MINIMUM DISTANCE APPROACH TO INFERENCE WITH MANY INSTRUMENTS*

Michal Kolesár†

Department of Economics and Woodrow Wilson School, Princeton University

This version 3.1.2, January 10, 2018

First circulated: November 2012.

Abstract

I analyze a linear instrumental variables model with a single endogenous regressor and many instruments. I use invariance arguments to construct a new minimum distance objective function. With respect to a particular weight matrix, the minimum distance estimator is equivalent to the random effects estimator of Chamberlain and Imbens (2004), and the estimator of the coefficient on the endogenous regressor coincides with the limited information maximum likelihood estimator. This weight matrix is inefficient unless the errors are normal, and I construct a new, more efficient estimator based on the optimal weight matrix. Finally, I show that when the minimum distance objective function does not impose a proportionality restriction on the reduced-form coefficients, the resulting estimator corresponds to a version of the bias-corrected two-stage least squares estimator. I use the objective function to construct confidence intervals that remain valid when the proportionality restriction is violated.

Keywords: Instrumental Variables, Minimum Distance, Incidental Parameters, Random Effects, Many Instruments, Misspecification, Limited Information Maximum Likelihood, Bias-Corrected Two-Stage Least Squares.
JEL Codes: C13; C26; C36

*Previously circulated under the title "Integrated Likelihood Approach to Inference with Many Instruments." I am deeply grateful to Guido Imbens and Gary Chamberlain for their guidance and encouragement. I also thank Joshua Angrist, Adam Guren, Whitney Newey, Jim Stock, Peter Phillips and participants at various seminars and conferences for their helpful comments and suggestions.

†Correspondence to: Department of Economics, Julis Romo Rabinowitz Building, Princeton University, Princeton, NJ 08544. Electronic correspondence: mkolesar@princeton.edu.

1 Introduction

This paper provides a principled and unified way of doing inference in a linear instrumental variables model with a single endogenous regressor and homoscedastic errors in which the number of instruments, k_n, is potentially large. To capture this feature in asymptotic approximations, I employ the many instrument asymptotics of Kunitomo (1980), Morimune (1983), and Bekker (1994) that allow k_n to increase in proportion with the sample size, n. I focus on the case in which collectively the instruments have substantial predictive power, so that the concentration parameter grows at the same rate as the sample size. I make no assumptions about the strength of individual instruments. I allow the rate of growth of k_n to be zero, in which case the asymptotics reduce to the standard few strong instrument asymptotics.

The presence of many instruments creates an incidental parameters problem (Neyman and Scott, 1948), as the number of first-stage coefficients, k_n, increases with the sample size. To directly address this problem, I use sufficiency and invariance arguments together with an assumption that the reduced-form errors are normally distributed to reduce the data to a pair of two-by-two matrices. In the absence of exogenous regressors, the first matrix can be written as $T = \begin{pmatrix} y & x \end{pmatrix}' P_Z \begin{pmatrix} y & x \end{pmatrix} / n$, where P_Z is the projection matrix of the instruments Z, and y and x are vectors corresponding to the outcome and the endogenous regressor. The second matrix, $S = \begin{pmatrix} y & x \end{pmatrix}' (I_n - P_Z) \begin{pmatrix} y & x \end{pmatrix} / (n - k_n)$, where I_n is the identity, corresponds to an estimator of the reduced-form covariance matrix. This solves the incidental parameters problem because the distribution of T and S depends on a fixed number of parameters even as $k_n \to \infty$: it depends on the first-stage coefficients only through the parameter λ_n, a measure of their collective strength.

I then drop the normality assumption and use a restriction on the first moment of T implied by the model to construct a minimum distance (MD) objective function. This restriction follows from the property of the instrumental variables model that the coefficients on the instruments in the first-stage regression are proportional to the coefficients in the reduced-form outcome regression. I use this MD objective function to derive three main results.

First, I show that minimizing the MD objective function with respect to the optimal weight

matrix yields a new estimator of β, the coefficient on the endogenous regressor, that exhausts the information in T and S. In particular, this efficient MD estimator is asymptotically more efficient than the limited information maximum likelihood (LIML) estimator when the reduced-form errors are not normal. Standard errors can easily be constructed using the usual sandwich formula for asymptotic variance of minimum distance estimators.[1] The MD approach thus gives a simple practical solution to the many-instrument incidental parameters problem.

Second, I compare the MD approach to that based on the invariant likelihood—the likelihood, under normality, based on T and S. I show that, when combined with a particular prior on λ_n, the likelihood is equivalent to the random-effects (RE) quasi-maximum likelihood of Chamberlain and Imbens (2004), and that maximizing it yields LIML. Therefore, the random-effects estimator of β is in fact equivalent to LIML. Furthermore, I show that the RE estimator of the model parameters also minimizes the MD objective function with respect to a particular weight matrix. This weight matrix is efficient under normality, but not in general.

Third, I consider minimum distance estimation that leaves the first moment of T unrestricted. This situation arises, for instance, when the instrumental variables model is used to estimate potentially heterogeneous causal effects, as in Angrist and Imbens (1995). When the causal effect is heterogeneous, the reduced-form coefficients are no longer proportional, so that the first moment of T is unrestricted. In this case, the instrumental variables estimand β can be interpreted as a weighted average of the marginal effect of the endogenous variable on the outcome (Angrist, Graddy and Imbens, 2000). I show that the unrestricted minimum distance estimator coincides with a version of the bias-corrected two-stage least squares estimator (Nagar, 1959; Donald and Newey, 2001), and use the MD objective function to construct confidence intervals that remain valid when the proportionality restriction is violated.

The MD objective function is also helpful in deriving a specification test that is robust to many instruments. By testing the restriction on the first moment of T, I derive a new test that is similar to that of Cragg and Donald (1993), but with an adjusted critical value. The adjustment ensures that the test is valid under few strong as well as many instrument asymptotics that also allow for many regressors. In contrast, when the number of regressors is allowed to increase

[1] Software implementing estimators and standard errors based on the MD objective function is available at `https://github.com/kolesarm/ManyIV`.

with the sample size, the size of the standard Sargan (1958) specification test converges to one, as does the size of the test proposed by Anatolyev and Gospodinov (2011).

The paper draws on two separate strands of literature. First, the literature on many instruments that builds on the work by Kunitomo (1980), Morimune (1983), Bekker (1994) and Chao and Swanson (2005). Like Anatolyev (2013), I relax the assumption that the dimension of regressors is fixed, and I allow them to grow with the sample size. Hahn (2002), Chamberlain (2007), Chioda and Jansson (2009), and Moreira (2009) focus on optimal inference with many instruments when the errors are normal and homoscedastic, and my optimality results build on theirs. Papers by Hansen, Hausman and Newey (2008), Anderson, Kunitomo and Matsushita (2010) and van Hasselt (2010) relax the normality assumption. Hausman, Newey, Woutersen, Chao and Swanson (2012), Chao, Swanson, Hausman, Newey and Woutersen (2012), Chao, Hausman, Newey, Swanson and Woutersen (2014) and Bekker and Crudu (2015) also allow for heteroscedasticity. An interesting new development is to employ shrinkage or regularization to solve the incidental parameters problem (see, for example, Belloni, Chen, Chernozhukov and Hansen, 2012, Gautier and Tsybakov, 2014, or Carrasco, 2012). When combined with additional assumptions on the model, these shrinkage estimators can be more efficient than the efficient MD estimator proposed here.

Second, the literature on incidental parameters dating back to Neyman and Scott (1948). Lancaster (2000) and Arellano (2003) discuss the incidental parameters problem in a panel data context. Chamberlain and Moreira (2009) relate invariance and random effects approaches to the incidental parameters problem in a dynamic panel data model. My results on the relationship between these two approaches in an instrumental variables model build on theirs. Sims (2000) proposes a similar random-effects solution in a dynamic panel data model. Moreira (2009) proposes to use invariance arguments to solve the incidental parameters problem.

The remainder of this paper is organized as follows. Section 2 sets up the instrumental variables model, and reduces the data to the T and S statistics. Section 3 considers likelihood-based approaches to inference under normality. Section 4 relaxes the normality assumption and considers the MD approach to inference. Section 5 considers MD estimation without imposing proportionality of the reduced-form coefficients. Section 6 studies tests of overidentifying

restrictions. Section 7 concludes. Proofs and derivations are collected in the appendix. The supplemental appendix contains additional derivations.

2 Setup

In this section, I first introduce the model, notation, and the many instrument asymptotic sequence that allows both the number of instruments and the number of exogenous regressors to increase in proportion with the sample size. I then reduce the data to the low-dimensional statistics T and S, and define the minimum distance objective function.

2.1 Model and Assumptions

There is a sample of individuals $i = 1, \ldots, n$. For each individual, we observe a scalar outcome y_i, a scalar endogenous regressor x_i, ℓ_n-dimensional vector of exogenous regressors w_i, and k_n-dimensional vector of instruments z_i^*. The instruments and exogenous regressors are treated as non-random.

It will be convenient to define the model in terms of an orthogonalized version of the original instruments. To describe the orthogonalization, let W denote the $n \times \ell_n$ matrix of regressors with ith row equal to w_i', and let Z^* denote the $n \times k_n$ matrix of instruments with ith row equal to $z_i^{*\prime}$. Let $\tilde{Z} = Z^* - W(W'W)^{-1}W'Z^*$ denote the residuals from regressing Z^* onto W. Then the orthogonalized instruments $Z \in \mathbb{R}^{n \times k_n}$ are given by $Z = \tilde{Z}R^{-1}$, where the upper-triangular matrix $R \in \mathbb{R}^{k_n \times k_n}$ is the Cholesky factor of $\tilde{Z}'\tilde{Z}$. Now, by construction, the columns of Z are orthogonal to each other as well as to the columns of W.[2]

Denote the ith row of Z by z_i', and let $Y \in \mathbb{R}^{n \times 2}$ with rows (y_i, x_i) pool all endogenous variables in the model. The reduced form regression of Y onto Z and W can be written as

$$Y = Z \begin{pmatrix} \pi_{1,n} & \pi_{2,n} \end{pmatrix} + W \begin{pmatrix} \psi_{1,n} & \psi_{2,n} \end{pmatrix} + V, \tag{1}$$

where $V \in \mathbb{R}^{n \times 2}$ with rows $v_i' = (v_{1i}, v_{2i})$ pools the reduced-form errors, which are assumed to

[2]This orthogonalization is sometimes called a standardizing transformation; see Phillips (1983) for discussion.

be mean zero and homoscedastic,

$$\mathbb{E}[v_i] = 0, \qquad \text{and} \qquad \mathbb{E}[v_i v_i'] = \Omega. \tag{2}$$

The reduced-form coefficients on the instruments are assumed to satisfy a proportionality restriction, and the parameter of interest, β, corresponds to the constant of proportionality:

Assumption PR (Proportionality restriction). $\pi_{1,n} = \pi_{2,n}\beta$.

The proportionality restriction implies that

$$y_i = x_i\beta + w_i'\beta_n^w + \epsilon_i, \tag{3}$$

where $\epsilon_i = v_{1i} - v_{2i}\beta$ is known as the structural error, and $\beta_n^w = \psi_{1,n} - \psi_{2,n}\beta$. This equation is known as the structural equation. In Section 5, I allow for certain violations of this assumption, such as when the effect of x_i on y_i is heterogeneous. Throughout the paper, I assume that $k_n > 1$, which implies that Assumption PR is testable; I discuss specification testing in Section 6. In order to employ sufficiency and invariance arguments, I assume that v_i has a normal distribution:

Assumption N (Normality). The errors v_i are i.i.d. and normally distributed.

This assumption has no effect on the consistency of estimators considered in this paper, although it does affect their asymptotic distribution and asymptotic efficiency. I drop this assumption in Sections 4 to 6 when I discuss a minimum-distance approach to inference.

To measure the strength of identification, I follow Chamberlain (2007) and Andrews, Moreira and Stock (2008), and use the parameter

$$\lambda_n = \pi_{2,n}'\pi_{2,n} \cdot a'\Omega^{-1}a/n, \qquad a = \begin{pmatrix} \beta \\ 1 \end{pmatrix}. \tag{4}$$

The goal is to construct inference procedures that work well even if the number of instruments k_n and the number of exogenous regressors ℓ_n is large relative to sample size. I therefore

follow Anatolyev (2013) and Kolesár, Chetty, Friedman, Glaeser and Imbens (2015) and allow both k_n and ℓ_n to potentially grow with the sample size:

Assumption MI (Many instruments). **(i)** $k_n/n = \alpha_k + o(n^{-1/2})$ and $\ell_n/n = \alpha_\ell + o(n^{-1/2})$ for some $\alpha_\ell, \alpha_k \geq 0$ such that $\alpha_k + \alpha_\ell < 1$; **(ii)** The matrix $\begin{pmatrix} W & Z \end{pmatrix}$ is non-random and full column rank $k_n + \ell_n$; **(iii)** $\sum_{i=1}^n ((z_i'\pi_{2,n})^4 + (z_i'\pi_{1,n})^4)/n^2 = o(1)$; and **(iv)** $\lambda_n = \lambda + o(1)$ for some $\lambda > 0$.

Assumption MI (i) weakens the many instrument sequence of Bekker (1994) by allowing ℓ_n to grow with the sample size. The motivation for this is twofold. First, often the presence of many instruments is the result of interacting a few basic instruments with many regressors (as in, for example Angrist and Krueger, 1991), in which case both ℓ_n and k_n are large. Second, oftentimes the instruments are valid only conditional on a large set of regressors w_i; for example, if the instruments are randomly assigned within a school, we need to condition on school fixed effects. By allowing $\alpha_k = \alpha_\ell = 0$, the assumption nests the standard few strong instrument asymptotic sequence in which the number of instruments and regressors is fixed. Parts (ii)–(iv) of Assumption MI are standard. Part (ii) is a normalization that requires excluding redundant columns from W and excluding columns of Z that are redundant or already included in W. It ensures that the reduced-form coefficients in (1) are well-defined. Part (iii) is used to verify the Lindeberg condition. Part (iv) is the many-instruments equivalent of the relevance assumption. It is equivalent to the assumption that the concentration parameter (Rothenberg, 1984), given by $\pi_{2,n}'\pi_{2,n}/\Omega_{22}$, grows at the same rate as the sample size.

2.2 Sufficient statistics and limited information likelihood

Under normality, the set of sufficient statistics is given by the least-squares estimators of the reduced-form coefficients $\Pi_n = \begin{pmatrix} \pi_{1,n} & \pi_{2,n} \end{pmatrix}$ and $\Psi_n = \begin{pmatrix} \psi_{1,n} & \psi_{2,n} \end{pmatrix}$,

$$\begin{pmatrix} \hat{\Pi} \\ \hat{\Psi} \end{pmatrix} = \begin{pmatrix} Z'Y \\ (W'W)^{-1}W'Y \end{pmatrix} \in \mathbb{R}^{(k_n+\ell_n)\times 2},$$

and an unbiased estimator of the reduced-form covariance matrix Ω based on the residual sum of squares,

$$S = \frac{1}{n - k_n - \ell_n} Y'(I_n - ZZ' - W'(W'W)^{-1}W)Y \in \mathbb{R}^{2\times 2}, \tag{5}$$

The advantage of working with the orthogonalized instruments is that now the rows of $\hat{\Pi}$ are mutually independent. Since the distribution of $\hat{\Psi}$ is unrestricted, we can drop it from the model and base inference on $\hat{\Pi}$ and S only as in Moreira (2003) and Chamberlain and Imbens (2004). This step eliminates the potentially high-dimensional nuisance parameter Ψ_n, so that the model parameters are now given by the triplet $(\beta, \pi_{2,n}, \Omega)$.

