy is still observable between One-step EL and Full-EL. Thus, Iter-EL provides a better approximation over One-step EL towards Full-EL. Secondly, applying the empirical Bartlett correction to the partitioned-out EL CI estimator would improve the coverage.

5 Extensions and Limitations

Since its seminal introduction by Owen (1988), Owen (1990), and Owen (1991), researchers have shown that empirical likelihood belongs to more general families of estimators. These include the generalized empirical likelihood (GEL) family introduced by Smith (1997) and the minimum discrepancy (MD) estimators introduced by Corcoran (1998), which have also been examined by Newey and Smith (2004). Kitamura (2006) provides a comprehensive review of minimum discrepancy estimation.

While there is no one-to-one correspondence between GEL and MD estimation, there is significant overlap between them. Newey and Smith (2004) show that any MD estimator based on Cressie-Read discrepancy has a GEL equivalent. The Cressie-Read family includes standard empirical likelihood, exponential tilting, and chi-square discrepancy, with the last being equivalent to continuously-updating GMM.

10. $\hat{\beta}_w$ denotes the EL solution from using the full sample.

Conceptually, our Algorithms 1 and 2 can be applied to GEL and MD estimators, with partitioning-out performed in the same way. However, extending the corresponding theorems in this paper to GEL and MD poses technical challenges. For example, the technical proofs of the theorems rely on positivity of the empirical likelihood weights to guarantee that the partitioning-out problem is well-posed in each iteration. Such positivity may not hold more generally for GEL and MD, since weights can be zero or negative (e.g., with chi-square discrepancy). However, we note that the partitioning-out problem may nevertheless be well-posed at each iteration in practice.

Additional extensions may also be possible in practice, including to Exponentially Tilted Empirical Likelihood (Schennach (2007)), adjusted empirical likelihood (Chen et al. (2008)), and empirical likelihood for partially linear models (Wang and Jing (2003)). Although, like the GEL and MD cases, the technical proofs may require substantial adjustment. We leave such analysis to future work.

6 Conclusion

We propose a Frisch-Waugh-Lovell-type theorem for empirical likelihood and develop an iterative partitioned empirical likelihood algorithm to implement the theorem in practice. Our algorithm can reduce the computational burden of empirical likelihood and improve numerical stability, with fast convergence to the full empirical likelihood estimate in practice. We further extend our iterative algorithm to simplify the computation of empirical likelihood confidence intervals. Monte Carlo simulations demonstrate our algorithms' superior behavior in small and large samples.

Finally, we note that a possible further research direction could be extending to the partially-linear model with the nuisance parameters estimated through a machine-learning-based first-stage estimation. Although, we leave such an extension to future research.

Acknowledgement

We thank seminar participants at The Ohio State University, and participants at the Asian Meeting of the Econometric Society 2022 China for their helpful comments and discussions. All errors are our own. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

		full-EL	Iter-EL	Analytical	Bootstrap	One-step EL	iter-EBC	One-step EBC
N = 50	Coverage	0.8440	0.8440	0.8580	0.8830	0.8450	0.9010	0.9020
	Length	0.3045	0.3045	0.2913	0.3525	0.3055	0.3798	0.3808
	Maximum diff, Iter vs Full					0.00016		
	Maximum diff, One-step vs Full					0.0758		
N = 100	Coverage	0.8790	0.8790	0.8780	0.8990	0.8800	0.9040	0.9050
	Length	0.2096	0.2096	0.2044	0.2224	0.2101	0.2291	0.2295
	Maximum diff, Iter vs Full					0.000008		
	Maximum diff, One-step vs Full					0.0111		

Table 7: **Setting 1** – Simulations for confidence interval of β_w, with nominal level $1 - \alpha = 90\%$.

		Full-EL	Iter-EL	Analytical	Bootstrap	One-step EL	iter-EBC	One-step EBC
N = 50	Coverage	0.9020	0.9020	0.9130	0.9380	0.9050	0.9500	0.9480
	Length	0.3713	0.3713	0.3492	0.4443	0.3727	0.4723	0.4736
	Maximum diff, Iter vs Full					0.00016		
	Maximum diff, One-step vs Full					0.0980		
N = 100	Coverage	0.9400	0.9400	0.9440	0.9480	0.9390	0.9550	0.9550
	Length	0.2527	0.2527	0.2443	0.2692	0.2534	0.2769	0.2776
	Maximum diff, Iter vs Full					0.000008		
	Maximum diff, One-step vs Full					0.0132		

Table 8: **Setting 1** – Simulations for confidence interval of β_w, with nominal level $1 - \alpha = 95\%$.

		Full-EL	Iter-EL	Analytical	Bootstrap	One-step EL	iter-EBC	One-step EBC
N = 50	Coverage	0.8420	0.8420	0.8600	0.8970	0.8420	0.9229	0.9218
	Length	0.2883	0.2883	0.2702	0.3403	0.2897	0.4181	0.4196
	Maximum diff, Iter vs Full					0.000160		
	Maximum diff, One-step vs Full					0.3203		
N = 100	Coverage	0.8780	0.8780	0.8950	0.8960	0.8770	0.9130	0.9140
	Length	0.2043	0.2043	0.1963	0.2177	0.2052	0.2417	0.2436
	Maximum diff, Iter vs Full					0.000799		
	Maximum diff, One-step vs Full					0.0390		

[1] In around 0.5% of simulations, extreme values occurs in estimating $\hat{\beta}_c$, which makes the EL estimation close to infeasible, and we just threw out those simulations. Similar results apply to the 95% confidence interval.

Table 9: **Setting 2** – Simulations for confidence interval of β_w, with nominal level $1 - \alpha = 90\%$.

		Full-EL	Iter-EL	Analytical	Bootstrap	One-step EL	iter-EBC	One-step EBC
N = 50	Coverage	0.9100	0.9100	0.9290	0.9460	0.9130	0.9618	0.9649
	Length	0.3559	0.3559	0.3240	0.4312	0.3589	0.5315	0.5462
	Maximum diff, Iter vs Full					0.00080		
	Maximum diff, One-step vs Full					0.5027		
N = 100	Coverage	0.9330	0.9330	0.9430	0.9480	0.9330	0.9560	0.9550
	Length	0.2479	0.2479	0.2347	0.2649	0.2494	0.2978	0.3036
	Maximum diff, Iter vs Full					0.00016		
	Maximum diff, One-step vs Full					0.0515		

Table 10: **Setting 2** – Simulations for confidence interval of β_w, with nominal level $1 - \alpha = 95\%$.

References

Aguirregabiria, V., and M. Marcoux. 2021. "Imposing Equilibrium Restrictions in the Estimation of Dynamic Discrete Games." *Quantitative Economics* 12 (4): 1223–1271.

Baum, C., M. Schaffer, and S. Stillman. 2007. "Enhanced Routines for Instrumental Variables/Generalized Method of Moments Estimation and Testing." *The Stata Journal* 7 (4): 465–506.

Chen, J., A. Variyath, and B. Abraham. 2008. "Adjusted Empirical Likelihood and its Properties." *Journal of Computational and Graphical Statistics* 17 (2): 426–443.

Chen, S. 1993. "On the Accuracy of Empirical Likelihood Confidence Regions for Linear Regression Model." *Annals of the Institute of Statistical Mathematics* 45 (4): 621–637.

Chen, S., and H. Cui. 2007. "On the Second-Order Properties of Empirical Likelihood with Moment Restrictions." *Journal of Econometrics* 141 (2): 492–516.

Chen, S., and P. Hall. 1993. "Smoothed Empirical Likelihood Confidence Intervals For Quantiles." *The Annals of Statistics* 21 (3): 1166–1181.

Chen, S., and I. Keilegom. 2009. "A Review on Empirical Likelihood Methods for Regression." *Test* 3:415–447.

Conlon, C. 2013. "The Empirical Likelihood MPEC Approach to Demand Estimation." Working Paper, https://chrisconlon.github.io/site/elmpecblp.pdf, September.