Estimators considered in this paper will only depend on $\hat{\Pi}$ through the statistic

$$T = \frac{1}{n}\hat{\Pi}'\hat{\Pi} = \frac{1}{n}Y'ZZ'Y \in \mathbb{R}^{2\times 2}. \tag{6}$$

Define the following functions of the statistics T and S:

$$Q_S(\beta, \Omega) = \frac{b'Tb}{b'\Omega b}, \qquad Q_T(\beta, \Omega) = \frac{a'\Omega^{-1}T\Omega^{-1}a}{a'\Omega^{-1}a}, \qquad b = \begin{pmatrix} 1 \\ -\beta \end{pmatrix},$$

and let $m_{\min}$ and $m_{\max}$ denote the minimum and maximum eigenvalues of the matrix $S^{-1}T$.

The likelihood of the model (1)–(2) under Assumptions PR and N is known as the limited information likelihood (Anderson and Rubin, 1949). The limited information maximum likelihood (LIML) estimator of β is given by

$$\hat{\beta}_{\text{LIML}} = \operatorname*{argmax}_{\beta} Q_T(\beta, S) = \operatorname*{argmin}_{\beta} Q_S(\beta, S) = \frac{T_{12} - m_{\min}S_{12}}{T_{22} - m_{\min}S_{22}}. \tag{7}$$

It turns out that $\hat{\beta}_{\text{LIML}}$ is consistent and asymptotically normal under Assumption MI despite the incidental parameters problem (Bekker, 1994). I will give some insight into this result in Section 3.

Due to the incidental parameters problem, the (β, β) block of the inverse information matrix of the limited information likelihood does not yield a consistent estimate of the asymptotic variance of LIML. The asymptotic distribution of $\hat{\beta}_{\text{LIML}}$ under Assumptions PR, N and MI is

given by (see Bekker, 1994 and Kolesár *et al.*, 2015 for derivation)

$$\sqrt{n}\left(\hat{\beta}_{\text{LIML}} - \beta\right) \Rightarrow \mathcal{N}\left(0, \mathcal{V}_{\text{LIML},N}\right), \tag{8}$$

where $\Rightarrow$ denotes convergence in distribution, and

$$\mathcal{V}_{\text{LIML},N} = \frac{b'\Omega b \cdot a'\Omega^{-1}a}{\lambda}\left(1 + \frac{\alpha_k(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell}\frac{1}{\lambda}\right). \tag{9}$$

In contrast, the (β,β) block of the inverse information matrix is given by $b'\Omega b \cdot a'\Omega^{-1}a/(n\lambda_n)$, missing the correction factor in parentheses (see supplemental appendix for derivation). This correction factor can be substantial even when the ratio of instruments to sample size, α_k, is small if λ is small.

2.3 Using invariance to reduce the dimension of the parameter space

To reduce the dimension of the parameter space, I follow Andrews, Moreira and Stock (2006), Chamberlain (2007), Chioda and Jansson (2009), and Moreira (2009), and require decision rules (procedures used for constructing point estimates and confidence intervals from the data) to be invariant with respect to rotations of the instruments. In other words, changing the co-ordinate system for the instruments should not affect inference about β—if we re-order the instruments, or use a different orthogonalization procedure to construct Z, we should get the same point estimate and confidence interval for β. A decision rule is invariant under rotations of instruments if it remains unchanged under the transformation $(\hat{\Pi}, S) \mapsto (g\hat{\Pi}, S)$, where $g \in \mathcal{O}(k_n)$, the group of $k_n \times k_n$ orthogonal matrices. A necessary and sufficient condition for a decision rule to be invariant is that it depends on the data only through a maximal invariant (Eaton, 1989, Theorem 2.3). A statistic $m(\hat{\Pi}, S)$ is maximal invariant if (i) $m(\hat{\Pi}, S) = m(g\hat{\Pi}, S)$ for all $g \in \mathcal{O}(k_n)$; and (ii) whenever $m(\dot{\Pi}, \dot{S}) = m(\ddot{\Pi}, \ddot{S})$ for some $\dot{\Pi}, \ddot{\Pi} \in \mathbb{R}^{k_n \times 2}$ and $\dot{S}, \ddot{S} \in \mathbb{R}^{2\times 2}$, then $(\dot{\Pi}, \dot{S}) = (g\ddot{\Pi}, \ddot{S})$ for some $g \in \mathcal{O}(k_n)$. It is straightforward to check that the pair of matrices (S, T) is a maximal invariant statistic. The distribution of (S, T) depends on $\pi_{2,n}$ only through $\pi_{2,n}'\pi_{2,n}$, or equivalently through $\lambda_n = \pi_{2,n}'\pi_{2,n} \cdot a'\Omega^{-1}a/n$.

This reduces the parameter space to $(\beta, \lambda_n, \Omega)$, which has a fixed dimension.[3]

There are two general approaches to constructing invariant decision rules based on the maximal invariant (S, T). First, one can use the likelihood based on S and T, called the invariant likelihood, $\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T)$. I consider this approach in detail in Section 3. The disadvantage of this approach is that the validity of inference based on the invariant likelihood is sensitive to Assumption N.

The second approach is to use moment restrictions on S and T implied by the model. In particular, the reduced form (1)–(2) without any further assumptions implies

$$\mathbb{E}[S] = \Omega, \tag{10a}$$

$$\mathbb{E}[T - (k_n/n)S] = \Xi_n, \qquad \text{where} \quad \Xi_n = \frac{1}{n}\begin{pmatrix} \pi_{1,n} & \pi_{2,n} \end{pmatrix}' \begin{pmatrix} \pi_{1,n} & \pi_{2,n} \end{pmatrix}. \tag{10b}$$

Under Assumption PR, the matrix of second moments of the reduced-form coefficients, Ξ_n, has reduced rank,

$$\Xi_n = \Xi_{22,n} a a' = \Xi_{22,n} \begin{pmatrix} \beta^2 & \beta \\ \beta & 1 \end{pmatrix}, \tag{11}$$

with $\Xi_{22,n} = \pi_{2,n}' \pi_{2,n}/n = \lambda_n/(a'\Omega^{-1}a)$. This rank restriction can be used to build a minimum distance objective function[4]

$$\mathcal{Q}_n(\beta, \Xi_{22,n}; \hat{W}_n) = \operatorname{vech}\left(T - (k_n/n)S - \Xi_{22,n} a a'\right)' \hat{W}_n \operatorname{vech}\left(T - (k_n/n)S - \Xi_{22,n} a a'\right), \tag{12}$$

where $\hat{W}_n \in \mathbb{R}^{3\times 3}$ is some weight matrix. Since the nuisance parameter Ω only appears in the moment condition (10a), which is unrestricted, we can exclude the moment condition from the objective function (12) without any loss of information (Chamberlain, 1982, Section 3.2). I consider this approach in detail in Sections 4 to 6, where I show that this approach is more attractive once Assumption N is relaxed.

[3]Similar arguments can be used to generalize the results in this paper to the case with more than one endogenous variable. In particular, if $\dim(x_i) = J$ and $\dim(Y) = n \times (J+1)$, then one can use invariance arguments to reduce the data to the same pair of matrices (S, T) defined in (5) and (6), now with dimension $(J+1) \times (J+1)$.

[4]The operator $\operatorname{vech}(A)$ transforms the lower-triangular part of A into a single column—when A is symmetric, as is the case here, the operator can be thought of as vectorizing A while removing the duplicates.

3 Likelihood-based estimation and inference

This section shows that by combining the invariant likelihood with a particular prior on λ_n, we can construct a likelihood with a simple closed form that addresses the incidental parameters problem. I show that this likelihood is equivalent to the random effects likelihood of Chamberlain and Imbens (2004), and that maximizing it yields the LIML estimator of β.

First consider maximizing the invariant likelihood. To state the result, let $\omega_n = \pi_{2,n}/\|\pi_{2,n}\|$, so that $\hat{\Pi}$ and S can be parametrized by $(\beta, \omega_n, \lambda_n, \Omega)$. The parameter ω_n lies on the unit sphere $\mathbb{S}^{k_n-1}$ in $\mathbb{R}^{k_n}$ and it can be thought of as measuring the relative strength of the individual instruments.

Lemma 1. The invariant likelihood $\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T)$ is maximized over β at $\hat{\beta}_{\text{LIML}}$. This result also holds if λ_n is fixed at an arbitrary value. Furthermore,

$$\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T) = \int_{\mathbb{S}^{k_n-1}} \mathcal{L}_{\text{LI},n}(\beta, \lambda_n, \omega_n, \Omega; \hat{\Pi}, S)\, \mathrm{d}F_{\omega_n}(\omega_n), \tag{13}$$

where $\mathcal{L}_{\text{LI},n}$ denotes the limited information likelihood, and $F_{\omega_n}(\cdot)$ denotes the uniform distribution on the unit sphere $\mathbb{S}^{k_n-1}$.

The first part of Lemma 1 generalizes the result in Moreira (2009) that the maximum invariant likelihood estimator for β coincides with LIMLK when Ω is known. It also explains why the limited information likelihood produces an estimator that is robust to many instruments, even though the number of parameters in the likelihood increases with sample size: it is because LIML happens to coincide with the maximum invariant likelihood estimator.

The last part of Lemma 1 shows that the invariant likelihood is equivalent to the integrated (marginal) likelihood that puts a uniform prior on ω_n. This observation will allow me to build the connection between the invariant likelihood and the random-effects likelihood of Chamberlain and Imbens (2004). In particular, consider integrating the limited information likelihood with respect to the following prior on λ_n, in addition to the uniform prior on ω_n:

$$\lambda_n \sim \frac{\lambda}{k_n}\chi^2(k_n). \tag{14}$$

The hyperparameter λ corresponds to the limit of λ_n under Assumption MI. I allow it to be determined by the data, so that the prior will be dominated in large samples. The two priors on ω_n and λ_n are equivalent to a single normal prior over the scaled first-stage coefficients $\eta_n = \pi_{2,n}\sqrt{a'\Omega^{-1}a/n}$,

$$\eta_n \sim \mathcal{N}(0, \lambda/k_n \cdot I_{k_n}). \tag{15}$$

This prior is the random-effects prior proposed in Chamberlain and Imbens (2004). Therefore, the integrated likelihood obtained after integrating the limited information likelihood over the uniform prior on ω_n and the chi-square prior on λ_n coincides with the RE likelihood that integrates the limited information likelihood over the normal prior (15). The RE likelihood has a simple closed form:[5]

$$\begin{aligned}
\mathcal{L}_{\text{RE},n}(\beta, \lambda, \Omega) &= \int_{\mathbb{R}^{k_n}} \mathcal{L}_{\text{LI},n}(\beta, \eta_n, \omega_n, \Omega; \hat{\Pi}, S)\, \mathrm{d}F_{\eta_n|\lambda}(\eta_n \mid \lambda) \\
&= \int_{\mathbb{R}} \int_{\mathbb{S}^{k_n-1}} \mathcal{L}_{\text{LI},n}(\beta, \lambda_n, \omega_n, \Omega; \hat{\Pi}, S)\, \mathrm{d}F_{\omega_n}(\omega_n)\, \mathrm{d}F_{\lambda_n|\lambda}(\lambda_n \mid \lambda) \\
&= |S|^{\frac{n-k_n-\ell_n-3}{2}} \cdot \left(1 + \frac{n}{k_n}\lambda\right)^{-k_n/2} |\Omega|^{-\frac{n-\ell_n}{2}} e^{-\frac{1}{2}\left(\operatorname{tr}(\Omega^{-1}\tilde{S}) - \frac{n\lambda Q_T(\beta,\Omega)}{k_n/n+\lambda}\right)},
\end{aligned} \tag{16}$$

where the last equality holds up to a normalizing constant, and $\tilde{S} = (n - k_n - \ell_n)S + nT$. Chamberlain and Imbens (2004) motivate the RE prior as a modeling tool: since the prior has zero mean, it intuitively captures the idea that the individual instruments may not be very relevant. This motivation leaves it unclear however, whether inference based on the RE is asymptotically valid when the first-stage parameters are viewed as fixed. The equivalence (16) implies that one can indeed use the RE likelihood for inference. In particular, since the invariant likelihood has a fixed number of parameters and the invariant model is locally asymptotically normal (Chioda and Jansson, 2009), inference based on it will be asymptotically valid by standard arguments. Since the prior on λ_n gets dominated in large samples, this implies that inference based on the RE likelihood will also be asymptotically valid. Furthermore, since by Lemma 1 constraining λ_n does not affect the maximum invariant likelihood estimator for β,

[5]Chamberlain and Imbens (2004) also consider putting a random effects prior only on some coefficients; the coefficients on the remaining instruments are then assumed to be fixed. When referring to the random-effects likelihood, I assume that we put a random-effects prior on all coefficients.

integrating the invariant likelihood with respect to the chi-square prior (14) will not affect it either: it will still be given by $\hat{\beta}_{\text{LIML}}$. The next proposition summarizes and formalizes these results.

Proposition 1.

(i) *The* RE *likelihood* (16) *is maximized at*

$$\begin{aligned}
\hat{\beta}_{\text{RE}} &= \hat{\beta}_{\text{LIML}}, \\
\hat{\lambda}_{\text{RE}} &= \max\{m_{\max} - k_n/n, 0\}, \\
\hat{\Omega}_{\text{RE}} &= \frac{n-k_n-\ell_n}{n-\ell_n} S + \frac{n}{n-\ell_n}\left(T - \frac{\hat{\lambda}_{\text{RE}}}{\hat{a}_{\text{RE}}' S^{-1} \hat{a}_{\text{RE}}} \hat{a}_{\text{RE}} \hat{a}_{\text{RE}}'\right), \qquad \hat{a}_{\text{RE}} = \begin{pmatrix} \hat{\beta}_{\text{RE}} \\ 1 \end{pmatrix}.
\end{aligned}$$

(ii) *If $m_{\max} > k_n/n$, the (1,1) element of the inverse Hessian of the* RE *likelihood* (16)*, evaluated at $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$, is given by:*

$$\hat{\mathcal{H}}^{11}_{\text{RE}} = \frac{\hat{b}_{\text{RE}}' \hat{\Omega}_{\text{RE}} \hat{b}_{\text{RE}} (\hat{\lambda}_{\text{RE}} + k_n/n)}{n \hat{\lambda}_{\text{RE}}} \left(\hat{Q}_{\mathcal{S}} \hat{\Omega}_{\text{RE},22} - T_{22} + \frac{\hat{c}}{1-\hat{c}} \frac{\hat{Q}_{\mathcal{S}}}{\hat{a}_{\text{RE}}' \hat{\Omega}_{\text{RE}}^{-1} \hat{a}_{\text{RE}}} \right)^{-1},$$

where $\hat{Q}_{\mathcal{S}} = Q_{\mathcal{S}}(\hat{\beta}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$, $\hat{c} = \frac{\hat{\lambda}_{\text{RE}} \hat{Q}_{\mathcal{S}}}{(k_n/n + \hat{\lambda}_{\text{RE}})(1-\ell_n/n)}$, and $\hat{b}_{\text{RE}} = (1, -\hat{\beta}_{\text{RE}})'$.

(iii) *Under Assumptions PR, N and MI, $-n\hat{\mathcal{H}}^{11}_{\text{RE}} \xrightarrow{p} \mathcal{V}_{\text{LIML},N}$, with $\mathcal{V}_{\text{LIML},N}$ given in Equation* (9)*.*

Part (i) of Proposition 1 formalizes the claim that the estimator of β remains unchanged under the additional chi-square prior for λ_n. Part (ii) derives the expression for the inverse Hessian. The condition $m_{\max} > k_n/n$ makes sure that the constraint $\lambda \geq 0$ does not bind, otherwise the Hessian is singular. It holds with probability approaching one under Assumption MI (iv), as $m_{\max} - k_n/n \xrightarrow{p} \lambda > 0$. Part (iii) proves that the extra prior on λ_n gets dominated in large samples so that the inverse Hessian can be used to estimate the asymptotic variance of $\hat{\beta}_{\text{LIML}}$ (one could also use the inverse Hessian of the invariant likelihood, although this involves numerical optimization since maximum invariant likelihood estimates of λ_n and Ω are not available in closed form). It is important that the prior on λ_n is chosen such that the prior is dominated in large samples. For example, Lancaster (2002) suggests integrating the orthogonalized incidental parameters out with respect to a uniform prior. Here such prior

corresponds to a flat prior on η_n, which is equivalent to a uniform prior on ω_n, and an improper prior on λ_n, obtained by taking the limit of (14) as $\lambda \to \infty$. However, this improper prior on λ_n will never get dominated by the data, and as a result, it can be shown that the resulting likelihood will fail to produce valid confidence intervals.