Corcoran, S. 1998. "Bartlett Adjustment of Empirical Discrepancy Statistics." *Biometrika* 85 (4): 967–972.

Dempster, A., N. Laird, and D. Rubin. 1977. "Maximum Likelihood from incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38.

Dong, L., and D. Giles. 2007. "An Empirical Likelihood Ratio Test for Normality." *Communications in Statistics - Simulation and Computation* 36 (1): 197–215.

Eryuruk, G. 2010. "The Time Series and Cross-Section Asymptotics of Empirical Likelihood Estimator in Dynamic Panel Data Models." Working Paper, http://allman.rhon.itam.mx/~gunce.eryuruk/Eryuruk_article.pdf, August.

Evdokomiv, K., and Y. Kitamura. 2011. "MATLAB/STATA codes for EL." Https://kitamura.sites.yale.edu/matlabstata-codes-el.

Frisch, R., and F. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1 (4): 387–401.

Giles, D. 1984. "Instrumental Variables Regressions Involving Seasonal Data." *Economics Letters* 14 (2): 339–343.

Guggenberger, P., and J. Hahn. 2005. "Finite Sample Properties of the Two-Step Empirical Likelihood Estimator." *Econometric Reviews* 24 (3): 247–263.

Hansen, B., and S. Lee. 2021. "Inference for Iterated GMM Under Misspecification." *Econometrica* 89 (3): 1419–1447.

Hansen, L., J. Heaton, and A. Yaron. 1996. "Finite-Sample Properties of Some Alternative GMM Estimators." *Journal of Business & Economic Statistics* 14 (3): 262–280.

Hausman, J., R. Lewis, K. Menzel, and W. Newey. 2011. "Properties of the CUE estimator and a modification with moments." *Journal of Econometrics* 165 (1): 45–57.

Kasahara, H., and K. Shimotsu. 2008. "Pseudo-Likelihood Estimation and Bootstrap Inference for Structural Discrete Markov Decision Models." *Journal of Econometrics* 146 (1): 92–106.

Kitamura, Y. 2006. *Empirical Likelihood Methods in Econometrics: Theory and Practice.* Discussion Paper 1569. Cowles Foundation.

La Cruz, W., J. Martínez, and M. Raydan. 2006. "Spectral Residual Method Without Gradient Information for Solving Large-scale Nonlinear System of Equations." *Mathematics of Computation* 75 (255): 1429–1448.

Lovell, M. 1963. "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis." *Journal of the American Statistical Association* 53 (304): 993–1010.

Newey, W., and R. Smith. 2004. "Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators." *Econometrica* 72 (1): 219–255.

Owen, A. 1988. "Empirical Likelihood Ratio Confidence Intervals for a Single Functional." *Biometrika* 75 (2): 237–249.

———. 1990. "Empirical Likelihood Confidence Regions." *The Annals of Statistics* 18 (1): 90–120.

Owen, A. 1991. "Empirical Likelihood for Linear Models." *The Annals of Statistics* 19 (4): 1725–1747.

Priam, R. 2021. "A brief survey of numerical procedures for empirical likelihood." Working Paper, https://hal.archives-ouvertes.fr/hal-03095014/document, January.

Qin, J., and J. Lawless. 1994. "Empirical Likelihood and General Estimating Equations." *The Annals of Statistics* 22 (1): 300–325.

Reich, G., and K. Judd. 2020. "Efficient Likelihood Ratio Confidence Intervals using Constrained Optimization." Working Paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3455484, December.

Schennach, S. 2007. "Point Estimation with Exponentially Tilted Empirical Likelihood." *The Annals of Statistics* 35 (2): 634–672.

Smith, R. 1997. "Alternative Semiparametric Likelihood Approaches to Generalized Method of Moments Estimation." *Economic Journal* 107:503–519.

Wang, Q.-H., and B.-Y. Jing. 2003. "Empirical Likelihood for Partial Linear Models." *Annals of the Institute of Statistical Mathematics* 55 (3): 585–595.

White, H. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50 (1): 1–25.

Zhang, J., S. Feng, G. Li, and H. Lian. 2011. "Empirical likelihood inference for partially linear panel data models with fixed effects." *Economics Letters* 113 (2): 165–167.

Technical Appendices

A. Duality Results

A.1 Duality for the EL Estimation Problem

The empirical likelihood problem takes the form

$$\begin{aligned}\left(\hat{\beta}^{EL}, \hat{\pi}^{EL}\right) = \arg\max_{\beta,\pi} \quad & \frac{1}{N}\sum_{i=1}^{N} \ln \pi_i \\ s.t. \quad & \sum_{i=1}^{N} \pi_i = 1 \\ & \sum_{i=1}^{N} \pi_i g_i(\beta) = 0\end{aligned}$$

An alternative formulation that uses Lagrange multipliers is as follows:

$$\begin{aligned}\left(\hat{\beta}^{EL}, \hat{\lambda}^{EL}\right) = \arg\max_{\beta,\lambda} \quad & \frac{1}{N}\sum_{i=1}^{N} \ln \frac{1}{1+\lambda' g_i(\beta)} \\ s.t. \quad & \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+\lambda' g_i(\beta)} g_i(\beta) = 0\end{aligned}$$

and we have $\hat{\pi}^{EL} = 1/(N(1+\hat{\lambda}^{EL'} g_i(\hat{\beta}^{EL}))$. In our full empirical likelihood problem, $\beta = (\beta_w, \beta_x)$ and $g_i = (y_i - w_i'\beta_w - x_i'\beta_x) \otimes (z_i', x_i')'$. In the partitioned problem, $\beta = \beta_w$ and $g_i = \tilde{g}_i = (\tilde{y}_i - \tilde{w}_i'\beta_w) \otimes \tilde{z}_i$.

A.2 Quasi-Duality for the EL Confidence Interval Problem

The general upper bound problem takes the form

$$\begin{aligned}\overline{C} = \max_{\beta,\pi} \quad & T(\beta) \\ s.t. \quad & \sum_{i=1}^{N} \pi_i = 1 \\ & \sum_{i=1}^{N} \pi_i g_i(\beta) = 0 \\ & \sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} - \sum_{i=1}^{N} \ln \pi_i \leq \chi_1^2(1-\alpha)/2\end{aligned}$$

Further analysis demonstrates that the inequality constraint must be binding for the optimal solution so that the corresponding Lagrange multiplier $\gamma \neq 0$. Following Qin and Lawless (1994), we solve for the expression of π_i, and insert it back to obtain the alternative formulation with Lagrange multiplier. Suppose we are interested in β_j's confidence interval, then the optimization problem becomes,

$$\begin{aligned}
\overline{C} \;= \min_{\beta,\lambda,\gamma} \quad & -T(\beta) \\
s.t. \quad & \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1-(1/\gamma)\lambda' g_i(\beta)} g_i(\beta) = 0 \\
& \sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} - \sum_{i=1}^{N} \ln \frac{1}{N} \frac{1}{1-(1/\gamma)\lambda' g_i(\beta)} = \chi_1^2(1-\alpha)/2 \\
& \sum_{i=1}^{N} \left[\frac{1}{N(1-(1/\gamma)\lambda' g_i(\beta))} \left(\frac{\partial g_i(\beta)}{\partial \beta_j} \right)' \lambda \right] - \frac{\partial T(\beta)}{\partial \beta_j} = 0
\end{aligned} \tag{5}$$

The last equality is to pin down the relationship between λ and γ. Otherwise, λ and γ would not be identified separately. This alternative representation generates the same set of FOC as the primal problem. To show this, consider the Lagrange representation,

$$\begin{aligned}
\mathcal{L} = & -T(\beta) + \kappa_1 \left[\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1-(1/\gamma)\lambda' g_i(\beta)} g_i(\beta) \right] + \kappa_2 \left[\sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} + \sum_{i=1}^{N} \ln N(1-(1/\gamma)\lambda' g_i(\beta)) \right. \\
& \left. - \chi_1^2(1-\alpha)/2 \right] + \kappa_3 \left[\sum_{i=1}^{N} \left[\frac{1}{N(1-(1/\gamma)\lambda' g_i(\beta))} \left(\frac{\partial g_i(\beta)}{\partial \beta_j} \right)' \lambda \right] - \frac{\partial T(\beta)}{\partial \beta_j} \right]
\end{aligned}$$

With the FOC, $\partial \mathcal{L}/\partial \lambda = 0$ gives $\kappa_1 = 0$ and $\kappa_3 = 0$, leading to

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = -\frac{\partial T(\beta)}{\partial \beta_j} + \kappa_2 \left[\sum_{i=1}^{N} \frac{1}{N(1-(1/\gamma)\lambda' g_i(\beta))} \left(\frac{\partial g_i(\beta)}{\partial \beta_j} \right)' \frac{\lambda}{\gamma} \right] = 0$$

Compare it with the last equality constraint, we have $\kappa_2 = \gamma$. Therefore the extra $1/\gamma$ in $\partial \mathcal{L}/\partial \beta$ is canceled out, and as shown in, we retrieve the FOC as in the primal problem.