4 Minimum distance estimation and inference

In this section, I first show that the random effects estimator is in fact equivalent to a minimum distance estimator that uses a particular weight matrix. This weight matrix weights the restrictions efficiently under normality, but not otherwise. I derive a new estimator of β based on the efficient weight matrix that is more efficient than LIML when the normality assumption is dropped. Moreover, unlike inference based on the random effects likelihood, minimum-distance-based inference will be asymptotically valid even if the reduced-form errors are not normally distributed.

To simplify the expressions in this section, let $D_d \in \mathbb{R}^{d^2 \times d(d+1)/2}$, $L_d \in \mathbb{R}^{d(d+1)/2 \times d^2}$, and $N_d \in \mathbb{R}^{d^2 \times d^2}$ denote the duplication matrix, the elimination matrix, and the symmetrizer matrix, respectively (see Magnus and Neudecker (1980) for definitions of these matrices). The symmetrizer matrix has the property that for any $d \times d$ matrix A, $N_d \operatorname{vec}(A) = (1/2)\operatorname{vec}(A + A')$. The duplication matrix transforms the vech operator into a vec operator, and the elimination operator performs the reverse operation, so that for a symmetric $d \times d$ matrix A, $D_d \operatorname{vech}(A) = \operatorname{vec}(A)$, and $L_d \operatorname{vec}(A) = \operatorname{vech}(A)$.

4.1 Random effects and minimum distance

The random effects likelihood (16) and the minimum distance objective function (12) both leverage the rank restriction (11) to construct an estimator of β. There should therefore exist a weight matrix such that the random effects estimator of $(\beta, \Xi_{22,n})$ is asymptotically equivalent to a minimum distance estimator with respect to this weight matrix. The next proposition shows that if the weight matrix is appropriately chosen, the minimum distance and random effects estimators are in fact *identical*.

Proposition 2. Consider the minimum distance objective function (12) *with respect to the weight matrix* $\hat{W}_{\text{RE}} = D_2'(S^{-1} \otimes S^{-1})D_2$. ***(i)*** *The objective function is minimized at* $(\hat{\beta}_{\text{RE}}, \hat{\Xi}_{22,\text{RE}})$*, where* $\hat{\Xi}_{22,\text{RE}} = \hat{\lambda}_{\text{RE}}/(\hat{a}_{\text{RE}}'\hat{\Omega}_{\text{RE}}^{-1}\hat{a}_{\text{RE}})$ ***(ii)*** *Under Assumptions PR, N and MI, the weight matrix* $\hat{W}_{\text{RE}}$ *is asymptotically optimal.*

The second part of Proposition 2 shows that if the errors are normally distributed, then the random effects weight matrix $\hat{W}_{\text{RE}}$ weights the moment condition (10b) efficiently under many-instrument asymptotics, even though $\hat{W}_{\text{RE}}$ doesn't converge to the inverse of the asymptotic variance of the moment condition. The proof shows that the inverse of the asymptotic variance is not the unique optimal weight matrix, but that there exists a whole class of optimal weight matrices, and that this class includes $\hat{W}_{\text{RE}}$.[6] As I show in the next subsection, this optimality result is sensitive to Assumption N.

The equivalence between minimum distance and RE estimators is related to the observation in Bekker (1994) that LIML can be thought of as a method-of-moments estimator in the sense that it satisfies $(T - m_{\min}S)(1, -\hat{\beta}_{\text{LIML}})' = 0$, which is similar to a first-order condition of the objective function (12) when the weight $\hat{W}_{\text{RE}}$ is used. It is also related to Goldberger and Olkin (1971), who consider a minimum distance objective function based on the proportionality restriction PR,

$$\mathcal{Q}_{\text{GO},n}(\beta, \pi_{2,n}) = \text{vec}\left(\hat{\Pi} - \pi_{2,n}a'\right)' \left(S^{-1} \otimes I_{k_n}\right) \text{vec}\left(\hat{\Pi} - \pi_{2,n}a'\right). \tag{17}$$

Goldberger and Olkin (1971) show that this objective function is minimized at $\hat{\beta}_{\text{LIML}}$. However, the number of parameters in this objective function diverges to infinity under Assumption MI, so it cannot be used for inference.

4.2 Minimum distance estimation under non-normal errors

The efficiency of $\hat{\beta}_{\text{LIML}}$ as well as the expression for the asymptotic distribution of $\hat{\beta}_{\text{LIML}}$ given in (9) depend on Assumption N. This sensitivity to the normality assumption is similar to

[6] The standard condition that the weight matrix converges to the inverse of the asymptotic covariance matrix of the moment conditions is sufficient, but not necessary for asymptotic efficiency (Newey and McFadden, 1994, Section 5.2).

the efficiency results for the maximum likelihood estimator in panel-data models in which identification is based on covariance restrictions (Arellano, 2003, Chapter 5.4).

In order to derive the optimal weight matrix as well as the correct asymptotic variance formulae under non-normality, we first need the limiting distribution of the moment condition (10b). The moment condition depends on the data through the three-dimensional statistic $\text{vech}(T - (k_n/n)S)$. It can be seen from the definition of T and S given in Equations (5) and (6) that this statistic can be written as a quadratic form

$$T - \frac{k_n}{n}S = \frac{1}{n}Y'HY = \frac{1}{n}(Z\pi_{2,n}a' + V)'H(Z\pi_{2,n}a' + V),$$

where

$$H = ZZ' - \frac{k_n}{n - k_n - \ell_n}(I_n - W(W'W)^{-1}W' - ZZ').$$

We need to impose some regularity conditions on the components of the quadratic form. Let $\text{diag}(A)$ denote the n-vector consisting of diagonal elements of an n-by-n matrix A, and let $\delta_n = \text{diag}(H)'\,\text{diag}(H)/k_n$.

Assumption RC (Regularity conditions). **(i)** The errors v_i are i.i.d, with finite 8th moments; **(ii)** For some $\delta, \mu \in \mathbb{R}$, as $n \to \infty$, $\delta_n \to \delta$, and $\pi'_{2,n}Z'\,\text{diag}(H)/\sqrt{nk_n} \to \mu$

Part (i) relaxes the normality assumption on the errors. Part (ii) ensures the asymptotic covariance matrix is well-defined.

Lemma 2. Under Assumptions PR, MI and RC:

(i) $\sqrt{n}\,\text{vech}(T - (k_n/n)S - \Xi_{22,n}aa') \Rightarrow \mathcal{N}(0, \Delta)$, *with* $\Delta = L_2(\Delta_1 + \Delta_2 + \Delta_3 + \Delta_3')L_2'$, *where*

$$\begin{aligned}
\Delta_1 &= 2N_2\left(\Xi_{22}aa' \otimes \Omega + \Omega \otimes \Xi_{22}aa' + \tau\Omega \otimes \Omega,\right), \qquad \tau = \alpha_k(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell),\\
\Delta_2 &= \alpha_k\delta\left[\Psi_4 - \text{vec}(\Omega)\,\text{vec}(\Omega)' - 2N_2(\Omega \otimes \Omega)\right], \qquad \Psi_4 = \mathbb{E}[(v_iv_i') \otimes (v_iv_i')],\\
\Delta_3 &= 2N_2(\sqrt{\alpha_k}\mu\Psi_3' \otimes a), \qquad\qquad\qquad\qquad \Psi_3 = \mathbb{E}[(v_iv_i') \otimes v_i],
\end{aligned}$$

and $\Xi_{22} = \lambda/(a'\Omega^{-1}a)$.

(ii) *Let* $M = I_n - ZZ' - W(W'W)^{-1}W'$, *let* $\hat{V} = MY$ *with rows* $\hat{v}_i'$ *denote estimates of the reduced*

form errors, and let $\hat{\pi}_2$ denote the second column of $\hat{\Pi}$. Then

$$\hat{\Psi}_3 = \frac{\sum_i [(\hat{v}_i \hat{v}_i') \otimes \hat{v}_i]}{\sum_{i,j} M_{ij}^3} \xrightarrow{p} \Psi_3,$$
$$\hat{\Psi}_4 = \frac{\sum_i (\hat{v}_i \hat{v}_i') \otimes (\hat{v}_i \hat{v}_i') - \left[\sum_i M_{ii}^2 - \sum_{i,j} M_{ij}^4\right] (2N_2 \hat{\Omega} \otimes \hat{\Omega} + \mathrm{vec}(\hat{\Omega}) \mathrm{vec}(\hat{\Omega})')}{\sum_{i,j} M_{ij}^4} \xrightarrow{p} \Psi_4,$$

and $\hat{\mu} = \hat{\pi}_2 Z' \operatorname{diag}(H) / \sqrt{nk_n} \xrightarrow{p} \mu$.

Part (i) shows that the asymptotic variance consists of three distinct terms. If the errors are normally distributed, then $\Delta_2 = \Delta_3 = 0$. The term Δ_2 accounts for excess kurtosis of the errors, and the term Δ_3 accounts for skewness. Part (ii) provides consistent estimators for the third and fourth moments of the errors, and for μ. Since the probability limits of S and T do not depend on Assumption N, the other components of Δ_1, Δ_2 and Δ_3 can be consistently estimated by $\hat{\beta}_{\text{RE}}$, $\hat{\Omega}_{\text{RE}}$, and $\hat{\Xi}_{22,\text{RE}} = \hat{\lambda}_{\text{RE}} / (\hat{a}_{\text{RE}}' \hat{\Omega}_{\text{RE}}^{-1} \hat{a}_{\text{RE}})$. Therefore, a consistent estimator of the asymptotic covariance matrix Δ is given by

$$\hat{\Delta} = L_2 (\hat{\Delta}_1 + \hat{\Delta}_2 + \hat{\Delta}_3 + \hat{\Delta}_3') L_2', \tag{18}$$

where the terms $\hat{\Delta}_1$, $\hat{\Delta}_2$, and $\hat{\Delta}_3$ are given by replacing β, Ξ_{22}, and Ω in the definitions of Δ_1, Δ_2, and Δ_3 by their random-effects estimators, replacing Ψ_3 and Ψ_4 by $\hat{\Psi}_3$ and $\hat{\Psi}_4$, and replacing δ and μ by δ_n and $\hat{\mu}$.

Inference based on LIML Since $\hat{\beta}_{\text{LIML}}$ is a minimum distance estimator, its asymptotic variance is given by the (1,1) element of the matrix

$$(G'WG)^{-1} G'W\Delta WG (G'WG)^{-1}, \tag{19}$$

where $W = D_2' (\Omega^{-1} \otimes \Omega^{-1}) D_2 = \operatorname{plim} \hat{W}_{\text{RE}}$, and G is the derivative of the moment condition (10b),

$$G = L_2 \begin{pmatrix} \Xi_{22} \left(a \otimes \binom{1}{0} + \binom{1}{0} \otimes a \right) & a \otimes a \end{pmatrix}.$$

This element evaluates as

$$\mathcal{V}_{\text{LIML}} = \mathcal{V}_{\text{LIML},N} + \frac{2\sqrt{\alpha_k}\mu}{\Xi_{22}^2}\mathbb{E}[(v_{2i} - \gamma\epsilon_i)\epsilon_i^2] + \frac{\alpha_k\delta}{\Xi_{22}^2}\mathbb{E}[\epsilon_i^2(v_{2i} - \gamma\epsilon_i)^2 - |\Omega|],$$

where $\epsilon_i = v_{1i} - v_{2i}\beta$ is the structural error, and γ is regression coefficient from projecting v_{2i} onto it,

$$\gamma = (\Omega_{12} - \Omega_{22}\beta)/(b'\Omega b), \tag{20}$$

The term $\mathcal{V}_{\text{LIML},N}$ (given in Equation (9)) corresponds to the asymptotic variance of $\hat{\beta}_{\text{LIML}}$ under normal errors. The two remaining terms are corrections for skewness and excess kurtosis. Anatolyev (2013) derives the same asymptotic variance expression by working with the explicit definition of $\hat{\beta}_{\text{LIML}}$. If $\alpha_\ell = 0$, then $\mathcal{V}_{\text{LIML}}$ reduces to the asymptotic variance given in Hansen *et al.* (2008), Anderson *et al.* (2010), and van Hasselt (2010). Due to the presence of the two extra terms, the inverse Hessian will no longer estimate the asymptotic variance consistently. However, a consistent plug-in estimator of (19) can easily be computed by replacing Δ by $\hat{\Delta}$ and replacing a, Ξ_{22}, and Ω in the expressions for G and W by $\hat{a}_{\text{RE}}$, $\hat{\Xi}_{22,\text{RE}}$ and $\hat{\Omega}_{\text{RE}}$.

Efficient minimum distance estimator Using the inverse of the variance estimator (18) as a weight matrix in the minimum distance objective function yields an efficient minimum distance (EMD) estimator

$$(\hat{\beta}_{\text{EMD}}, \hat{\Xi}_{22,\text{EMD}}) = \underset{\beta,\Xi_{22}}{\operatorname{argmin}}\, \mathcal{Q}_n(\beta, \Xi_{22,n}; \hat{\Delta}^{-1}).$$

Since the objective function is a fourth-order polynomial in two arguments, the solution can be easily found numerically. It then follows by standard arguments (see, for example, Newey and McFadden (1994)), that when $\alpha_k > 0$,

$$\sqrt{n}(\hat{\beta}_{\text{EMD}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_{\text{EMD}}),$$

where $\mathcal{V}_{\text{EMD}}$ corresponds to (1,1) element of the matrix $(G'\Delta^{-1}G)^{-1}$, which evaluates as

$$\mathcal{V}_{\text{EMD}} = \mathcal{V}_{\text{LIML}} - \frac{1}{\Xi_{22}^2(b'\Omega b)^2}\frac{\left(\sqrt{\alpha_k}\mu\mathbb{E}[\epsilon_i^3] + \alpha_k\delta\mathbb{E}[(v_{2i} - \gamma\epsilon_i)\epsilon_i^3]\right)^2}{2\tau + \alpha_k\delta\kappa}, \tag{21}$$

where

$$\kappa = \mathbb{E}[\epsilon_i^4 / (b'\Omega b)^2 - 3] \tag{22}$$

measures excess kurtosis of ϵ_i. A consistent plug-in estimator of $\mathcal{V}_{\text{EMD}}$ can be easily constructed by replacing Δ by $\hat{\Delta}$, and replacing Ξ_{22} and β in the expression for G by their random-effects, or EMD estimators.

There is a slightly stronger sense in which $\hat{\beta}_{\text{EMD}}$ is efficient than just being efficient in the class of minimum distance estimators: it exhausts the information available in (S, T). In particular, as argued in van der Ploeg and Bekker (1995), the efficiency bound for estimators that are smooth functions of (S, T) is given by the efficient minimum distance estimator based on the moment conditions (10a) and (10b). However, since the nuisance parameter Ω only appears in the first moment condition (10a), which is unrestricted, we can exclude it from the objective function, and the minimum distance estimator of β with respect to an efficient weight matrix will achieve the same asymptotic variance (Chamberlain, 1982, Section 3.2).

Hahn (2002) shows that when the errors are restricted to be normal, an estimator that exhausts the information in (S, T) will have variance given by $\mathcal{V}_{\text{LIML}}$. Anderson *et al.* (2010) generalize this result by allowing the errors to belong to the family of elliptically contoured distributions.[7] Equation (21) shows that this is not true in general. Indeed, the Anderson *et al.* (2010) result obtains as a special case, since for elliptically contoured distributions, $\Psi_3 = 0$, so that $\mathbb{E}[\epsilon_i^3] = 0$ and Ψ_4 is proportional to $\text{vec}(\Omega)\,\text{vec}(\Omega)' + 2N_2\Omega \otimes \Omega$ (Wong and Wang, 1992), which implies $\mathbb{E}[(v_{2i} - \gamma\epsilon_i)\epsilon_i^3] = 0$, so that the second term in (21)—the efficiency gain over LIML—equals zero.

The other special case in which the efficiency gain is zero is when $\delta = 0$, which by the Cauchy-Schwarz inequality, $\mu^2 \leq \delta\Xi_{22}$, implies $\mu = 0$. The term δ measures the balance of the design matrices Z and W. If the diagonal elements of the projection matrices $(ZZ')_{ii}$ and $(W(W'W)^{-1}W)_{ii}$, called the leverage of i, are constant across i, then $\delta_n = 0$, and δ_n, and hence δ, generally increases with the variability of the leverages. Suppose, for instance, that

[7] A mean-zero random vector has an elliptically contoured distribution if its characteristic function can be written as $\varphi(t'\mathcal{V}t)$, for some matrix $\mathcal{V}$ and some function φ. The multivariate normal distribution is a special case, with $\varphi(t) = e^{-t't/2}$.

each observation i belongs to one of $k_n + 1$ groups, numbered $0, \ldots, k_n$, and that there are n_j observations in group j. Let the k_n-vector of instruments z_i^* correspond to a vector of group indicators, with group 0 excluded, and suppose that the only covariate is an intercept.[8] Then, $(W(W'W)^{-1}W')_{ii} = 1/n$, and $(ZZ')_{ii} = 1/n_{j(i)} - 1/n$, where $j(i)$ denotes the group that i belongs to (see supplemental appendix for derivation). It then follows from the definition of δ_n that $\delta_n = 0$ if and only if all groups have equal size, and that the magnitude of δ_n increases with $\sum_{j=0}^{k_n} 1/n_j$, which can be thought of as a measure of group size variability.

Recently, Cattaneo, Crump and Jansson (2012), in the context of few strong instrument asymptotics, proposed a modification of LIML that is more efficient than LIML when the distribution of the reduced-form errors is not normal. The modification was to use a more efficient estimator of the reduced-form coefficients Π_n than $\hat{\Pi}$. Cattaneo *et al.* (2012) use a two-step estimator that uses a kernel estimator of the distribution of the reduced-form errors in the first step Under Assumption MI, when the number of regressors in the reduced-form regression increases with sample size however, this kernel estimator will not be consistent, and so this estimator is unlikely to perform well in settings with many instruments. In contrast, $\hat{\beta}_{\text{EMD}}$ uses the same estimator $\hat{\Pi}$ of Π_n as LIML, but combines the information about β in $\hat{\Pi}$ in a more efficient way. On the other hand, $\hat{\beta}_{\text{EMD}}$ requires $\alpha_k > 0$ for the efficiency gain to be non-zero.

5 Minimum distance estimation without rank restriction

Assumption PR implies that the matrix Ξ_n is reduced rank. In particular, it implies that there are two sources of information for estimating β,

$$\Xi_{11,n} = \Xi_{12,n}\beta, \qquad \text{and} \tag{23}$$

$$\Xi_{12,n} = \Xi_{22,n}\beta. \tag{24}$$

[8] This setup arises when individuals are randomly assigned to groups. For example, if a defendant is randomly assigned to one of $k_n + 1$ judges who differ in their sentencing severity, then one can use judge indicators as instruments for the length of sentence of the defendant, as in Aizer and Doyle (2015) or Dobbie and Song (2015).

The minimum distance objective function (12) weights both sources of identification. In this section, I consider estimation without imposing the rank restriction that $\Xi_{11,n}/\Xi_{12,n} = \Xi_{12,n}/\Xi_{22,n}$. I show that a version of the bias-corrected two-stage least squares estimator (Nagar, 1959; Donald and Newey, 2001) is equivalent to a minimum distance estimator that only uses (24) to estimate β, and derive standard errors that remain valid when (23) does not hold.

5.1 Motivation for relaxing the rank restriction

There are two important cases in which the ratios $\Xi_{12,n}/\Xi_{22,n}$ and $\Xi_{11,n}/\Xi_{12,n}$, which correspond to estimands of the reverse two-stage least squares and two-stage least squares estimators under standard asymptotics (Kolesár, 2013), are not necessarily equal to each other, but $\Xi_{12,n}/\Xi_{22,n}$, is still of interest.

The first case arises when the effect of x_i on y_i is heterogeneous, as in Imbens and Angrist (1994). Let $y_i(x)$ denote the potential outcome of individual i when assigned $x_i = x$, and similarly let $x_i(z)$ denote the potential value of the endogenous variable if the individual was assigned $z_i = z$. We observe $y_i = y_i(x_i)$ and $x_i = x_i(z_i)$. For simplicity, suppose there are no regressors w_i beyond a constant. Suppose that **(i)** z_i affects the outcome only through its effect on x_i: $\{y_i(x)\}_{x\in\mathcal{X}}$ is independent of z_i, where $\mathcal{X}$ denotes the support of x_i; and **(ii)** Monotonicity holds: for any pair (z_1, z_2), $\mathbb{P}(x_i(z_1) \geq x_i(z_2))$ equals either zero or one. Then $\Xi_{12,n}/\Xi_{22,n}$ can be written as a particular weighted average of average partial derivatives $\beta(z) = \mathbb{E}[\partial y_i(x_i(z))/\partial x]$ (see Angrist and Imbens (1995) and Angrist *et al.* (2000) for details). However, unless $\beta(z)$ is constant, the rank restriction will not hold, and the ratio $\Xi_{11,n}/\Xi_{12,n}$ may be outside of the convex hull of the average partial derivatives (Kolesár, 2013), which makes it hard to interpret.

The second case arises when instruments have a direct effect on the outcome. In this case, the coefficient $\pi_{1,n}$ in the reduced-form regression of the outcome on instruments is given by $\pi_{1,n} = \pi_{2,n}\beta + \beta_n^z$, where β_n^z measures the strength of the direct effect. The structural equation (3) no longer holds—instead we have $y_i = x_i\beta + w_i'\beta_n^w + z_i'\beta_n^z + \epsilon_i$. Without any restrictions on β_n^z, the parameter β is no longer identified. However, Kolesár *et al.* (2015) show that if the direct effects are orthogonal to the effects of the instruments on the endogenous variable in the sense that

$\pi_{2,n}'\beta_n^z/n \to 0$ as $n \to \infty$, then β can still be consistently estimated. In particular, under this condition $\Xi_{12,n}/\Xi_{22,n} = \beta + \beta_n^{z\prime}\pi_{2,n}/\pi_{2,n}'\pi_{2,n} \to \beta$. In contrast, $\Xi_{11,n}/\Xi_{12,n} \to \beta$ only if direct effects disappear asymptotically so that $\beta_n^{z\prime}\beta_n^z/n \to 0$.

5.2 Unrestricted minimum distance estimation

To relax the rank restriction on Ξ_n, define β_n simply as the ratio $\Xi_{12,n}/\Xi_{22,n}$, and consider the objective function

$$\mathcal{Q}_n^{\text{UMD}}(\beta, \Xi_{11}, \Xi_{22}; \hat{W}_n) = \\ \operatorname{vech}\left(T - (k_n/n)S - \begin{pmatrix} \Xi_{11} & \Xi_{22}\beta \\ \Xi_{22}\beta & \Xi_{22} \end{pmatrix}\right)' \hat{W}_n \operatorname{vech}\left(T - (k_n/n)S - \begin{pmatrix} \Xi_{11} & \Xi_{22}\beta \\ \Xi_{22}\beta & \Xi_{22} \end{pmatrix}\right), \quad (25)$$

where $\hat{W}_n \in \mathbb{R}^{3\times 3}$ is some weight matrix. If we restrict $\Xi_{11,n}$ to equal to $\Xi_{22,n}\beta_n^2$, then minimizing this objective function is equivalent to minimizing the original objective function (12). If $\Xi_{11,n}$ is unrestricted, the weight matrix does not matter since then the model is exactly identified. The unrestricted minimum distance estimators will be given by their sample counterparts,

$$\hat{\Xi}_{22,\text{UMD}} = T_{22} - (k_n/n)S_{22}, \qquad \hat{\Xi}_{11,\text{UMD}} = T_{11} - (k_n/n)S_{11},$$

and

$$\hat{\beta}_{\text{UMD}} = \frac{T_{12} - (k_n/n)S_{12}}{T_{22} - (k_n/n)S_{22}}.$$

The unrestricted minimum distance estimator for β_n coincides with the modified bias-corrected two-stage least squares estimator (Kolesár *et al.*, 2015), a version of the bias-corrected two-stage least squares estimator. The version proposed by Donald and Newey (2001) multiplies S_{12} and S_{22} by $\frac{k_n-2}{n}\frac{n-k_n-\ell_n}{n-k_n+2}$ instead of k_n/n. The motivation for the version in Kolesár *et al.* (2015) was to modify the Donald and Newey estimator to make it consistent when $\alpha_\ell > 0$. However, it can also be viewed as a minimum distance estimator that puts no restrictions on the reduced form. The next proposition derives its large-sample properties.

Proposition 3. Suppose that Assumption MI(i)–(iii) and Assumption RC hold, $\Xi_n = \Xi + o(1)$ *where* Ξ *is a positive semi-definite matrix with* $\Xi_{22} > 0$*, and that* $(\pi_{1,n} - \pi_{2,n}\beta_n)'Z' \operatorname{diag}(H)/\sqrt{nk_n} =$

$\tilde{\mu} + o(1)$ *for some* $\tilde{\mu}$*. Then*

$$\sqrt{n}\left(\hat{\beta}_{\text{UMD}} - \beta_n\right) \Rightarrow \mathcal{N}(0, V_{\text{UMD}}),$$

where,

$$V_{\text{UMD}} = V_{\text{EMD}} + V_\Delta + \frac{|\Xi|\Omega_{22}/\Xi_{22} + 2\sqrt{\alpha_k}\tilde{\mu}\mathbb{E}[v_{2i}^2\epsilon_i]}{\Xi_{22}^2}$$
$$V_\Delta = \frac{\left((2\tau + \alpha_k\delta\kappa)\gamma(b'\Omega b)^2 + \sqrt{\alpha_k}\mu\mathbb{E}[\epsilon_i^3] + \alpha_k\delta\mathbb{E}[\epsilon_i^3(v_{2i} - \gamma\epsilon_i)]\right)^2}{(b'\Omega b)^2(2\tau + \alpha_k\delta\kappa)\Xi_{22}^2}$$

where κ, γ *are defined as in* (20) *and* (22)*, with* $\beta = \Xi_{12}/\Xi_{22}$*, and* $\epsilon_i = v_{2i} - v_{1i}\beta$*.*

The asymptotic variance V_{UMD} corresponds to the (1,1) element of the matrix $G_{\text{UMD}}^{-1}\Delta_{\text{UMD}}G_{\text{UMD}}^{-1\prime}$, where G_{UMD} is the derivative of the moment condition,

$$G_{\text{UMD}} = \begin{pmatrix} 1 & 1 & 0 \\ \Xi_{22} & 0 & \beta \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{so that} \quad G_{\text{UMD}}^{-1\prime}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{\Xi_{22}}(0, 1, -\beta),$$

and, as shown in the proof, $\Delta_{\text{UMD}} = L_2(\Delta_{1,\text{UMD}} + \Delta_2 + \Delta_{3,\text{UMD}} + \Delta_{3,\text{UMD}}')L_2'$ is the asymptotic variance of the moment condition (10b), with

$$\Delta_{1,\text{UMD}} = 2N_2(\Xi \otimes \Omega + \Omega \otimes \Xi + \tau\Omega \otimes \Omega), \qquad \Delta_{3,\text{UMD}} = 2\sqrt{\alpha_k}N_2\left(\Psi_3' \otimes \begin{pmatrix} \tilde{\mu} + \mu\beta \\ \mu \end{pmatrix}\right),$$

and Δ_2 given in Lemma 2. If $\Xi = \Xi_{22}aa'$, then the expressions for $\Delta_{1,\text{UMD}}$ and $\Delta_{3,\text{UMD}}$ reduce to those for Δ_1 and Δ_3 given in Lemma 2.

The asymptotic variance consists of three components. The first term coincides with the asymptotic variance of EMD given in Equation (21). The second component, V_Δ, represents the asymptotic efficiency loss relative to $\hat{\beta}_{\text{EMD}}$ when the rank restriction holds; it quantifies the price for not using information contained in (23) when the rank restriction holds. Unlike the efficiency loss of LIML, the term is positive even when the errors are normal, in which case it simplifies to $2\tau(\Omega_{12} - \Omega_{22}\beta)^2/\Xi_{22}^2$, which is only zero if there is no endogeneity (that is, $\mathbb{E}[x_i\epsilon_i] = 0$). Finally, the last component represents the increase in asymptotic variance due to

the failure of rank restriction; when PR holds, $|\Xi| = 0$ and $\tilde{\mu} = 0$, and this term drops out.

The asymptotic variance can be easily consistently estimated by

$$\hat{V}_{\text{UMD}} = \frac{1}{\hat{\Xi}^2_{22,\text{UMD}}}(0, 1, -\hat{\beta}_{\text{UMD}})'\hat{\Delta}_{\text{UMD}}(0, 1, -\hat{\beta}_{\text{UMD}}),$$

where $\hat{\Delta}_{\text{UMD}}$ is a plug-in estimator based on $\hat{\Xi}_{\text{UMD}} = T - k_n/nS$, $\hat{\beta}_{\text{UMD}}$, $\hat{\Omega} = S$, and estimators of Ψ_3 and Ψ_4 given in Lemma 2. Confidence intervals based on $\hat{\beta}_{\text{UMD}}$ and $\hat{V}_{\text{UMD}}$ will then be robust to both many instruments, and failure of the proportionality restriction (1).

It is possible to reduce the asymptotic mean-squared error of the minimum distance estimator by minimizing the minimum distance objective function subject to the constraint that Ξ_n be positive semi-definite (which has to be the case since Ξ_n is a matrix of second moments of $\pi_{2,n}$ and $\pi_{1,n}$), which is equivalent to the constraint $\Xi_{11,n} \geq \beta_n^2 \Xi_{22,n}$. If the weight matrix $\hat{W}_n$ is used, then the resulting estimator will be a mixture between $\hat{\beta}_{\text{UMD}}$, and the restricted minimum distance estimator that minimizes (12) with respect to $\hat{W}_n$: when $T - (k_n/n)S$ is positive semi-definite, then the estimator equals $\hat{\beta}_{\text{UMD}}$; otherwise, the minimum distance objective is minimized at a boundary, and the estimator equals the restricted minimum distance estimator. When Ξ_n is full rank, then the constraint won't bind in large samples, and the estimator will be asymptotically equivalent to $\hat{\beta}_{\text{UMD}}$. However, when Ξ_n is reduced-rank, the mixing will deliver a smaller asymptotic mean-squared error. The disadvantage is that the estimator will be asymptotically biased, which makes inference more complicated. See supplemental appendix for details.

6 Tests of overidentifying restrictions

The proportionality restriction PR is testable. In this section, I discuss a simple test based on the minimum distance objective function, and compare it to some alternatives previously proposed in the literature.