Moreover, a further investigation indicates that, for a given β, there exists some π such that (β, π) is feasible for the above optimization problem if and only if $\sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} - \sum_{i=1}^{N} \ln \pi_i^{\max}(\beta) \leq$

$\chi^2_1(1-\alpha)/2$ where

$$\begin{aligned} \pi^{\max}(\beta) \; = \arg\max_{\pi} \quad & \frac{1}{N}\sum_{i=1}^{N} \ln \pi_i \\ s.t. \quad & \sum_{i=1}^{N} \pi_i = 1 \\ & \sum_{i=1}^{N} \pi_i g_i(\beta) = 0. \end{aligned} \tag{6}$$

This sub-problem in (6) has a convenient dual representation. One can instead solve for $\lambda(\beta) \in \mathbb{R}^{\dim(g_i)}$ such that

$$\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+\tilde{\lambda}' g_i(\beta)} g_i(\beta) = 0$$

and then compute $\pi_i^{\max}(\beta) = 1/(N(1+\lambda(\beta)' g_i(\beta))$. This leads to the quasi-dual confidence interval bound problem.

$$\begin{aligned} \overline{C} \; = \max_{\beta,\tilde{\lambda}} \quad & T(\beta) \\ s.t. \quad & \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+\tilde{\lambda}' g_i(\beta)} g_i(\beta) = 0 \\ & \sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} - \sum_{i=1}^{N} \ln\left(\frac{1}{N\left(1+\tilde{\lambda}' g_i(\beta)\right)}\right) \leq \chi^2_1(1-\alpha)/2. \end{aligned} \tag{7}$$

This "quasi-dual" problem is not the standard dual problem to the full confidence interval bound problem. Nevertheless, it is a convenient simplification to a general-purpose constrained optimization solver since it has a lower dimension than the full problem. Unlike the full problem, the size of the quasi-dual problem does not grow with the sample size.

In addition, the alternative representation in (5) and in (7) would generate exact same solution. Consider an optimal solution (β, λ, γ) for (5), and the Lagrange representation for (7) is

$$\mathcal{L} = -T(\beta) + \psi_1\left[\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+\tilde{\lambda}' g_i(\beta)} g_i(\beta)\right] + \psi_2\left[\sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} + \sum_{i=1}^{N} \ln N(1+\tilde{\lambda}' g_i(\beta)) - \chi^2_1(1-\alpha)/2\right]$$

It is straightforward to show that (β, λ, γ) in (5) satisfies the FOC for (7) if $\psi_1 = 0, \psi_2 = -\gamma$, and $\tilde{\lambda} = \lambda/\gamma$. Then need to show that (β, λ, γ) is indeed an optimal solution. Suppose not, and instead, the optimal solution for (7) is $\bar{\lambda}, \bar{\beta}$, where $\bar{\lambda} \neq \lambda/\gamma$, then we can construct a new γ by setting $\gamma = -\psi_2$ and a new λ by setting $\lambda = \bar{\lambda}/\gamma$. Combined with $\kappa_1 = \gamma$, they are feasible for

(5), but generate a smaller value for the objective function. This is contradicted to the assumption that (β, λ, γ) is the solution for (5). Similar logic applies from the opposite direction. Thus, the optimization problem in (5) and (7) gives the same results for the CI bound estimation.

Our argument supporting the quasi-dual formulation operates through the constraints of the full problem. In particular, we do not require that the initial problem be a maximization problem to continue. Our results still carry through when the initial problem is a constrained minimization. Thus, a similar alternative formulation is valid for computing $\underline{C}$.

$$\begin{aligned}
\underline{C} \;=\; & \min_{\beta,\lambda,\gamma} \quad T(\beta) \\
& s.t. \quad \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1-(1/\gamma)\lambda' g_i(\beta)} g_i(\beta) = 0 \\
& \qquad \sum_{i=1}^{N} \ln \hat{\pi}_i^{EL} - \sum_{i=1}^{N} \ln \frac{1}{N} \frac{1}{1-(1/\gamma)\lambda' g_i(\beta)} = \chi_1^2(1-\alpha)/2 \\
& \qquad \sum_{i=1}^{N} \left[\frac{1}{N(1-(1/\gamma)\lambda' g_i(\beta))} \left(\frac{\partial g_i(\beta)}{\partial \beta_j} \right)' \lambda \right] + \frac{\partial T(\beta)}{\partial \beta_j} = 0
\end{aligned}$$

B. Proofs Omitted in the Main Text

Before getting into the main proofs, we first introduce some matrix notations. Let Y, W, X, and Z be defined as the data matrix for dependent variable, endogenous regressors, exogenous regressors and instruments. Let $\Pi = \text{diag}(\pi)$, with $\Pi^{EL} = \text{diag}(\pi^{EL})$, $\hat{\Pi} = \text{diag}(\hat{\pi})$, etc.

Since the empirical likelihood problem takes ln of choice probabilities, we can restrict attention to π such that $\pi_i > 0$ for i=1, ..., N [11]. This allows for a useful invertibility result.

Lemma 1. *Under Assumption 1, $X'\Pi X = \sum_{i=1}^{N} \pi_i x_i x_i'$ is invertible when $\pi_i > 0$ for all i=1,...,N.*

Proof. First, note that $x_i x_i'$ is positive semi-definite, so that $v' x_i x_i' v \geq 0$ for all $v \in \mathbb{R}^M \backslash \{0\}$. For any $v \neq 0$, non-singularity of $N^{-1}\sum_{i=1}^{N} x_i x_i'$ requires $\sum_{i=1}^{N} v' x_i x_i' v > 0$. This implies that there is at least one j such that $v' x_j x_j' v > 0$. We have $\pi_i > 0$ for all i=1,..., N by assumption, so $\pi_j > 0$

11. As for the family of the generalized empirical likelihood method, it is possible to have some $\pi_i < 0$. For instance, the Continuous Updating Estimator proposed in Hansen et al. (1996) does not put any positive restrictions on weights. Optimization with negative weights is out of the range of this research since it would make the iterative process using WLS problematic.

and $v'\left(\sum_{i=1}^{N}\pi_i x_i x_i'\right)v = \sum_{i=1}^{N}\pi_i\left(v'x_i x_i' v\right) > 0$. Thus, $\sum_{i=1}^{N}\pi_i x_i x_i'$ is positive definite and must be non-singular. And since it is square, it is invertible. □

The invertibility result in Lemma 1 allows us to define the projection matrices $P_\pi \equiv X\left(X'\Pi X\right)^{-1}X'\Pi$ and $M_\pi \equiv I - P_\pi$. Residuals from weighted regression are $\tilde{Y}_\pi \equiv M_\pi Y$, $\tilde{W}_\pi = M_\pi W$, and $\tilde{Z}_\pi = M_\pi Z$.