In the invariant model, testing Assumption PR is equivalent to testing the null that Ξ_n is reduced-rank against the alternative that it is positive definite. A simple way to implement the test is to compare the value of the minimum distance objective function (25) minimized subject

to the restriction that $|\Xi_n|$ is reduced rank with its value when it is minimized subject to $|\Xi_n|$ being positive semi-definite. When $\hat{W}_{\text{RE}}$ is used as a weight matrix, the test statistic is given by (see supplemental appendix for derivation)

$$\begin{aligned}\hat{J}_{\text{MD}} &= \min_{\Xi_{11}=\Xi_{22}\beta^2} \mathcal{Q}_n^{\text{UMD}}(\beta, \Xi_{11}, \Xi_{22}; \hat{W}_{\text{RE}}) - \min_{\Xi_{11}\geq\Xi_{22}\beta^2} \mathcal{Q}_n^{\text{UMD}}(\beta, \Xi_{11}, \Xi_{22}; \hat{W}_{\text{RE}}) \\ &= \begin{cases} 0 & \text{if } m_{\min} \leq k_n/n, \\ (m_{\min} - k_n/n)^2 & \text{otherwise.} \end{cases}\end{aligned}$$

The test statistic depends on the data through the minimum eigenvalue of $S^{-1}T$. Because the weight matrix $\hat{W}_{\text{RE}}$ is not optimal, the large-sample distribution of $\hat{J}_{\text{MD}}$ is not pivotal under the null: if $k_n \to \infty$ as $n \to \infty$, then in large samples, $n^2\hat{J}_{\text{MD}}/k_n$ will be distributed as a mixture between a χ^2_1 distribution scaled by $\frac{2(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell} + \delta\kappa$ (with κ defined in Equation (22)), and a degenerate distribution with a point mass at 0. One solution would be to divide the test statistic by $\frac{2(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell} + \delta\kappa$ and use a critical value based on the 90% quantile of a χ^2_1 distribution, or, equivalently, reject whenever $(n/\sqrt{k_n})(m_{\min} - k_n/n)/\sqrt{\frac{2(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell} + \delta\kappa}$ is greater the 95% quantile of a standard normal distribution. However, since the asymptotic distribution changes when k_n is fixed, this won't yield valid inference when k_n is fixed. Using arguments similar to Anatolyev and Gospodinov (2011), the next proposition proposes a modification that ensures size control whether k_n is fixed or grows with the sample size.

Proposition 4. Suppose Assumptions PR, MI and RC hold. Suppose also that if $k_n = K$ is fixed, then $\sup_{n\geq 1} \max_{i=1,\ldots,n}(ZZ')^2_{ii} = o(1)$. Then:

- ***(i)*** *If $k_n \to \infty$, $\frac{n}{\sqrt{k_n}}(m_{\min} - k_n/n) \Rightarrow \mathcal{N}(0, \frac{2(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell} + \delta\kappa)$. If $k_n = K$ is fixed, then $nm_{\min} \Rightarrow \chi^2_{K-1}$.*
- ***(ii)*** *Let $\delta_n = \text{diag}(H)'\,\text{diag}(H)/k_n$, let $\hat{\kappa} = (\hat{b}_{\text{RE}} \otimes \hat{b}_{\text{RE}})'\hat{\Psi}_4(\hat{b}_{\text{RE}} \otimes \hat{b}_{\text{RE}})/(\hat{b}'_{\text{RE}}S\hat{b}_{\text{RE}})^2 - 3$, and let Φ denote cdf of a standard normal distribution. The test that rejects whenever $nm_{\min}$ is greater than the*

$$1 - \Phi\left(\sqrt{\tfrac{(n-k_n)}{n-k_n-\ell_n} + \tfrac{\delta_n\hat{\kappa}}{2}} \cdot \Phi^{-1}(ns)\right)$$

quantile of the $\chi^2_{k_n-1}$ distribution has asymptotic size equal to ns. This holds whether $k_n = K$ is fixed or $k_n \to \infty$.

When $k_n \to \infty$, the test is asymptotically equivalent to the test proposed in the previous paragraph. However, unlike that test, it also remains valid under the few strong instrument asymptotics with k_n fixed. In this case, it is asymptotically equivalent to the Cragg and Donald (1993) test, which is based on the minimum distance objective function (17), and rejects whenever $nm_{\min}$ is greater than the $1-$ns quantile of $\chi^2_{k_n}$ distribution. The test can therefore be interpreted as a Cragg-Donald test with a modified critical value that ensures size control under few strong, as well as many-instrument asymptotics.

It is interesting to compare this test to some other tests proposed in the literature. In the context of few strong instrument asymptotics, the most popular test is due to Sargan (1958). The test statistic can be written as $\hat{J}_{\text{S}} = \frac{m_{\min}}{1-k_n/n-\ell_n/n+m_{\min}}$, and the critical value is given by $1-$ns quantile of $\chi^2_{k_n}$. Anatolyev and Gospodinov (2011) show that if $\alpha_k > 0$ and $\alpha_\ell = 0$ and the errors are normal, the Sargan test is mildly conservative. With $\alpha_k = 0.1$ for example, the asymptotic size of the test with nominal size 0.05 is given by 0.04. Anatolyev and Gospodinov (2011) therefore propose an adjustment to the critical value similar to the one proposed here to match the asymptotic size with the nominal size. Unfortunately, this solution is not robust to allowing the number of exogenous regressors to increase with the sample size: if $\alpha_\ell > 0$, the asymptotic size of the Sargan test converges to one (see supplemental appendix for details). Lee and Okui (2012) propose a different modification of the Sargan test that controls size under conditions similar to Proposition 4, provided that, in addition, $\alpha_\ell = 0$ and $k_n \to \infty$. In contrast, the test proposed here will work irrespective of the number of regressors or instruments; the researcher doesn't have to determine what type of asymptotics are appropriate.

Another alternative to the test in Proposition 4 would be to use the efficient weight matrix instead of $\hat{W}_{\text{RE}}$ in the minimum distance objective function. Such a test would in general direct local asymptotic power to different alternatives, and, without specifying which local violations of the proportionality restriction are of interest, it is unclear which test should be preferred. However, an attractive feature of the test in Proposition 4 is its easy implementation, which only requires modifying the critical value of the Cragg-Donald test.

7 Conclusion

In this paper, I outlined a minimum distance approach to inference in a linear instrumental variables model with many instruments. I showed how estimation and inference based on the minimum distance objective function solves the incidental parameters problem that the large number of instruments create. When the efficient weight matrix is used, I obtain a new estimator that is in general more efficient than LIML. Moreover, depending on the weight matrix used, and whether a proportionality restriction on the reduced-form coefficients is imposed, the bias-corrected two-stage least squares estimator, the LIML estimator, and the random-effects estimator, which is shown to coincide with LIML, are obtained as particular minimum distance estimators. Standard errors can easily be constructed using the usual sandwich formula for asymptotic variance of minimum distance estimators.

The invariance argument underlying the construction of the minimum distance objective function relied on the assumption of homoscedasticity. It would be interesting to explore in future work how this approach can be adapted to deal with heteroscedasticity, and whether similar minimum distance construction can be used in other models with an incidental parameters problem.

Appendix

Appendix A states and proves some auxiliary Lemmata that are helpful for proving the main results. Proofs of lemmata and propositions stated in the text are given in Appendix B. Throughout the appendix, I use the following simple identifies that follow from simple algebra. For any positive definite matrix $\Omega \in \mathbb{R}^{2\times 2}$, vectors $a = (\beta, 1)'$ and $b = (1, -\beta)'$, $\beta \in \mathbb{R}$, and constants c_1, c_2:

$$Q_{\mathcal{S}}(\beta, \Omega) + Q_{\mathcal{T}}(\beta, \Omega) = \operatorname{tr}(\Omega^{-1}T), \tag{26a}$$

$$|\Omega| a\Omega^{-1}a = b'\Omega b, \tag{26b}$$

$$|c_1 T + c_2 S| = (c_1 m_{\max} + c_2)(c_1 m_{\min} + c_2)|S|. \tag{26c}$$

Appendix A Auxiliary Lemmata

Lemma A.1. ***(i)*** *If for some invertible matrix $V \in \mathbb{R}^{d\times d}$, $N_d V = V N_d$, then $(L_d N_d V L_d')^{-1} = D_d' V_d^{-1} D_d$.* ***(ii)*** *For an invertible matrix $V \in \mathbb{R}^{d\times d}$, a vector $m \in \mathbb{R}^d$ and a constant c,*

$$\left(V \otimes V + c(mm') \otimes (mm')\right)^{-1} = V^{-1} \otimes V^{-1} - \frac{c(V^{-1}mm'V^{-1}) \otimes (V^{-1}mm'V^{-1})}{1 + c(m'V^{-1}m)^2},$$

Proof. **(i)** It follows from Lemmata 3.5(i) and 3.6(ii) in Magnus and Neudecker (1980) that $L_d N_d D_d = I_{d(d+1)/2}$. Also, by Lemma 3.5(ii) in Magnus and Neudecker (1980), $D_d L_d N_d = N_d$. Thus, $(D_d' V^{-1} D_d)(L_d N_d V L_d') = D_d' V^{-1} N_d V L_d' = D_d' V^{-1} V N_d L_d' = D_d' N_d L_d' = I_{d(d+1)/2}$. **(ii)** Follows from direct calculation. □

Lemma A.2. Consider the quadratic form $Q_n = (M_n + U_n)' P_n (M_n + U_n)$, where $P_n \in \mathbb{R}^{n\times n}$ is a symmetric matrix with non-random elements, $U_n, M_n \in \mathbb{R}^{n\times G}$, M_n is non-random, and the rows u_{in}' of U_n are i.i.d. with zero mean, variance Ω_n, and finite fourth moments.

Let $\Lambda_n = M_n' P_n P_n M_n$, $\delta_n = \operatorname{diag}(P_n)' \operatorname{diag}(P_n)$, $\overline{m}_n = M_n' P_n \operatorname{diag}(P_n)$, $p_{ij} = (P_n)_{ij}$, and let e_{in} denote an n-vector of zeros with 1 in the ith position. Then

(i) *The variance of Q_n is given by*

$$\operatorname{var}(\operatorname{vec}(Q_n)) = 2N_G \left(\Omega_n \otimes \Lambda_n + \Lambda_n \otimes \Omega_n + \operatorname{tr}(P_n^2)\Omega_n \otimes \Omega_n\right)$$

$$+ \delta_n \left(\mathbb{E}[u_{in}u_{in}' \otimes u_{in}u_{in}'] - \text{vec}(\Omega_n)\,\text{vec}(\Omega_n)' - 2N_G\Omega_n \otimes \Omega_n \right) + 2N_G(\Psi_3' \otimes \overline{m}_n) + 2(\Psi_3' \otimes \overline{m}_n)'N_G,$$

where $\Psi_3 = E[(u_{in}u_{in}') \otimes u_{in}]$*, and the last two lines are equal to zero if the distribution of* u_{in} *is normal.*

(ii) *Suppose that (a)* $\text{var}(\text{vec}(Q_n))$ *converges, and* δ_n *and* Λ_n *are bounded; (b)* $\sup_n \mathbb{E}[\|u_{in}\|^8] < \infty$*; (c)* $\sum_{i=1}^n |p_{ii}|^4 = o(1)$*,* $\sum_{i=1}^n \left(\sum_{j=1}^n p_{ij}^2\right)^2 = o(1)$*, and* $\sum_{i<j<k<\ell} p_{ik}p_{i\ell}p_{jk}p_{j\ell} = o(1)$*; and (d)* $\sum_{i=1}^n \|e_{in}'P_nM_n\|^4 = o(1)$*. Then:*

$$\text{vec}(Q_n - M_n'P_nM_n - \text{tr}(P_n)\Omega_n) \Rightarrow \mathcal{N}(0, \lim_{n\to\infty}(\text{var}(\text{vec}(Q_n)))).$$

Proof. Proof of Part (i) follows from a tedious, but straightforward calculation. Proof of Part (ii) is a generalization of the central limit theorems in Chao *et al.* (2012) and Hansen *et al.* (2008), and is proved using similar arguments. Full proof is given in the supplemental appendix. □

Lemma A.3. Let $P_n = (A_n + \nu_n B_n)/\sqrt{m_n}$*, where* $m_n \to \infty$ *as* $n \to \infty$*,* $\nu_n = O(1)$*,* $A_n, B_n \in \mathbb{R}^{n\times n}$ *are projection matrices such that* $A_nB_n = 0$*, and for* $j > 1$*,* $\text{tr}(A_n/m_n^j) = o(1)$ *and* $\text{tr}(\nu_nB_n/m_n^j) = o(1)$*. Then condition (ii)c of Lemma A.2 holds.*

Proof. Denote the (i,j) elements of A_n and B_n by a_{ij} and b_{ij}. The first condition follows from the bound, for $j > 2$, $\sum_i p_{ii}^j \le 2^{j-1}(\sum_i a_{ii}^j + \nu_n^j \sum_i b_{ii}^j)/m_n^{j/2} \le 2^{j-1}(\sum_i a_{ii} + \nu_n^j \sum_i b_{ii})/m_n^{j/2} = o(1)$ The second condition follows from $\sum_i \left(\sum_j p_{ij}^2\right)^2 \le \sum_i \left(2\sum_j a_{ij}^2 + 2\nu_n^2 \sum_j b_{ij}^2\right)^2/m_n^2 = \sum_i \left(2a_{ii} + 2\nu_n^2 b_{ii}\right)^2/m_n^2 = o(1)$.

It therefore remains to show that $\sum_{i<j<k<\ell} p_{ik}p_{i\ell}p_{jk}p_{j\ell} = o(1)$. This can be shown using arguments similar to those in the proof of Lemma B.2 in Chao *et al.* (2012). Let D denote a diagonal matrix with elements $D_{ii} = (P_n)_{ii}$, let $S_n = \sum_{i<j<k<\ell}(p_{ik}p_{i\ell}p_{jk}p_{j\ell} + p_{ij}p_{i\ell}p_{jk}p_{k\ell} + p_{ij}p_{ik}p_{j\ell}p_{k\ell})$, and let $\|\cdot\|_F$ denote the Frobenius norm. Note that

$$\|(P_n - D)^2\|_F \le \|P_n^2\|_F + \|D^2\|_F + 2\|DP_n\|_F = o(1), \tag{27}$$

where the last equality follows from $\|D^2\|_F^2 = \sum_i p_{ii}^4 = o(1)$, $\|P_n^2\|_F^2 = (\text{tr}\, A_n + \nu_n^4 \,\text{tr}\, B_n)/m_n^2 = o(1)$, and $\|DP_n\|_F^2 = \sum_i p_{ii}^2 \sum_j p_{ij}^2 \le m_n^{-2}\sum_i (a_{ii} + \nu_n b_{ii})(a_{ii} + \nu_n^2 b_{ii}) = o(1)$. On the other hand, expanding the left-hand side in (27) yields

$$\|(P_n - D)^2\|_F^2 = 2\sum_{i<j} p_{ij}^4 + 4\sum_{i<j<\ell}\left(p_{ij}^2p_{i\ell}^2 + p_{ij}^2p_{j\ell}^2 + p_{j\ell}^2p_{i\ell}^2\right) + 8S_n.$$

Since the first four terms are bounded by $\sum_i (\sum_j p_{ij}^2)^2 = o(1)$, it follows that $S_n = o(1)$. Define

$$\Delta_2 = \sum_{i<j<k} \left(p_{ij}p_{ik}\epsilon_j\epsilon_k + p_{ij}p_{jk}\epsilon_i\epsilon_k \right),$$
$$\Delta_3 = \sum_{i<j<k} \left(p_{ik}p_{jk}\epsilon_i\epsilon_j \right),$$

and let $\Delta_1 = \Delta_2 + \Delta_3$. Then

$$\mathbb{E}[\Delta_3^2] = \sum_{i<j<k,\ell} p_{ik}p_{jk}p_{i\ell}p_{j\ell} = \sum_{i<j<k} p_{ik}^2 p_{jk}^2 + 2\sum_{i<j<k<\ell} p_{ik}p_{jk}p_{i\ell}p_{j\ell} = 2\sum_{i<j<k<\ell} p_{ik}p_{jk}p_{i\ell}p_{j\ell} + o(1),$$
$$\mathbb{E}[\Delta_1^2] = \mathrm{tr}((P_n - D)^4) - 2\sum_{i<j} p_{ij}^4 = o(1),$$
$$\mathbb{E}[\Delta_2^2] = \sum_{i<j<k} (p_{ij}^2 p_{ik}^2 + p_{ij}^2 p_{jk}^2) + 2S_n = o(1).$$

Thus, by the Cauchy-Schwarz inequality,

$$\mathbb{E}[\Delta_3^2] \leq 2\mathbb{E}[\Delta_1^2] + 2\mathbb{E}[\Delta_2^2] = o(1),$$

which proves the result. □

Corollary A.1. Consider the model (1)–(2)*, and suppose Assumptions PR, N and MI hold. Then:*

$$\sqrt{n}\,\mathrm{vec}\,(S - \Omega) \Rightarrow \mathcal{N}_4\left(0, \frac{1}{1-\alpha_k-\alpha_\ell} 2N_2(\Omega\otimes\Omega)\right)$$
$$\sqrt{n}\,\mathrm{vec}\left(T - \alpha_k\Omega - \frac{\lambda_n}{a'\Omega^{-1}a}aa'\right) \Rightarrow \mathcal{N}_4\left(0, 2N_2(\alpha_k\Omega\otimes\Omega + \Omega\otimes M + M\otimes\Omega)\right),$$

where $M = \frac{\lambda}{a'\Omega^{-1}a}aa'$.

Proof. The result follows from Lemmata A.2 and A.3, with $P_n = (I - ZZ' - W(W'W)^{-1}W)/\sqrt{n}$, and $P_n = (ZZ')/\sqrt{n}$. □

Corollary A.2. Consider the model (1)–(2)*, and suppose Assumption MI(i)–(iii), and Assumption RC hold,* $\Xi_n = \Xi + o(1)$ *where* Ξ *is a positive semi-definite matrix with* $\Xi_{22} > 0$*, and that* $(\pi_{1,n} - \pi_{2,n}\beta_n)'Z'\,\mathrm{diag}(H)/\sqrt{nk_n} = \tilde{\mu} + o(1)$ *for some* $\tilde{\mu}$*. Let* $\overline{m} = (\tilde{\mu} + \mu(\Xi_{12}/\Xi_{22}), \mu)'$*. Then*

$$\sqrt{n}\,\mathrm{vec}(T - (k_n/n)S - \Xi_n) \Rightarrow \mathcal{N}(0, \Delta),$$

where

$$\Delta = 2N_2 \left(\Omega \otimes \Xi + \Xi \otimes \Omega + (\alpha_k(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell) - \alpha_k\delta)\Omega \otimes \Omega\right) \\ + \alpha_k\delta\left(\mathbb{E}[v_i v_i' \otimes v_i v_i'] - \operatorname{vec}(\Omega)\operatorname{vec}(\Omega)'\right) + \alpha_k^{1/2}(2N_2(\Psi_3' \otimes \overline{m}) + (\Psi_3' \otimes \overline{m})' N_2),$$

and $\Psi_3 = E[(v_i v_i') \otimes v_i]$.

Proof. The result follows from Lemmata A.2 and A.3, with $P_n = H/\sqrt{n}$. □

Appendix B Proofs

Proof of Lemma 1. To ensure that the densities of T and $\hat{\Pi}$ are expressed with respect to compatible dominating measures, I will express the density of T with respect to the measure

$$\mu_T(\mathrm{d}t) = \frac{n^{k_n}\pi^{k_n - 1/2}}{\Gamma(k_n/2)\Gamma((k_n-1)/2)}|t|^{(k_n-3)/2}\lambda(\mathrm{d}t),$$

where λ is the Lebesgue measure on the sample space of T, and Γ denotes the gamma function. μ_T is the measure induced by the Lebesgue measure μ on the sample space of $\hat{\Pi}$ in the sense that for any measurable set B, $\mu_T(B) = \mu(\delta^{-1}(B))$, where $\delta(\hat{\Pi}) = \hat{\Pi}'\hat{\Pi}/n$ is the function defining T (Eaton, 1989, Example 5.1). The statistic T has the same distribution as the statistic W_N in Moreira (2009, Section 4), with the parameters λ_N and Σ in that paper corresponding to $\lambda_n/(a'\Omega^{-1}a)$ and Ω. Hence, by Theorem 4.1 in Moreira (2009), the density of T with respect to μ_T is given by

$$f_T(T \mid \beta, \lambda_n, \Omega) = \mathcal{K}_1 e^{-\frac{n}{2}(\lambda_n + \operatorname{tr}(\Omega^{-1}T))}|\Omega|^{-k_n/2}(n\lambda_n^{1/2}Q_{\mathcal{T}}(\beta,\Omega)^{1/2})^{-\frac{k_n-2}{2}} I_{(k_n-2)/2}(n\lambda_n^{1/2}Q_{\mathcal{T}}(\beta,\Omega)^{1/2}), \tag{28}$$

where $\mathcal{K}_1 = \Gamma(k_n/2)\pi^{-k_n}2^{-k_n/2-1}$ and $I_\nu(\cdot)$ is modified Bessel function of the first kind of order ν. $I_\nu(\cdot)$ has the integral representation (Abramowitz and Stegun, 1965, Equation 9.6.18, p. 376)

$$I_\nu(t) = \frac{(t/2)^\nu}{\pi^{1/2}\Gamma(\nu+1/2)}G_{2\nu+2}(t), \qquad \text{where} \qquad G_k(t) = \int_{[-1,1]} e^{ts}(1-s^2)^{(k-3)/2}\,\mathrm{d}s.$$

The density (28) can therefore be written as:

$$f_T(T \mid \beta, \lambda_n, \Omega) = \frac{2^{-k_n}\Gamma(k_n/2)}{\pi^{k_n+1/2}\Gamma((k_n-1)/2)} \cdot e^{-\frac{n}{2}(\lambda_n + \operatorname{tr}(\Omega^{-1}T))}|\Omega|^{-k_n/2}G_{k_n}(n\lambda_n^{1/2}Q_{\mathcal{T}}(\beta,\Omega)^{1/2}).$$

Combining this expression with the density for S (with respect to Lebesgue measure), which is given by

$$f_S(S;\Omega) = C_{n-k_n-\ell_n} \cdot |S|^{(n-k_n-\ell_n-3)/2}|\Omega|^{-n-k_n-\ell_n/2}e^{-\frac{n-k_n-\ell_n}{2}\operatorname{tr}(\Omega^{-1}S)},$$

where

$$C_\nu^{-1} = (2/\nu)^\nu \pi^{1/2} \Gamma(\nu/2)\Gamma((\nu-1)/2) \tag{29}$$

yields the invariant likelihood

$$\begin{aligned}\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T) &= \frac{2^{-k_n}\Gamma(k_n/2)}{\pi^{k_n+1/2}\Gamma((k_n-1)/2)} \cdot e^{-\frac{n}{2}(\lambda_n + \text{tr}(\Omega^{-1}T))} |\Omega|^{-k_n/2} G_{k_n}(n\lambda_n^{1/2} Q_\mathcal{T}(\beta,\Omega)^{1/2}) \cdot f_S(S;\Omega) \\ &\propto \exp\left[-\frac{1}{2}\left((n-\ell_n)\log|\Omega| + \text{tr}(\Omega^{-1}\tilde{S}) + n\lambda_n - 2\log G_{k_n}(n\sqrt{\lambda_n Q_\mathcal{T}(\beta,\Omega)})\right)\right],\end{aligned} \tag{30}$$

where $\tilde{S} = (n - k_n - \ell_n)S + nT$. The derivative with respect to Ω is given by:

$$\frac{\partial \log \mathcal{L}_{\text{INV},n}}{\partial \Omega} = \frac{1}{2}\Omega^{-1}\left[\tilde{S} - (n-\ell_n)\Omega - \frac{G'_{k_n}(n\sqrt{\lambda_n Q_\mathcal{T}(\beta,\Omega)})}{G_{k_n}(n\sqrt{\lambda_n Q_\mathcal{T}(\beta,\Omega)})} \frac{n\lambda_n^{1/2}}{Q_\mathcal{T}(\beta,\Omega)^{1/2}}\left(T - \frac{Q_\mathcal{S}(\beta,\Omega)}{b'\Omega b}\Omega b b'\Omega\right)\right]\Omega^{-1}, \tag{31}$$

where the derivative $\partial Q_\mathcal{T}(\beta,\Omega)/\partial\Omega$ is computed using the identity (26a). Note that $b'\Omega b > 0$ for Ω positive definite, $Q_\mathcal{T}(\beta,\Omega) > 0$ with probability one for Ω positive definite, and $G_{k_n}(t) > 0$ for $t > 0$ (Abramowitz and Stegun, 1965, p. 374), so that the denominators in (31) are non-zero at any point in the parameter space for $(\beta, \Omega, \lambda_n)$ with probability one.

Fix λ_n. Denote the ML estimates of β and Ω given λ_n by $(\hat{\beta}_{\lambda_n}, \hat{\Omega}_{\lambda_n})$. Since $G(\cdot)$ is a monotone increasing function, it follows from (30) that:

$$\hat{\beta}_{\lambda_n} = \underset{\beta}{\text{argmax}}\, Q_\mathcal{T}(\beta, \hat{\Omega}_{\lambda_n}) = \underset{\beta}{\text{argmin}}\, Q_\mathcal{S}(\beta, \hat{\Omega}_{\lambda_n}). \tag{32}$$

Secondly, the derivative (31) evaluated at $(\hat{\beta}_{\lambda_n}, \hat{\Omega}_{\lambda_n})$ has to be equal to zero. Pre-multiplying and post-multiplying Equation (31) by $\hat{b}'_{\lambda_n}\hat{\Omega}_{\lambda_n}$ and $\hat{\Omega}_{\lambda_n}\hat{b}_{\lambda_n}$ therefore yields

$$(n-\ell_n)\hat{b}'_{\lambda_n}\hat{\Omega}_{\lambda_n}\hat{b}_{\lambda_n} = \hat{b}'_{\lambda_n}\tilde{S}\hat{b}_{\lambda_n}. \tag{33}$$

Therefore,

$$\hat{\beta}_{\lambda_n} = \underset{\beta}{\text{argmin}}\, Q_\mathcal{S}(\beta, \hat{\Omega}_{\lambda_n}) = \underset{\beta}{\text{argmin}}\, Q_\mathcal{S}(\beta, \tilde{S}) = \text{argmin}\, Q_\mathcal{S}(\beta, S) = \hat{\beta}_{\text{LIML}},$$

where the first equality follows by (32), the second by (33), the third by $Q_\mathcal{S}(\beta,\tilde{S})^{-1} = (n-k_n-\ell_n)Q_\mathcal{S}(\beta,S)^{-1} + n$, and the last equality follows from definition of $\hat{\beta}_{\text{LIML}}$. By similar arguments, Equations (32) and (33) must also hold when the likelihood is maximized over λ_n as well, so that $\hat{\beta}_{\text{INV}} = \hat{\beta}_{\text{LIML}}$.

It remains to show (13). This result follows from the fact that F_{ω_n} corresponds to the invariant prior distribution induced by the Haar probability measure ν_H on $\mathcal{O}(k_n)$ (which is unique since $\mathcal{O}(k_n)$ is compact) via the group action $\omega_n \mapsto g\omega_n$, $g \in \mathcal{O}(k_n)$, in the sense that for any measurable set B, $F_{\omega_n}(B) = \nu_H(g^{-1}B)$, and arguments in Eaton (1989, pp. 87–88). For convenience, I give a direct argument. Since

$$\text{vec}(\hat{\Pi}) \sim \mathcal{N}_{2k_n}\left((a'\Omega^{-1}a/n)^{-1/2} a \otimes \eta_n, \Omega \otimes I_{k_n}\right),$$

it follows that the limited information likelihood is given by

$$\mathcal{L}_{\text{LI},n}(\beta, \omega_n, \lambda_n, \Omega) = (2\pi)^{-k_n} |\Omega|^{-\frac{k_n}{2}} e^{-\frac{n}{2}\left(\operatorname{tr}(T\Omega^{-1}) + \lambda_n\right)} e^{n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta,\Omega)}\omega_n' A(\beta,\Omega,\hat{\Pi})} f_S(S;\Omega),$$

where $A(\beta, \Omega, \hat{\Pi}) = \frac{\hat{\Pi}\Omega^{-1}a}{(na'\Omega^{-1}aQ_{\mathcal{T}}(\beta,\Omega))^{1/2}}$. To integrate the likelihood, we use the result that for all $t \in \mathbb{R}$, $\alpha \in \mathbb{S}^{k_n-1}$, and $k_n \geq 2$ (see Stroock, 1999, pp. 88–89)

$$\int_{\mathbb{S}^{k_n-1}} e^{t\alpha'\omega} \, \mathrm{d}F_{\omega_n}(\omega) = \frac{\Gamma(k_n/2)}{\pi^{1/2}\Gamma((k_n-1)/2)} G_{k_n}(t),$$

Applying this result with $t = n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta,\Omega)}$ and $\alpha = A(\beta, \Omega, \hat{\Pi})$ gives

$$\int \mathcal{L}_{\text{LI},n}(\beta, \omega, \lambda_n, \Omega) \, \mathrm{d}F_{\omega_n}(\omega) = \frac{2^{-k_n}\Gamma(k_n/2)}{\pi^{k_n+1/2}\Gamma((k_n-1)/2)} |\Omega|^{-\frac{k_n}{2}} e^{-\frac{n}{2}\left(\operatorname{tr}(T\Omega^{-1}) + \lambda_n\right)} G_{k_n}(n\lambda_n^{1/2} Q_{\mathcal{T}}(\beta,\Omega)^{1/2}) \cdot f_S(S;\Omega),$$

which in view of (30) completes the proof. □

Proof of Proposition 1. Let $\nu = n - k_n - \ell_n$, and to prevent clutter, I use the notation $(\hat{\beta}, \hat{\lambda}, \hat{\Omega})$ rather than $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$. Consider first maximizing the likelihood with respect to Ω, holding β and λ fixed. Let $\hat{\Omega}_{\beta,\lambda}$ denote the resulting estimator. The derivative of the log-likelihood with respect to Ω is given by

$$\frac{\partial \log \mathcal{L}_{\text{RE},n}(\beta, \lambda, \Omega)}{\partial \Omega} = \frac{1}{2}\left[\Omega^{-1}\tilde{S}\Omega^{-1} - (n-\ell_n)\Omega^{-1} - d(\lambda)\left(\Omega^{-1}T\Omega^{-1} - \frac{Q_{\mathcal{S}}(\beta,\Omega)}{b'\Omega b} bb'\right)\right],$$

where $\tilde{S} = nT + \nu S$ and $d(\lambda) = \frac{n\lambda}{k_n/n+\lambda}$ and the derivative $\partial Q_{\mathcal{T}}(\beta,\Omega)/\partial\Omega$ is computed using the identity (26a). Since the derivative equals zero at $\hat{\Omega}_{\beta,\lambda}$, this implies

$$\hat{\Omega}_{\beta,\lambda}^{-1}\tilde{S} = (n-\ell_n)I_2 + d(\lambda)\left(\hat{\Omega}_{\beta,\lambda}^{-1}T - \frac{Q_{\mathcal{S}}(\beta,\hat{\Omega}_{\beta,\lambda})}{b'\hat{\Omega}_{\beta,\lambda}b} bb'\hat{\Omega}_{\beta,\lambda}\right). \tag{34}$$

Taking a trace on both sides of the equation and using the identity (26a) then yields

$$\operatorname{tr}(\hat{\Omega}_{\beta,\lambda}^{-1}\tilde{S}) - d(\lambda)Q_{\mathcal{T}}(\beta,\hat{\Omega}_{\beta,\lambda}) = 2(n-\ell_n). \tag{35}$$

Pre- and post-multiplying Equation (34) by $b'\hat{\Omega}_{\beta,\lambda}$ and b; and by $\hat{\Omega}_{\beta,\lambda}$ and b yields

$$b'\hat{\Omega}_{\beta,\lambda}b = b'\tilde{S}b/(n-\ell_n),$$
$$(n-\ell_n)\hat{\Omega}_{\beta,\lambda}b = \frac{1}{1-d(\lambda)Q_{\mathcal{S}}(\beta,\tilde{S})}(\tilde{S} - d(\lambda)T)b.$$

Plugging these expressions back into Equation (34) and pre-multiplying the resulting expression by $\hat{\Omega}_{\beta,\lambda}$ yields

$$(n-\ell_n)\hat{\Omega}_{\beta,\lambda} = \tilde{S} - d(\lambda)T + \frac{1}{b'(\tilde{S} - d(\lambda)T)b}\frac{d(\lambda)Q_{\mathcal{S}}(\beta,\tilde{S})}{1-d(\lambda)Q_{\mathcal{S}}(\beta,\tilde{S})}(\tilde{S} - d(\lambda)T)bb'(\tilde{S} - d(\lambda)T). \tag{36}$$