Remark 1. $P_\pi X = X$ and $M_\pi X = 0$, and therefore $M_{\pi^1}P_{\pi^2} = 0$ for arbitrary π^1 and π^2 (with all elements positive). Also, $M_\pi'\Pi X = 0$ and therefore $M_{\pi^1}'\Pi_1 P_{\pi^2} = 0$ for arbitrary π^1 and π^2. $P_\pi'\Pi M_\pi = 0$, for arbitrary π.

Proof of these simple equations are straightforward, from the definition of each matrix. For instance,

$$\begin{aligned}
P_\pi'\Pi M_\pi &= \Pi X(X'\Pi X)^{-1}X'\Pi(I - X(X'\Pi X)^{-1}X'\Pi) \\
&= \Pi X(X'\Pi X)^{-1}X'\Pi - \Pi X(X'\Pi X)^{-1}X'\Pi X(X'\Pi X)^{-1}X'\Pi \\
&= \Pi X(X'\Pi X)^{-1}X'\Pi - \Pi X(X'\Pi X)^{-1}X'\Pi = 0
\end{aligned}$$

Remark 2. The moment constraints for the full empirical likelihood problem in matrix form are $[Z, X]'\,\Pi\,(Y - W\beta_w - X\beta_x) = 0$.

Remark 3. Suppose that partition is performed with weights $\check{\pi}$. Then the moment conditions for the partitioned-out empirical likelihood problem in matrix form are $\tilde{Z}_{\check{\pi}}'\Pi\left(\tilde{Y}_{\check{\pi}} - \tilde{W}_{\check{\pi}}\beta_w\right) = 0$.

With these results in place, we can now prove Theorem 1.

B.1 Proof of Theorem 1

First, we note that Theorem 1.1 can be re-stated as follows: $[Z, X]'\,\Pi\,(Y - W\beta_w - X\beta_x) = 0$ implies $\tilde{Z}_\pi'\Pi\left(\tilde{Y}_\pi - \tilde{W}_\pi\beta_w\right) = 0$. This is what we will prove.

Suppose that $[Z, X]'\,\Pi\,(Y - W\beta_w - X\beta_x) = 0$. The full empirical likelihood moment conditions must be satisfied at (β_w, β_x, π), so that $X'\Pi\,(Y - W\beta_w - X\beta_x) = 0$ and

$$\begin{aligned}
0 \;=\; & Z'\Pi\,(Y - W\beta_w - X\beta_x) = \left(Z'M_\pi' + Z'P_\pi'\right)\Pi\,(Y - W\beta_w - X\beta_x) \\
=\; & Z'M_\pi'\Pi\,(Y - W\beta_w - X\beta_x) = Z'M_\pi'\Pi\left(\tilde{Y}_\pi - \tilde{W}_\pi\beta_w + P_\pi Y - P_\pi W\beta_w - X\beta_x\right) \\
=\; & Z'M_\pi'\Pi\left(\tilde{Y}_\pi - \tilde{W}_\pi\beta_w\right) = \tilde{Z}_\pi'\Pi\left(\tilde{Y}_\pi - \tilde{W}_\pi\beta_w\right).
\end{aligned}$$

The second, fourth, and sixth equalities arises from the definitions of $M_\pi, P_\pi, \tilde{Y}_\pi, \tilde{W}_\pi$, and $\tilde{Z}_\pi$. The third equality arises from the form of P_π and the requirement that $X'(Y - W\beta_w - X\beta_x) = 0$. The fifth equality arises because $M_\pi'\Pi X = 0$ and $M_\pi'\Pi P_\pi = 0$ (Remark 1)

Theorem 1.2 can be restated as, given the fixed point $(\check{\beta}_w, \check{\pi})$ out of Algorithm 1, the β_x, constructed following $\beta_x = (X'\check{\Pi}X)^{-1}X'\check{\Pi}(Y - W\check{\beta}_w)$ would also have the property such that, $Z'\check{\Pi}(Y - W\check{\beta}_w - X\beta_x) = 0$, since,

$$\begin{aligned} Z'\check{\Pi}(Y - W\check{\beta}_w - X\beta_X) &= Z'\check{\Pi}(Y - W\check{\beta}_w - P_{\check{\pi}}(Y - W\check{\beta}_w)) \\ &= (Z'M_{\check{\pi}}' + Z'P_{\check{\pi}}')\check{\Pi}M_{\check{\pi}}(Y - W\check{\beta}_w) \\ &= Z'M_{\check{\pi}}'\check{\Pi}M_{\check{\pi}}(Y - W\check{\beta}_w) = \tilde{Z}_{\check{\pi}}'\check{\Pi}(\tilde{Y}_{\check{\pi}} - \tilde{W}_{\check{\pi}}\check{\beta}_w) = 0 \end{aligned}$$

The first equality holds since $X\beta_x = P_{\check{\pi}}(Y - W\check{\beta}_w)$, which comes from the expression of β_x and the definition of P_π directly. The third equality comes from Remart 1 where $P_\pi'\Pi M_\pi = 0$ for arbitrarily non-negative π. The last equation equals 0 since $\check{\pi}$ is a fixed point for the partitioned problem.

Theorem 1.2 shows that Algorithm 1 gives a well-defined solution to the full problem in the finite sample since one can construct a feasible β_x from the fixed point containing only parts of the variables [12].

Before formally going to the asymptotic proof for Theorem 1.3, a few notation needs to be clarified. If converges, Algorithm 1 would end up in a fixed point $(\check{\beta}_w, \check{\pi})$ in the primal problem, or a dual problem with $(\check{\beta}_w, \check{\lambda}_z)$, where λ_z is the Lagrange multiplier corresponding to the excluded instruments. We do not restrict the fixed point to be unique for the partitioned problem. Among those fixed points, we are interested in the one with the highest primal objective function value, denoted as (β_w^P, λ_z^P).

The primal objective function is $Q_N(\beta, \lambda) = \sum_{i=1}^N \ln \pi_i(\beta, \lambda) = \sum_{i=1}^N N^{-1} \ln(1/[1 + \lambda' g_i(\beta)])$. The idea here is to show that $\hat{\beta}_w^{EL}$ and $\hat{\lambda}_z^{EL}$ are asymptotically the β_w^P and λ_z^P through

(1) If $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ satisfy the FOC and SOC for the problem partitioned-out by $\hat{\pi}^{EL}$, then it is a fixed point of Algorithm 1.

(2) If $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ is a fixed point of the partitioned problem, and by Assumption 2 the full

12. The One-step partitioned EL would not have this property, so one may not be able to retrieve a β_x satisfying the moment restrictions from the full problem.

EL problem has a unique solution, then there does not exist other fixed point $(\check{\beta}_w, \check{\lambda})$ that would have $Q_N(\check{\beta}_w, \check{\lambda}_z) \geq Q_N(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$. Therefore, $\hat{\beta}_w^{EL}$ and $\hat{\lambda}_z^{EL}$ are the fixed points that generates the largest objective function value to the partitioned problem, or the β_w^P and λ_Z^P.

(3) As $N \to \infty$, asymptotically (1) would hold almost surely.

(1) comes directly from the definition of the fixed point. A fixed point in Algorithm 1 refers to the case where the weights derived from the optimal solution equals the weights used for partialling-out. So that $\hat{\pi}^{EL}$ is a fixed point means $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ are optimal solutions for the problem partialled-out by $\hat{\pi}^{EL}$.

Proof of (2). By contradiction. suppose there exists another fixed point with $\check{\beta}_w \neq \hat{\beta}_w^{EL}$ or $\check{\lambda}_z \neq \hat{\lambda}_z^{EL}$, but with $Q_N(\check{\beta}_w, \check{\lambda}_z) \geq Q_N(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$, then by the Theorem 1.2, we can construct a $\check{\beta}_x$ from just-identification and then $(\check{\beta}_w, \check{\beta}_x)$ is a feasible solution to the full EL with a larger primal function value. This is contradicted to Assumption 2 that $(\hat{\beta}^{EL}, \hat{\lambda}^{EL})$ is the unique solution to full EL. □

To prove (3), we need more assumptions.