Taking a determinant on both sides of the equation, using the matrix determinant lemma $|A + cVV'| = |A|(1 +$

$cV'A^{-1}V$) with $A = \tilde{S} - d(\lambda)T$ and $V = (\tilde{S} - d(\lambda)T)b$, and the identity (26c) yields

$$(n-\ell_n)^2|\hat{\Omega}_{\beta,\lambda}| = \frac{|\tilde{S} - d(\lambda)T|}{1 - d(\lambda)Q_{\mathcal{S}}(\beta,\tilde{S})} = \frac{|\nu S + k_n/(k_n/n+\lambda)\cdot T|}{k_n/n + \lambda\nu/(\nu + nQ_{\mathcal{S}}(\beta,S))}(k_n/n+\lambda)$$
$$= \frac{|S|}{k_n/n+\lambda}\frac{(k_n m_{\max} + \nu(k_n/n+\lambda))(k_n m_{\min} + \nu(k_n/n+\lambda))}{k_n/n + \lambda\nu/(\nu + nQ_{\mathcal{S}}(\beta,S))}$$

Plugging this expression and the expression (35) back into the likelihood then yields that the log-likelihood with Ω concentrated out is given by

$$\log \mathcal{L}_{\text{RE},n}(\lambda,\beta,\hat{\Omega}_{\beta,\lambda}) \propto -\frac{1}{2}\left[k_n \log\left(\frac{k_n}{n}+\lambda\right) + (n-\ell_n)\log\left(\frac{(k_n(m_{\max}+\frac{\nu}{n})+\nu\lambda)(k_n(m_{\min}+\frac{\nu}{n})+\nu\lambda)}{(k_n/n+\lambda)(k_n/n+\lambda\nu/(\nu+nQ_{\mathcal{S}}(\beta,S)))}\right)\right]. \quad (37)$$

Since this expression depends on β only through $Q_{\mathcal{S}}(\beta,S)$, and is decreasing in $Q_{\mathcal{S}}(\beta,S)$ for any $\lambda > 0$, it follows that the maximum likelihood estimate of β with λ fixed at any positive value is given by $\hat{\beta}_\lambda = \operatorname{argmin}_\beta Q_{\mathcal{S}}(\beta,S) = \hat{\beta}_{\text{LIML}}$. If $\lambda = 0$, then the expression doesn't depend on β, and we can in particular set $\hat{\beta}_{\lambda=0} = \hat{\beta}_{\text{LIML}}$, so that $\hat{\beta} = \hat{\beta}_{\text{LIML}}$. The log-likelihood with Ω and β both concentrated out is thus given by

$$\log \mathcal{L}_{\text{RE},n}(\lambda,\hat{\beta}_\lambda,\hat{\Omega}_{\hat{\beta}_\lambda,\lambda}) \propto -\frac{1}{2}\left[k_n \log(k_n/n+\lambda) + (n-\ell_n)\log\left(\frac{k_n m_{\max}}{k_n/n+\lambda}+\nu\right)\right].$$

The derivative equals zero at $\lambda = m_{\max} - k_n/n$, and is negative for $\lambda > m_{\max} - k_n/n$, which implies that $\hat{\lambda} = \max\{m_{\max} - k_n/n, 0\}$. Plugging in the expressions for $\hat{\lambda}$ and $\hat{\beta}$ into (36) then yields

$$(n-\ell_n)\hat{\Omega} = \tilde{S} - d(\hat{\lambda})\left(T - \frac{1}{\hat{b}'T\hat{b}}T\hat{b}\hat{b}'T\right) = \tilde{S} - d(\hat{\lambda})\frac{\hat{a}\hat{a}'|T|}{\hat{b}'T\hat{b}} = \tilde{S} - d(\hat{\lambda})m_{\max}\frac{\hat{a}\hat{a}'}{\hat{a}'S^{-1}\hat{a}},$$

where $\hat{b} = (1,-\hat{\beta})'$, the first equality uses the identity $T\hat{b} = m_{\min}S\hat{b}$, the second equality uses the identity $b'Mb \cdot M = \hat{a}\hat{a}'|B| + Mbb'M$ that holds for any matrix M, and the last equality uses $\hat{b}'T\hat{b} = m_{\min}\hat{b}'S\hat{b}$ and (26b).

Next I derive the inverse Hessian. Let $e_2 = (0,1)'$. The score equations based on the RE likelihood (16) are given by:

$$\mathcal{S}_\beta(\beta,\lambda,\Omega) = d(\lambda)\frac{e_2'(T - Q_{\mathcal{S}}(\beta,\Omega)\Omega)b}{b'\Omega b}, \quad (38)$$
$$\mathcal{S}_\lambda(\beta,\lambda,\Omega) = -\frac{1}{2}\frac{k_n}{k_n/n+\lambda}\left(1 - \frac{Q_{\mathcal{T}}(\beta,\Omega)}{k_n/n+\lambda}\right), \quad (39)$$
$$\mathcal{S}_\Omega(\beta,\lambda,\Omega) = \frac{1}{2}D_2'\operatorname{vec}\left[\Omega^{-1}\tilde{S}\Omega^{-1} - (n-\ell_n)\Omega^{-1} - d(\lambda)\left(\Omega^{-1}T\Omega^{-1} - \frac{Q_{\mathcal{S}}(\beta,\Omega)}{b'\Omega b}bb'\right)\right]. \quad (40)$$

Let $\hat{Q}_{\mathcal{S}} = Q_{\mathcal{S}}(\hat{\beta},\hat{\Omega})$. If $m_{\max} \leq k_n/n$, then the Hessian, evaluated at $(\hat{\beta},\hat{\lambda},\hat{\Omega})$, is singular. Otherwise, it is given by:

$$\mathcal{H}_{\text{RE}}(\hat{\beta},\hat{\lambda},\hat{\Omega}) = \begin{pmatrix} \frac{d(\hat{\lambda})}{\hat{b}'\hat{\Omega}\hat{b}}(\hat{Q}_{\mathcal{S}}\hat{\Omega}_{22} - T_{22}) & 0 & \hat{\mathcal{H}}_{1,3:5} \\ 0 & -\frac{1}{2}\frac{k_n}{(k_n/n+\hat{\lambda})^2} & \hat{\mathcal{H}}_{2,3:5} \\ \hat{\mathcal{H}}'_{1,3:5} & \hat{\mathcal{H}}'_{2,3:5} & \hat{\mathcal{H}}_{3:5,3:5} \end{pmatrix},$$

where

$$\hat{\mathcal{H}}_{1,3:5} = \frac{1}{2}\frac{d(\hat{\lambda})\hat{Q}_{\mathcal{S}}}{\hat{b}'\hat{\Omega}\hat{b}}\left(2\frac{e_2'\hat{\Omega}\hat{b}}{\hat{b}'\hat{\Omega}\hat{b}}\hat{b}\otimes\hat{b}-\hat{b}\otimes e_2-e_2\otimes b\right)' D_2,$$

$$\hat{\mathcal{H}}_{2,3:5} = \frac{1}{2}\frac{k_n}{(k_n/n+\hat{\lambda})^2}\left(\frac{\hat{Q}_{\mathcal{S}}}{\hat{b}\hat{\Omega}\hat{b}}\hat{b}\otimes\hat{b}-\mathrm{vec}(\hat{\Omega}^{-1}T\hat{\Omega}^{-1})\right)' D_2 = -\frac{1}{2}\frac{k_n}{(k_n/n+\hat{\lambda})^2}\frac{m_{\max}}{\hat{a}S^{-1}\hat{a}}\left(\hat{\Omega}^{-1}\hat{a}\otimes\hat{\Omega}^{-1}\hat{a}\right)' D_2,$$

$$\hat{\mathcal{H}}_{3:5,3:5} = -\frac{(n-\ell_n)}{2}D_2'\left(\left(\hat{\Omega}^{-1}-\frac{\hat{c}\hat{b}\hat{b}'}{\hat{b}'\hat{\Omega}\hat{b}}\right)\otimes\left(\hat{\Omega}^{-1}-\frac{\hat{c}\hat{b}\hat{b}'}{\hat{b}'\hat{\Omega}\hat{b}}\right)-(2\hat{c}-\hat{c}^2)\frac{\hat{b}\hat{b}'}{\hat{b}'\hat{\Omega}\hat{b}}\otimes\frac{\hat{b}\hat{b}'}{\hat{b}'\hat{\Omega}\hat{b}}\right)D_2.$$

By the formula for block inverses, the upper 2×2 submatrix of the inverse Hessian is given by:

$$\hat{\mathcal{H}}^{1:2,1:2}(\hat{\beta},\hat{\lambda},\hat{\Omega}) = \left(\hat{\mathcal{H}}_{1:2,1:2}-\hat{\mathcal{H}}_{1:2,3:5}\hat{\mathcal{H}}_{3:5,3:5}^{-1}\hat{\mathcal{H}}_{1:2,3:5}'\right)^{-1}. \tag{41}$$

Applying Lemma A.1 and using the fact that $N_d(A\otimes A) = N_d(A\otimes A)N_d = (A\otimes A)N_d$ (Magnus and Neudecker, 1980, Lemma 2.1(v)) yields:

$$\hat{\mathcal{H}}_{3:5,3:5}^{-1} = -\frac{2}{n-\ell_n}L_2N_2\left[\left(\hat{\Omega}+\frac{\hat{c}}{1-\hat{c}}\frac{\hat{\Omega}\hat{b}\hat{b}'\hat{\Omega}}{\hat{b}'\hat{\Omega}\hat{b}}\right)\otimes\left(\hat{\Omega}+\frac{\hat{c}}{1-\hat{c}}\frac{\hat{\Omega}\hat{b}\hat{b}'\hat{\Omega}}{\hat{b}'\hat{\Omega}\hat{b}}\right)+\frac{\hat{c}^2-2\hat{c}}{(1-\hat{c})^2}\frac{\hat{\Omega}\hat{b}\hat{b}'\hat{\Omega}\otimes\hat{\Omega}\hat{b}\hat{b}'\hat{\Omega}}{(\hat{b}'\hat{\Omega}\hat{b})^2}\right]N_2L_2',$$

It follows that

$$\hat{\mathcal{H}}_{1,3:5}\hat{\mathcal{H}}_{3:5,3:5}^{-1}\hat{\mathcal{H}}_{1,3:5}' = -\frac{(n-\ell_n)\hat{c}^2}{1-\hat{c}}\frac{|\Omega|}{(b'\Omega b)^2}.$$

Finally, since $\hat{\mathcal{H}}_{2,3:5}\hat{\mathcal{H}}_{3:5,3:5}^{-1}\hat{\mathcal{H}}_{1,3:5}' = 0$, Equation (41) combined with the expression in the previous display yields

$$\hat{\mathcal{H}}_{\text{RE}}^{11} = \left(\hat{\mathcal{H}}_{11}-\hat{\mathcal{H}}_{1,3:5}\hat{\mathcal{H}}_{3:5,3:5}^{-1}\hat{\mathcal{H}}_{1,3:5}'\right)^{-1} = \frac{\hat{b}'\hat{\Omega}\hat{b}(\hat{\lambda}+k_n/n)}{n\hat{\lambda}}\left(\hat{Q}_{\mathcal{S}}\hat{\Omega}_{22}-T_{22}+\frac{\hat{c}}{1-\hat{c}}\frac{\hat{Q}_{\mathcal{S}}}{\hat{a}'\hat{\Omega}^{-1}\hat{a}}\right)^{-1},$$

which yields the result.

It remains to show that the inverse Hessian is consistent for $V_{\text{LIML},N}$. To this end, note that $m_{\min} = \frac{\hat{b}_{\text{LIML}}'T\hat{b}_{\text{LIML}}}{\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} \xrightarrow{p} \alpha_k$ by Corollary A.1 and consistency of $\hat{\beta}$. By continuity of the trace operator, and Corollary A.1

$$m_{\max} = \mathrm{tr}(S^{-1}T) - m_{\min} = \mathrm{tr}(S^{-1}T) - \frac{\hat{b}_{\text{LIML}}'T\hat{b}_{\text{LIML}}}{\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} \xrightarrow{p} 2\alpha_k+\lambda-\alpha_k = \lambda+\alpha_k.$$

Consistency of $\hat{\Omega}$ then follows by consistency of $\hat{\lambda}$ and $\hat{\beta}$, Corollary A.1, and Slutsky's Theorem. It also follows that

$$\hat{Q}_{\mathcal{S}} = (n-\ell_n)\frac{\hat{b}'T\hat{b}}{\hat{b}'\tilde{S}\hat{b}} = \frac{(n-\ell_n)\hat{b}'T\hat{b}}{(n-k_n-\ell_n)\hat{b}'S\hat{b}+n b'T\hat{b}} = \left(\frac{1}{1-\ell_n/n}+\frac{n-k_n-\ell_n}{(n-\ell_n)m_{\min}}\right)^{-1} \xrightarrow{p} \alpha_k.$$

Hence,

$$\frac{\hat{c}}{1-\hat{c}} \xrightarrow{p} \frac{\alpha_k\lambda}{\alpha_k(1-\alpha_\ell)+(1-\alpha_k-\alpha_\ell)\lambda},$$

so that

$$-n\hat{H}^{11}_{\text{RE}} \xrightarrow{p} -\frac{b'\Omega b(\alpha_K+\lambda)}{\lambda}\left(-\frac{\lambda}{a'\Omega^{-1}a}+\frac{\lambda\alpha_K^2}{a'\Omega^{-1}a\left((1-\alpha_K-\alpha_\ell)\lambda+(1-\alpha_\ell)\alpha_K\right)}\right)^{-1}$$
$$=\frac{b'\Omega ba'\Omega^{-1}a}{\lambda^2}\left(\lambda+\frac{(1-\alpha_\ell)\alpha_K}{1-\alpha_\ell-\alpha_K}\right)=\mathcal{V}_{\text{LIML},N},$$

which completes the proof. □

Proof of Proposition 2. The objective function evaluates as:

$$\mathcal{Q}_n(\beta,\Xi_{22,n};\hat{W}_{\text{RE}})=\text{tr}((TS^{-1}-(k_n/n)I_2)^2)+\Xi_{22,n}\cdot a'S^{-1}a\left[\Xi_{22,n}\cdot a'S^{-1}a-2Q_{\mathcal{T}}(\beta,S)+2k_n/n\right]. \tag{42}$$

Consider first minimizing the objective function with respect to $\Xi_{22,n}$, holding β fixed. Let $\hat{\Xi}_\beta$ denote the resulting estimator. Since the derivative $\partial\mathcal{Q}_n(\beta,\Xi_{22,n};\hat{W}_{\text{RE}})/\partial\Xi_{22,n}$ equals zero at $\Xi_{22,n}=(Q_{\mathcal{T}}(\beta,S)-k_n/n)/(a'S^{-1}a)$ and is positive for $\Xi_{22,n}\geq(Q_{\mathcal{T}}(\beta,S)-k_n/n)/(a'S^{-1}a)$, we get

$$\hat{\Xi}_\beta=\frac{\max\{Q_{\mathcal{T}}(\beta,S)-k_n/n,0\}}{a'S^{-1}a}. \tag{43}$$

Therefore, the objective function with $\Xi_{22,n}$ concentrated out is given by

$$\mathcal{Q}_n(\beta,\hat{\Xi}_\beta)=\text{tr}((TS^{-1}-(k_n/n)I_2)^2)-(Q_{\mathcal{T}}(\beta,S)-k_n/n)^2\cdot\mathbb{1}\{Q_{\mathcal{T}}(\beta,S)\geq k_n/n\},$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Since $\max_\beta Q_{\mathcal{T}}(\beta,S)=m_{\max}$, with the maximum attained at $\hat{\beta}_{\text{LIML}}$, it follows that if $m_{\max}>k_n/n$, the objective function (42) is uniquely minimized at $(\hat{\beta}_{\text{RE}},\hat{\lambda}_{\text{RE}}/\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}})$. If $m_{\max}\leq k_n/n$, then $\mathcal{Q}_n(\beta,\Xi_{22,n};\hat{W}_{\text{RE}})$ is minimized at $\Xi_{22,n}=0=\hat{\lambda}_{\text{RE}}/\hat{a}'_{\text{RE}}\hat{\Omega}^{-1}_{\text{RE}}\hat{a}_{\text{RE}}$ and an arbitrary $\hat{\beta}$, so that in particular we can set $\hat{\beta}_{\text{RE}}=\beta$. Therefore, Part (i) of Proposition 2 follows if we can show that if $m_{\max}>k_n/n$, then $\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}}=\hat{a}'_{\text{RE}}\hat{\Omega}^{-1}_{\text{RE}}\hat{a}_{\text{RE}}$. Using the notation $\tilde{S}=(n-k_n-\ell_n)S+nT$, we have

$$\begin{aligned}
\hat{a}'_{\text{RE}}\hat{\Omega}^{-1}_{\text{RE}}\hat{a}_{\text{RE}}&=(n-\ell_n)\hat{a}'_{\text{RE}}\left(\tilde{S}-n\frac{m_{\max}-k_n/n}{\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}}}\hat{a}_{\text{RE}}\hat{a}'_{\text{RE}}\right)^{-1}\hat{a}_{\text{RE}}\\
&=-(n-\ell_n)\frac{\hat{a}'_{\text{RE}}\tilde{S}^{-1}\hat{a}_{\text{RE}}\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}}}{n(m_{\max}-k_n/n)\hat{a}_{\text{RE}}\tilde{S}^{-1}\hat{a}_{\text{RE}}-\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}}}\\
&=-(n-\ell_n)\left(nm_{\max}-k_n-\frac{|\tilde{S}|}{|S|}\frac{\hat{b}'_{\text{RE}}S\hat{b}_{\text{RE}}}{\hat{b}_{\text{RE}}\tilde{S}\hat{b}_{\text{RE}}}\right)^{-1}\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}}\\
&=\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}},
\end{aligned} \tag{44}$$

where the first line follows from the definition of $\hat{\Omega}_{\text{RE}}$ and $\hat{\lambda}_{\text{RE}}$ given in Proposition 1, the second line follows by the Woodbury identity, the third line follows from Equation (26b), and the fourth line follows from Equation (26c).