Assumption 4. *At (β^*, λ^*), (i) $E[\frac{\partial \tilde{g}_i}{\partial \beta_w}]$ and $E[\frac{\partial g_i}{\partial \beta_w}]$ are non-singular, (ii) $E[\tilde{g}_i \tilde{g}_i']$ and $E[g_i g_i']$ are positive definite.*

Assumption 5. *(i) $\partial g_i(\beta^*)/\partial \beta$ is continuous in a neighborhood of the true value β^*; and $||\partial g_i(\beta)/\partial \beta||$ and $||g_i(\beta)||$ are bounded by some integrable function G in this neighborhood; (ii) $\partial^2 g(\beta)/\partial \beta \partial \beta'$ is continuous in β in a neighborhood of β^*, and $||\partial^2 g(\beta)/\partial \beta \partial \beta'||$ can be bounded by some integrable function G in this neighborhood.*

Assumption 4 contains the regularity conditions commonly presenting in empirical likelihood literature, and the conditions are for both the original conditions and the partitioned-out moment conditions. Asymptotically it would lead to SOC for optimization. Assumption 5 are standard and sufficient in proving the root-N consistency of the full EL solution, as in Qin and Lawless (1994).

Proof of (3). Here, we want to illustrate that, the full EL solution $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ satisfy the FOC of the partitioned problem in finite sample, and asymptotically for the SOC as $N \to \infty$ and $\hat{\lambda}_z^{EL} \to 0$. Dealing with SOC in constrained optimization is prohibitively cumbersome. Instead, we follow Kitamura (2006) and rewrite the constrained optimization as an unconstrained max-min saddle point problem,

$$\hat{\beta}_w^{EL} = \arg \max_{\beta_w \in \mathcal{B}_w} \min_{\lambda_z \in \mathbb{R}^L} -\frac{1}{N} \sum_{i=1}^{N} \ln(1 + \lambda_z' \tilde{g}_i)$$

where the partitioned-out moment condition is $\tilde{g}_i = (\tilde{y}_i - \tilde{w}_i'\beta_w) \otimes \tilde{z}_i$. The inner minimization problem regarding λ_z is convex so the FOC would be enough to justify a minimizer.

Denote $\mathscr{L} = \min_{\lambda_z \in \mathbb{R}^L} -N^{-1} \sum_{i=1}^N \ln(1 + \lambda_z' \tilde{g}_i)$. For the inner loop, given a fixed β, λ_z needs to satisfy the FOC,

$$\frac{\partial \mathscr{L}}{\partial \lambda_z} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{1 + \lambda_z' \tilde{g}_i(\beta)} \tilde{g}_i(\beta) = 0 \tag{8}$$

Or in other words, the optimal solution $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ for the full problem needs to be dual feasible for the partialling-out problem. By Theorem 1.1 and the FOC for the full problem

$$-\frac{1}{N} \sum_{i=1}^N \frac{1}{1 + \hat{\lambda}_z^{EL\prime} g_{z,i}(\hat{\beta}^{EL}) + \hat{\lambda}_x^{EL\prime} g_{x,i}(\hat{\beta}^{EL})} \tilde{g}_i(\hat{\beta}_w^{EL}) = 0$$

where $\beta = (\beta_w', \beta_x')'$, and the moment conditions are partitioned into two blocks, with $g_{z,i} = (y_i - w_i'\beta_w - x_i'\beta_x) \otimes z_i$, $g_{x,i} = (y_i - w_i'\beta_w - x_i'\beta_x) \otimes x_i$. From now on, we temporarily omit the superscript EL and hat to simplify the notation. A sufficient condition for (8) to hold is, $\forall i \in 1, \cdots, N$, $\lambda_z' g_{z,i}(\beta) + \lambda_x' g_{x,i}(\beta) = \lambda_z' \tilde{g}_i(\beta_w)$. To have this, we need,

$$\begin{aligned}
&\lambda_z'[(\tilde{y}_i - \tilde{w}_i'\beta_w) \otimes \tilde{z}_i] \\
&= \lambda_z'[(y_i - x_i'\hat{\alpha}_y - (w_i' - x_i'\hat{\alpha}_w)\beta_w) \otimes (z_i - \hat{\alpha}_z' x_i)] \\
&= \lambda_z'[(y_i - w_i'\beta_w - (x_i'(X'\Pi X)^{-1}X'\Pi Y - x_i'(X'\Pi X)^{-1}X'\Pi W\beta_w)) \otimes (z_i - \hat{\alpha}_z' x_i)] \\
&= \lambda_z'[(y_i - w_i'\beta_w - (x_i'(X'\Pi X)^{-1}X'\Pi(X\beta_x + \hat{\varepsilon}))) \otimes (z_i - \hat{\alpha}_z' x_i)] \\
&= \lambda_z'[(y_i - w_i'\beta_w - x_i'\beta_x) \otimes z_i] - \lambda_z'[(y_i - w_i'\beta_w - x_i'\beta_x) \otimes \hat{\alpha}_z' x_i] \\
&= \lambda_z' g_z - \lambda_z'[(y_i - w_i'\beta_w - x_i'\beta_x) \otimes \hat{\alpha}_z' x_i]
\end{aligned}$$

$\hat{\alpha}_y$, $\hat{\alpha}_w$ and $\hat{\alpha}_z$ are the partialling-out coefficients from WLS with weights $\hat{\pi}_i^{EL}$. Particularly, $\hat{\alpha}_z$ is a $M \times L$ matrix, and $\hat{\alpha}_w$ is a $M \times K$ matrix, stacking the estimated coefficients from WLS together. Compared with $\lambda_z' g_{z,i} + \lambda_x' g_{x,i}$, a sufficient condition for us to have the equality hold is to have

$$\lambda_z'[(y_i - w_i'\beta_w - x_i'\beta_x) \otimes \hat{\alpha}_z' x_i] = \lambda_x' g_{x,i}, \quad \text{or,} \quad \lambda_x = -\hat{\alpha}_z \lambda_z \tag{9}$$

This relationship between λ_x and λ_z is a direct result from FOC regarding β_x of the full max-min

saddle point problem,

$$(\beta_w, \beta_x) = \underset{\beta_w \in \mathcal{B}_w, \beta_x \in \mathcal{B}_x}{\arg\max} \min_{\lambda_z \in \mathbb{R}^L, \lambda_x \in \mathbb{R}^M} -\frac{1}{N}\sum_{i=1}^{N} \ln(1 + \lambda_x' g_{x,i} + \lambda_z' g_{z,i})$$

Denote $\mathcal{H} = \min_{\lambda_z \in \mathbb{R}^L, \lambda_x \in \mathbb{R}^M} -N^{-1}\sum_{i=1}^{N} \ln(1 + \lambda_x' g_{x,i} + \lambda_z' g_{z,i})$. By taking total derivative with respect to β_x, and use the Envelope Theorem, we have $d\mathcal{H}/d\beta_x = \partial\mathcal{H}/\partial\beta_x + (\partial\lambda/\partial\beta_x)'\partial\mathcal{H}/\partial\lambda$, and

$$\begin{aligned}
&\frac{d\mathcal{H}}{d\beta_x} = \frac{\partial\mathcal{H}}{\partial\beta_x} = -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' g_{z,i} + \lambda_x' g_{x,i}} [x_i z_i' \lambda_z + x_i x_i' \lambda_x] = 0 \\
&\Rightarrow -\Big(\sum_{i=1}^{N} \pi_i x_i z_i'\Big)\lambda_z = \Big(\sum_{i=1}^{N} \pi_i x_i x_i'\Big)\lambda_x \;\Rightarrow \lambda_x = -\hat{\alpha}_z \lambda_z
\end{aligned}$$

by $\hat{\alpha}_z = (X'\Pi X)^{-1} X'\Pi Z = \big(\sum_{i=1}^{N} \pi_i x_i x_i'\big)^{-1}\big(\sum_{i=1}^{N} \pi_i x_i z_i'\big)$. With this, we have equation (8) hold and the FOC of λ_z given β_w is satisfied by $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ in the finite sample.

Next, let us come to the FOC for β_w in the partitioned problem. First, we write the total derivative as,

$$\frac{d\mathcal{L}}{d\beta_w} = \frac{\partial\mathcal{L}}{\partial\beta_w} + \left(\frac{\partial\lambda_z}{\partial\beta_w}\right)' \frac{\partial\mathcal{L}}{\partial\lambda_z} = \frac{\partial\mathcal{L}}{\partial\beta_w} = -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} \left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)' \lambda_z$$

By the result from the FOC of λ, we have $\partial\mathcal{L}/\partial\lambda_z$ evaluated at $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ equal to 0. Further calculation shows,

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\beta_w} &= -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} \left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)' \lambda_z = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} \tilde{w}_i \tilde{z}_i' \lambda_z = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} w_i \tilde{z}_i' \lambda_z \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} w_i (z_i - \hat{\alpha}_z' x_i)' \lambda_z = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} [w_i z_i' \lambda_z - w_i x_i' \hat{\alpha}_z \lambda_z]
\end{aligned} \tag{10}$$

This comes from $W'\Pi\tilde{Z} = (W'P_\pi' + W'M_\pi')\Pi\tilde{Z} = W'M_\pi'\Pi\tilde{Z} = \tilde{W}'\Pi\tilde{Z}$ and $M_\pi'\Pi P_\pi = 0$. And since $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ satisfy the FOC for the full problem, we also have,

$$\begin{aligned}
\frac{d\mathcal{H}}{d\beta_w} &= \frac{\partial\mathcal{H}}{\partial\beta_w} = -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' g_{z,i} + \lambda_x' g_{x,i}} \left[\left(\frac{\partial g_{z,i}}{\partial\beta_w}\right)' \lambda_z + \left(\frac{\partial g_{x,i}}{\partial\beta_w}\right)' \lambda_x\right] = 0 \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' g_{z,i} + \lambda_x' g_{x,i}} [w_i z_i' \lambda_z + w_i x_i' \lambda_x] = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} [w_i z_i' \lambda_z - w_i x_i' \hat{\alpha}_z \lambda_z].
\end{aligned} \tag{11}$$

The last equality follows from equation (9) and the proof of FOC for $\hat{\lambda}_z^{EL}$ given $\hat{\beta}_w^{EL}$. Therefore, equation (10) = equation (11) = 0, and the full EL solution $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ satisfies the FOC for the

partialling-out problem when partialling-out by $\hat{\pi}_i^{EL}$.

And we still require SOC for an optimal solution. Denote the i-jth element in the vector-to-vector derivative matrix $\partial y/\partial x$ as $\partial y_i/\partial x_j$, for two column vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$. The total derivative with respect to β_w is [13]

$$\frac{d^2\mathcal{L}}{d\beta_w^2} = \frac{\partial^2\mathcal{L}}{\partial\beta_w^2} + \left(\frac{\partial^2\mathcal{L}}{\partial\lambda_z\partial\beta_w}\right)'\frac{\partial\lambda_z}{\partial\beta_w} + \left(\frac{\partial\mathcal{L}}{\partial\lambda_z}\right)'\frac{\partial^2\lambda_z}{\partial\beta_w^2} = \frac{\partial^2\mathcal{L}}{\partial\beta_w^2} + \left(\frac{\partial^2\mathcal{L}}{\partial\lambda_z\partial\beta_w}\right)'\frac{\partial\lambda_z}{\partial\beta_w}$$

Then calculate each items in the SOC as

$$\frac{\partial^2\mathcal{L}}{\partial\beta_w^2} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\left[\left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)'\lambda_z\lambda_z'\left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)\right] - \frac{1}{1+\lambda_z'\tilde{g}_i}\frac{\partial}{\partial\beta_w}\left[\left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)'\lambda_z\right]$$
$$\frac{\partial^2\mathcal{L}}{\partial\lambda_z\partial\beta_w} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\lambda_z\tilde{g}_i'\left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right) - \frac{1}{1+\lambda_z'\tilde{g}_i}\left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)$$

Since the partitioned moment $\tilde{g}_i$ is a linear function of β_w, the second term of $\partial^2\mathcal{L}/\partial\beta_w^2 = 0$. And using Implicit Function Theorem regarding the FOC of λ_z towards $h_1 = N^{-1}\sum_{i=1}^{N}(1+\lambda_z'\tilde{g}_i)^{-1}\tilde{g}_i = 0$ gives $\partial h_1/\partial\beta_w + (\partial h_1/\partial\lambda_z)\times(\partial\lambda_z/\partial\beta_w) = 0$. And by inserting equation (12)

$$\frac{\partial h_1}{\partial\beta_w} = \frac{\partial^2\mathcal{L}}{\partial\lambda_z\partial\beta}, \quad \frac{\partial h_1}{\partial\lambda_z} = \frac{\partial^2\mathcal{L}}{\partial\lambda_z^2} = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\tilde{g}_i\tilde{g}_i' \tag{12}$$

back, we get the

$$\frac{\partial\lambda_z}{\partial\beta_w} = -\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\tilde{g}_i\tilde{g}_i'\right]^{-1}\frac{\partial^2\mathcal{L}}{\partial\lambda_z\partial\beta}$$

Then, gathering everything and insert into the expression of SOC, we obtain

$$\frac{d^2\mathcal{L}}{d\beta_w^2} = \left\{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\left[\left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)'\lambda_z\lambda_z'\left(\frac{\partial\tilde{g}_i}{\partial\beta_w}\right)\right]\right\} - \left(\frac{\partial^2\mathcal{L}}{\partial\lambda_z\partial\beta}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\tilde{g}_i\tilde{g}_i'\right]^{-1}\frac{\partial^2\mathcal{L}}{\partial\lambda_z\partial\beta} \tag{13}$$

Unlike the original FWL theorem for linear least-square type estimator, the SOC from empirical likelihood is a positive semi-definite matrix minus a positive semi-definite matrix. It takes a complicated shape with no guarantee that $d^2\mathcal{L}/d\beta_w^2$ to be negative definite in the finite sample. Intuitively, if λ_z takes a small value, the first term would diminish faster, and we would have the desirable negative definiteness. For a minimum condition, as $N\to\infty$, by Assumption 4, 5 and Qin

13. Here $\partial^2\lambda_z/\partial\beta_w^2$ is a three-dimensional matrix.

and Lawless (1994), $\hat{\lambda}_z^{EL}$ is a root-N consistent estimator with $\hat{\lambda}_z^{EL} \to 0$, equation (13) becomes,

$$\frac{d^2\mathcal{L}}{d\beta_w^2} \to -E\left[\frac{\partial \tilde{g}_i}{\partial \beta_w}\right]' (E[\tilde{g}_i \tilde{g}_i'])^{-1} E\left[\frac{\partial \tilde{g}_i}{\partial \beta_w}\right]$$

By Assumption 4, the non-singularity of $E[\partial \tilde{g}_i/\partial \beta_w]$ preserves the positive definiteness of $E[\tilde{g}_i \tilde{g}_i']^{-1}$. Hence, the negative-definiteness second order condition for the maximization problem with respect to β_w is asymptotically satisfied almost surely by $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ as $N \to \infty$ [14]. □

B.3 Proof of Theorem 2

Proof. Suppose we denote the parameters in $\gamma = (\hat{\beta}_w, \hat{\lambda}_z)$. Consider a composition function from $\mathbb{R}^{K+L} \to \mathbb{R}^{K+L}$ with $\gamma^k = f_N(\gamma^{k-1})$. In Algorithm 1, we start at a fair starting value (usually 2SLS), and then iterating until γ with $\gamma = f_N(\gamma)$ is reached. According to Theorem 1 and Qin and Lawless (1994), our fixed point iteration converges to full EL solution, and full EL solution itself is a consistent estimator. Locally, for large N, we can expand around $\gamma^0 = (\beta_w^{EL}, \lambda_z^{EL}) \to (\beta_w^*, \lambda_z^*) = (\beta_w^*, 0)$, which yields,

$$\begin{aligned} \hat{\gamma}^k &= f_N(\hat{\gamma}^{k-1}) \\ &= \hat{\gamma}^{EL} + \bigtriangledown_\gamma f_N(\hat{\gamma}^{EL})(\hat{\gamma}^{k-1} - \hat{\gamma}^{EL}) + O_p(||\hat{\gamma}^{k-1} - \hat{\gamma}^{EL}||^2) \end{aligned} \tag{14}$$

Denote m_i as any of (y_i, w_i, z_i) and α_m as the parameters estimated in the WLS with m as dependent variable. For instance, α_Y denotes the vector of parameters obtained from the WLS regression with y_i as dependent variable and x_i as regressors. Two sets of FOC needs to be satisfied to pin down $f_N(\gamma^{k-1})$. With the superscript k omitted, we have

$$h_1 = -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} \tilde{g}_i = 0, \quad h_2 = -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + \lambda_z' \tilde{g}_i} \left(\frac{\partial \tilde{g}_i}{\partial \beta_w}\right)' \lambda_z \tag{15}$$

The first equation in (15) determines λ_z for a given β_w. Use the Implicit Function Theorem for α_m, we have $\partial h_1/\partial \alpha_m + \partial h_1/\partial \lambda_z \times \partial \lambda_z/\partial \alpha_m = 0$. From the proof of Theorem 1, we have $\partial h_1/\partial \lambda_z \neq 0$

14. Though the conclusion drawn for SOC is asymptotic, in our simulations, we notice that the violation of SOC at $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL})$ is very rare. For instance, in our simulations, we find that there are indeed some cases where a satisfaction in the full problem does not translate into the negative definiteness for the partitioned optimization. However, the situation is fairly rare. In the simulation, we have N = 30, K = 1, M = 1, L = 2. Only 5 out of 10000 repetitions ends up with problematic SOC when the SOC is satisfied for the full problem. If we increase the sample size N to 100, then we do not observe any violations. In summary, even when proof for finite sample is not attainable, the SOC would hold in practice with high probability, especially when we have moderate to large sample size.

when evaluated at $(\beta_w^*, 0)$. And for $\partial h_1/\partial \alpha_m$

$$\frac{\partial h_1}{\partial \alpha_m} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\tilde{g}_i\lambda_z'\frac{\partial \tilde{g}_i}{\partial \tilde{\alpha}_m} - \frac{1}{1+\lambda_z'\tilde{g}_i}\frac{\partial \tilde{g}_i}{\partial \alpha_m}\right]$$

From Theorem 1.3, as $N \to \infty$, $(\hat{\beta}_w^{EL}, \hat{\lambda}_z^{EL}) \to (\beta_w^*, 0)$ becomes a fixed point.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\tilde{g}_i\lambda_z'\frac{\partial \tilde{g}_i}{\partial \tilde{\alpha}_m}\Big|_{(\beta_w^0,0)} \to 0, \quad \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\lambda_z'\tilde{g}_i}\frac{\partial \tilde{g}_i}{\partial \alpha_m} \to E\left[\frac{\partial \tilde{g}_i}{\partial \alpha_m}\right]$$

And

$$E\left[\frac{\partial \tilde{g}_i}{\partial \alpha_m}\right] = \begin{cases} -E[\tilde{z}_i x_i'] = 0, & m_i \in \{y_i, w_i\} \\ -E[(\tilde{y}_i - \tilde{w}_i'\beta_w)\otimes x_i)] = 0, & m_i \in z_i \end{cases}$$

Hence we have $\partial\lambda_z/\partial\alpha_m \to 0$ as $N \to \infty$. The second equation h_2 in (15) pin down the β_w under the condition that λ_z is the optimal solution, and also by Implicit Function Theorem,

$$\frac{\partial h_2}{\partial \alpha_m} + \frac{\partial h_2}{\partial \beta_w}\frac{\partial \beta_w}{\partial \alpha_m} + \frac{\partial h_2}{\partial \lambda_z}\frac{\partial \lambda_z}{\partial \beta_w}\frac{\partial \beta_w}{\partial \alpha_m} = 0$$

From the proof of Theorem 1.3, as $N \to \infty$, $\partial h_2/\partial\beta_w + \partial h_2/\partial\lambda_z \times \partial\lambda_z/\partial\beta_w$ goes to a positive definite matrix by assumption. And

$$\frac{\partial h_2}{\partial \alpha_m} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(1+\lambda_z'\tilde{g}_i)^2}\left(\frac{\partial \tilde{g}_i}{\partial \beta_w}\right)'\lambda_z\lambda_z'\frac{\partial \tilde{g}_i}{\partial \alpha_m} + \frac{1}{1+\lambda_z'\tilde{g}_i}\frac{\partial}{\partial \alpha_m}\left[\left(\frac{\partial \tilde{g}_i}{\partial \beta_w}\right)'\lambda_z\right] \to 0$$

Therefore, we have $\partial\beta_w/\partial\alpha_m \to 0$. By the Delta method, $\bigtriangledown_\gamma f_N(\gamma^{EL}) \to 0$ at the rate root-N. As a result, equation (14) becomes,

$$\left\|\hat{\gamma}_w^k - \hat{\gamma}_w^{EL}\right\| = O_p\left(N^{-1/2}\left\|\hat{\gamma}_w^{k-1} - \hat{\gamma}_w^{EL}\right\| + \left\|\hat{\gamma}_w^{k-1} - \hat{\gamma}_w^{EL}\right\|^2\right)$$

□

B.4 Proofs for Theorem 3

Proof. Here, what we want to show is, the solution $(\hat{\beta}^{CI}, \hat{\pi}^{CI})$ that satisfy the FOC for the full problem of $\overline{C}$, would also satisfy the FOC for the partitioned problem, when partitioned by the π^{CI}. Trivially, (β^{CI}, π^{CI}) is a feasible solution to the primal partitioned problem. There are both exogenous regressors x_i (to be partitioned out) and endogenous regressors w_i. And we try to

calculate the confidence interval for β_{wj}, the jth element of β_w. The Lagrange expression for the full problem becomes,

$$\begin{aligned}\mathcal{H} = &- T(\beta) + t\Big[1 - \sum_{i=1}^{N} \pi_i\Big] - N\lambda_x'\Big[\sum_{i=1}^{N} \pi_i g_{x,i}\Big] - N\lambda_z'\Big[\sum_{i=1}^{N} \pi_i g_{z,i}\Big] + \\ &\gamma\Big[\sum_{i=1}^{N} \ln(\hat{\pi}_i^{EL}) - \chi_1^2(1-\alpha)/2 - \sum_{i=1}^{N} \ln(\pi_i)\Big]\end{aligned}$$

Take derivative with respect to π_i following Qin and Lawless (1994) and obtain

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial \pi_i} &= -t - N\lambda_x' g_{x,i} - N\lambda_z' g_{z,i} - \gamma\frac{1}{\pi_i} = 0 \\ \sum_{i=1}^{N} \pi_i \frac{\partial \mathcal{H}}{\partial \pi_i} &= -t\sum_{i=1}^{N} \pi_i - N\lambda_x'\Big[\sum_{i=1}^{N} \pi_i g_{x,i}\Big] - N\lambda_z'\Big[\sum_{i=1}^{N} \pi_i g_{z,i}\Big] - N\gamma = 0 \\ &- t - N\gamma = 0 \Rightarrow t = -N\gamma\end{aligned} \tag{16}$$

The inequality constraint $\sum_{i=1}^{N} \ln(\hat{\pi}_i^{EL}) - \chi_1^2(1-\alpha)/2 - \sum_{i=1}^{N} \ln(\pi_i) \leq 0$ needs to be binding to have a bounded optimization problem. Also this can be seen from the complementary slackness. If the inequality constraint is inactive, then $\gamma = 0$. By equation (16), we have t = 0, and correspondingly $\lambda_x = 0$ and $\lambda_z = 0$. This would not generate a feasible solution since from $\partial\mathcal{H}/\partial\beta_{wj}$, $N\lambda_x(-\sum_{i=1}^{N} \pi_i w_i x_i') + N\lambda_z(-\sum_{i=1}^{N} \pi_i w_i z_i') = 1$.

Then we are left with a binding equality constraint and $\gamma \neq 0$. Insert t back,

$$N\gamma - N\lambda_x' g_{x,i} - N\lambda_z' g_{z,i} - \gamma\frac{1}{\pi_i} = 0 \Rightarrow \pi_i = \frac{1}{N}\frac{1}{1 - (1/\gamma) \times [\lambda_x' g_{x,i} + \lambda_z' g_{z,i}]} \tag{17}$$

With (17), the set of FOC justifying the solution can be expressed as

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial \beta_w} &= -\frac{\partial T(\beta)}{\partial \beta_w} - N\lambda_x'\Big[\sum_{i=1}^{N} \pi_i w_i x_i'\Big] - N\lambda_z'\Big[\sum_{i=1}^{N} \pi_i w_i z_i'\Big] = 0 \\ \frac{\partial \mathcal{H}}{\partial \beta_x} &= -N\lambda_x'\Big[\sum_{i=1}^{N} \pi_i x_i x_i'\Big] - N\lambda_z'\Big[\sum_{i=1}^{N} \pi_i x_i z_i'\Big] = 0 \\ \frac{\partial \mathcal{H}}{\partial \gamma} &= \sum_{i=1}^{N} \ln(\hat{\pi}_i^{EL}) - \chi_1^2(1-\alpha)/2 - \sum_{i=1}^{N} \ln(\pi_i) = 0 \\ \frac{\partial \mathcal{H}}{\partial \lambda_x} &= \sum_{i=1}^{N} \pi_i g_{x,i} = 0, \quad \frac{\partial \mathcal{H}}{\partial \lambda_z} = \sum_{i=1}^{N} \pi_i g_{z,i} = 0, \quad \pi_i = \frac{1}{N}\frac{1}{1 - (1/\gamma) \times [\lambda_x' g_{x,i} + \lambda_z' g_{z,i}]}\end{aligned} \tag{18}$$

Then compare this with the partitioned FOC when partitioned out by $\hat{\pi}_i^{CI}$, and the corresponding Lagrange expression for β_w^P, π^P is

$$\mathcal{L} = -T(\beta_w) + t\Big[1 - \sum_{i=1}^{N} \pi_i^P\Big] - N\lambda_z'\Big[\sum_{i=1}^{N} \pi_i^P \tilde{g}_{z,i}\Big] + \gamma\Big[\sum_{i=1}^{N} \ln(\hat{\pi}_i^{EL}) - \chi_1^2(1-\alpha)/2 - \sum_{i=1}^{N} \ln(\pi_i^P)\Big]$$

And the corresponding FOC,

$$\begin{aligned} &\frac{\partial \mathcal{L}}{\partial \beta_w} = -\frac{\partial T(\beta)}{\partial \beta_w} + N\lambda_z'(-\sum_{i=1}^{N} \pi_i^P \tilde{w}_i \tilde{z}_i'), \quad \frac{\partial \mathcal{L}}{\partial \gamma} = \sum_{i=1}^{N} \ln(\hat{\pi}_i^{EL}) - \chi_1^2(1-\alpha)/2 - \sum_{i=1}^{N} \ln(\pi_i^P) \\ &\frac{\partial \mathcal{L}}{\partial \lambda_z} = \sum_{i=1}^{N} \pi_i^P \tilde{g}_{z,i}, \quad \pi_i^P = \frac{1}{N} \frac{1}{1 - (1/\gamma) \times [\lambda_z' \tilde{g}_{z,i}]} \end{aligned} \tag{19}$$

What we want to show is, if the solution $(\pi_i^{CI}, \beta_w^{CI}, \lambda_z^{CI}, \gamma^{CI})$ satisfies (18), then also satisfies (19). This is straightforward. Overall, the confidence interval is constructed using empirical likelihood, hence some major steps in proving point estimation still work here. Firstly, from $\partial \mathcal{H}/\partial \beta_x = 0$ in (18), we have $\lambda_x = -\hat{\alpha}_z \lambda_z$, and $\hat{\alpha}_z = (\sum_{i=1}^{N} \pi_i x_i x_i')^{-1} \sum_{i=1}^{N} \pi_i x_i z_i'$. Thus, $\hat{\pi}_i^{CI} = \pi_i^P$, since $\lambda_x' g_{x,i} + \lambda_z' g_{z,i} = \lambda_z' \tilde{g}_{z,i}$ following the proof in Theorem 1.3. Also $\partial \mathcal{L}/\partial \gamma = \partial \mathcal{H}/\partial \gamma = 0$ since they are only a function of $\ln(\pi_i)$. Then we also have $\partial \mathcal{H}/\partial \lambda_z = \partial \mathcal{L}/\partial \lambda_z$ and $\partial \mathcal{H}/\partial \beta_w = \partial \mathcal{L}/\partial \beta_w = 0$ – both of these from the analysis of Theorem 1.3.

C. Computational Method

Our algorithms, like Algorithm 1, includes fixed point iteration over π. However, simply replacing π^k with the π^{k+1} from $(\beta_w^{k+1}, \lambda_z^{k+1})$, could be numerically unstable, especially in small samples. Therefore, we adopt the spectral algorithm from La Cruz et al. (2006). Suppose the fixed point $\breve{\pi}$ satisfies the equation, $\phi(\pi) - \pi = 0$, where $\phi(\pi)$ is a composite function containing the procedure of WLS, partitioned EL, and the retrieve of π using β and λ. We calculate π^{k+1} following $\pi^{k+1} = \pi^k - \tau_k(\phi(\pi^k) - \pi^k)$.

The π^{k+1} is updated by a weighted sum of π^k and the residuals of the equations $\phi(\pi^k) - \pi^k$. And α_k is defined as

$$\tau_k = \frac{(\pi^k - \pi^{k-1})'(\pi^k - \pi^{k-1})}{(\pi^k - \pi^{k-1})'[\phi(\pi^k) - \pi^k - (\phi(\pi^{k-1}) - \pi^{k-1})]}$$

Note that if τ_k equals to -1, we would have the simple updates $\pi^{k+1} = \phi(\pi^k)$. But usually, a different τ_k would be chosen. Besides, π_i^{k+1} still add up to one since $\sum_{i=1}^{N} \pi_i^{k+1} = \sum_{i=1}^{N} \pi_i^k - \tau_k(\phi_i(\pi^k) - \pi_i^k) =$

$(1+\tau_k)\sum_{i=1}^{N}\pi_k - \tau_k\sum_{i=1}^{N}\phi_i(\pi_i^k) = 1+\tau_k-\tau_k = 1$. To have $\pi_i > 0$, $\forall i \in 1,\cdots,N$, we also need $\pi_i^{k+1} = \pi_i^k - \tau_k(\phi_i(\pi^k) - \pi_i^k) > 0$; or, $\forall i \in 1,\cdots,N$ with $\phi_i(\pi^k) - \pi_i^k$ and τ_k has the same sign, $|\tau_k| < \sup_{i\in\{1,\cdots,N\}} |\pi_i^k/(\phi_i(\pi^k) - \pi_i^k)|$. The latter holds when we start from a set of π that is not far from the fixed point.

As pointed out by La Cruz et al. (2006) and Aguirregabiria and Marcoux (2021), this type of spectral algorithm possesses several advantages. Firstly, it combines several previously-proposed methods' strengths and keeps a well-defined problem in each step. Secondly, it offers a non-monotone global path for line search and a better chance to escape stagnation. Thirdly, it does not need the analytical formation for the gradients – actually, the gradients are approximated by the formula calculating α_k. Our Monte Carlo simulations give us better numerical convergence than simply using the fixed-point iteration, especially with small sample. □