To prove the second part of Proposition 2, I show that whenever the weight matrix satisfies

$$\hat{W}_n\xrightarrow{p}\tilde{c}D_2'\Phi_t^{-1}D_2,\qquad\text{where}\qquad\Phi_t=\Omega\otimes\Omega+\Omega\otimes tmm'+tmm'\otimes\Omega,$$

for some constants $\bar{c} > 0$ and $t \geq 0$, with $m = \Xi_{22}^{1/2} a$, then it is asymptotically optimal. Since we can write $\Phi_t = ((\Omega + tmm') \otimes (\Omega + tmm') - t^2(mm') \otimes (mm'))$, by Lemma A.1, and the identity $\lambda = m'\Omega^{-1}m$,

$$\Phi_t^{-1} = (\Omega^{-1} \otimes \Omega^{-1}) \left[\left(\Omega - \frac{tmm'}{1+t\lambda} \right) \otimes \left(\Omega - \frac{tmm'}{1+t\lambda} \right) + \frac{t^2(mm') \otimes (mm')}{(1+2t\lambda)(1+t\lambda)^2} \right] (\Omega^{-1} \otimes \Omega^{-1}).$$

By Corollary A.1, the asymptotic variance of the moment condition

$$\text{vech}(T - (k_n/n)S - \Xi_{22,n} aa') \tag{45}$$

is given by

$$\Delta = 2L_2 N_2 \left[\tau\Omega \otimes \Omega + \Omega \otimes (mm') + (mm') \otimes \Omega \right] L_2', \tag{46}$$

where $\tau = \alpha_k(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell)$. Suppose first that $\tau > 0$. Then Δ is invertible, and by Lemma A.1(i), its inverse is given by $\Delta^{-1} = \frac{1}{2\tau} D_2' \Phi_{1/\tau}^{-1} D_2$. A necessary and sufficient condition for optimality is that for some matrix C_t (Newey and McFadden, 1994, Section 5.2),

$$(D_2'\Phi_t^{-1}D_2)G = \Delta^{-1}GC_t = \frac{1}{2\tau} D_2'\Phi_{1/\tau}^{-1} D_2 G C_t, \tag{47}$$

where G is the derivative of the moment condition (45), given by:

$$G = -L_2 M, \qquad M = \begin{pmatrix} \Xi_{22}^{1/2}(m \otimes e_1 + e_1 \otimes m) & \frac{1}{\Xi_{22}} m \otimes m \end{pmatrix},$$

where $e_1 = (1,0)'$. Since for a symmetric matrix $A \in \mathbb{R}^{2\times 2}$, $D_2 L_2 \,\text{vec}(A) = \text{vec}(A)$ (Magnus and Neudecker, 1980, p. 427), it follows that $D_2 G = -M$, so that

$$\Phi_t^{-1} D_2 G = -(\Omega^{-1} \otimes \Omega^{-1}) \begin{pmatrix} \frac{\Xi_{22}^{1/2}}{1+t\lambda} \left(m \otimes e_1 + e_1 \otimes m - \frac{2tm'\Omega^{-1}e_1}{1+2t\lambda} m \otimes m \right) & \frac{1}{\Xi_{22}(1+2t\lambda)} m \otimes m \end{pmatrix}.$$

It then follows that (47) holds with

$$C_t = 2\tau \begin{pmatrix} \frac{1+\lambda/\tau}{1+\lambda t} & 0 \\ \frac{2m'\Omega^{-1}e_1\Xi^{3/2}}{1+t\lambda} \left(\frac{1}{\tau} - t\frac{1+2\lambda/\tau}{1+2\lambda t} \right) & \frac{1+2\lambda/\tau}{1+2\lambda t} \end{pmatrix}.$$

If $\tau = 0$, then the asymptotic variance Δ given in Equation (46) is degenerate, since one of the three moment conditions given in Equation (45) is asymptotically redundant: the first moment condition equals 2β times the second minus β^2 times the third. In this case, any weight matrix that puts positive weight on at least two of the moment conditions will be optimal, and in particular $\hat{W}_n$ is optimal. □

Proof of Lemma 2. Part (i) of the Lemma follows from Corollary A.2. Consistency of $\hat{\Psi}_3$ and $\hat{\Psi}_4$ follows from

Lemma A.7 in Anatolyev (2013). Finally, since ZZ' is a projection matrix, $\|ZZ' \operatorname{diag}(H)\| \leq \|\operatorname{diag}(H)\|$, so that

$$\operatorname{var}(\hat{\mu}) = \frac{1}{nk_n}\Omega_{22}\|ZZ' \operatorname{diag}(H)\|^2 \leq \frac{1}{nk_n}\Omega_{22} \operatorname{diag}(H)' \operatorname{diag}(H) = \frac{\Omega_{22}}{n}\delta(1+o(1)) = o(1).$$

Thus, $\hat{\mu} \xrightarrow{p} \mu$ by Markov inequality. □

Proof of Proposition 3. It follows by Corollary A.2 and the Delta method that

$$\sqrt{n}\left(\hat{\beta}_{\text{UMD}} - \Xi_{12,n}/\Xi_{22,n}\right) \Rightarrow \mathcal{N}(0, V_{\text{UMD}}),$$

where, letting $A = (e_2 b' + b e_2')/2\Xi_{22}$, $e_2 = (0,1)'$,

$$\begin{aligned}
V_{\text{UMD}} &= \operatorname{vec}(A)'\Delta \operatorname{vec}(A) \\
&= \operatorname{tr}(4A\Omega A\Xi + 2\tau A\Omega A\Omega) + \alpha_k\delta\left(\mathbb{E}[(v_i'Av_i)^2] - \operatorname{tr}(\Omega A)^2 - 2\operatorname{tr}(A\Omega A\Omega)\right) + 4\alpha_k^{1/2}(\tilde{\mu} + \mu\beta, \mu)\mathbb{E}[Av_i v_i'Av_i] \\
&= V_{\text{LIML},N} + \frac{2\tau(e_2'\Omega b)^2 + 2\alpha_k^{1/2}\mu\mathbb{E}[v_{2i}\epsilon_i^2] + \alpha_k\delta(\mathbb{E}[v_{2i}^2\epsilon_i^2] - 3(e_2'\Omega b)^2 - |\Omega|)}{\Xi_{22}^2} + \frac{\Omega_{22}|\Xi|/\Xi_{22} + 2\alpha_k^{1/2}\tilde{\mu}\mathbb{E}[v_{2i}^2\epsilon_i]}{\Xi_{22}^2} \\
&= V_{\text{LIML}} + \frac{(2\tau + \alpha_k\delta\kappa)(e_2'\Omega b)^2 + 2\gamma\left(\alpha_k\delta\mathbb{E}[(v_{2i} - \gamma\epsilon_i)\epsilon_i^3] + \alpha_k^{1/2}\mu\mathbb{E}[\epsilon_i^3]\right)}{\Xi_{22}^2} + \frac{\Omega_{22}|\Xi|/\Xi_{22} + 2\alpha_k^{1/2}\tilde{\mu}\mathbb{E}[v_{2i}^2\epsilon_i]}{\Xi_{22}^2},
\end{aligned}$$

from which the result follows. □

Proof of Proposition 4. We have:

$$\begin{aligned}
(m_{\min} - \alpha_k) &= \frac{\hat{b}_{\text{LIML}}'(T - \alpha_k S)\hat{b}_{\text{LIML}}}{\hat{b}_{\text{LIML}}' S \hat{b}_{\text{LIML}}} \\
&= \frac{\hat{b}_{\text{LIML}}'(T - \alpha_k S - \lambda_n aa'/(a'\Omega^{-1}a))\hat{b}_{\text{LIML}}}{\hat{b}_{\text{LIML}}' S \hat{b}_{\text{LIML}}} + \frac{\lambda_n(a'\hat{b}_{\text{LIML}})^2}{(a'\Omega^{-1}a)\hat{b}_{\text{LIML}}' S \hat{b}_{\text{LIML}}} \\
&= \frac{(\hat{b}_{\text{LIML}} \otimes \hat{b}_{\text{LIML}})' \operatorname{vec}\left(T - \alpha_k S - \lambda_n aa'/(a'\Omega^{-1}a)\right)}{\hat{b}_{\text{LIML}}' S \hat{b}_{\text{LIML}}} + \frac{\lambda_n(\hat{b}_{\text{LIML}} - \beta)^2}{(a'\Omega^{-1}a)\hat{b}_{\text{LIML}}' S \hat{b}_{\text{LIML}}} \\
&= \frac{(\hat{b}_{\text{LIML}} \otimes \hat{b}_{\text{LIML}})' \operatorname{vec}\left(T - (k_n/n)S - \Xi_{22,n} aa'\right)}{\hat{b}_{\text{LIML}}' S \hat{b}_{\text{LIML}}} + O_p(n^{-1}) \\
&= \frac{(b \otimes b)' \operatorname{vec}\left(T - (k_n/n)S - \Xi_{22,n} aa'\right)}{b'\Omega b} + O_p(n^{-1}),
\end{aligned}$$

where the first line follows from the identity $m_{\min} = Q_S(\hat{\beta}_{\text{LIML}}, S)$, the second and third line follows by algebra, the fourth line and the last line follow from $\sqrt{n}$-rate of convergence $\hat{\beta}_{\text{LIML}}$ and $T - (k_n/n)S$. Expanding the numerator then yields

$$\frac{n}{\sqrt{k_n}}(m_{\min} - \alpha_k) = \frac{(b \otimes b)' \operatorname{vec}(V'(H/\sqrt{k_n})V)}{b'\Omega b} + O_p(k_n^{-1/2}) = \frac{\epsilon'(H/\sqrt{k_n})\epsilon}{b'\Omega b} + O_p(k_n^{-1/2}).$$

If $k_n \to \infty$, then by Lemmata A.2 and A.3, with $P_n = H/\sqrt{k_n}$,

$$\frac{n}{\sqrt{k_n}}(m_{\min} - \alpha_k) \Rightarrow \mathcal{N}(0, 2(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell) + \delta\kappa).$$

If $k_n = K$ is fixed, then $\text{vec}(\hat{\Pi} - \Pi) = \sum_i v_i \otimes z_i \Rightarrow \mathcal{N}(0, \Omega \cdot I_K)$, since the Lyapunov condition is implied by $\sum_{i=1}^n \|z_i\|^{2+\nu} = \sum_i (ZZ')_{ii}^{1+\nu/2} \leq \max_i (ZZ')_{ii}^{\nu/2} \operatorname{tr}(ZZ') = o(1)$ for any $\nu > 0$. It then follows by standard arguments that $nm_{\min} \Rightarrow \chi^2_{K-1}$, which proves the first part.

To prove the second part, I use the approximation from Peiser (1943) (see also Anatolyev and Gospodinov, 2011) that as $k \to \infty$,

$$q^{\chi^2_k}_{1-\text{ns}} = k + \Phi^{-1}(1-\text{ns})\sqrt{2k} + O(1),$$

where $q^{\chi^2_k}_{1-\text{ns}}$ denotes the $1-\text{ns}$ quantile of a χ^2 distribution with k degrees of freedom. Therefore, if $k_n \to \infty$, letting $c = \Phi(\sqrt{\frac{(1-\alpha_k)}{1-\alpha_k-\alpha_\ell}} + \frac{\delta\kappa}{2}\Phi^{-1}(\text{ns}))$

$$\begin{aligned}
\mathbb{P}\left(nm_{\min} \geq q^{\chi^2_{k_n-1}}_{1-c}\right) &= \mathbb{P}\left(nm_{\min}/\sqrt{k_n} \geq \sqrt{k_n} + \Phi^{-1}(1-c)\sqrt{2} + o(1)\right) \\
&= \mathbb{P}\left(nm_{\min}/\sqrt{k_n} - \sqrt{k_n} \geq \Phi^{-1}(1-c)\sqrt{2} + o(1)\right) \\
&= \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \Phi^{-1}(1-c)\sqrt{\frac{2(1-\alpha_k-\alpha_\ell)}{2(1-\alpha_\ell)+(1-\alpha_k-\alpha_\ell)\delta\kappa}} + o(1)\right) \\
&= \Phi\left(\Phi^{-1}(c)\sqrt{\frac{2(1-\alpha_k-\alpha_\ell)}{2(1-\alpha_\ell)+(1-\alpha_k-\alpha_\ell)\delta\kappa}}\right) + o(1) \\
&= \text{ns} + o(1).
\end{aligned}$$

If k_n is fixed, then, since $\sum_{i=1}^n (ZZ')_{ii}^2 \leq \max_j (ZZ')_{jj} \operatorname{tr}(ZZ')$, it follows that $\delta_n = \sum_{i=1}^n (ZZ')_{ii}^2 / k_n + o(1) = o(1)$. Thus, $c = \text{ns} + o(1)$, and so $\mathbb{P}\left(nm_{\min} \geq q^{\chi^2_{k_n-1}}_{1-c}\right) = \text{ns} + o(1)$. Since δ_n and $\hat{\kappa}$ are consistent estimators of δ and κ, the assertion of the theorem follows. □

References

Abramowitz, M. and Stegun, I. A. (eds.) (1965). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. New York: Dover. 31, 32

Aizer, A. and Doyle, J. J., Jr. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, **130** (2), 759–803. 20

Anatolyev, S. (2013). Instrumental variables estimation and inference in the presence of many exogenous regressors. *The Econometrics Journal*, **16** (1), 27–72. 4, 7, 18, 38

— and Gospodinov, N. (2011). Specification testing in models with many instruments. *Econometric Theory*, **27** (2), 427–441. 4, 25, 26, 39

Anderson, T. W., Kunitomo, N. and Matsushita, Y. (2010). On the asymptotic optimality of the LIML estimator with possibly many instruments. *Journal of Econometrics*, **157** (2), 191–204. 4, 18, 19

— and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, **20** (1), 46–63. 8

Andrews, D. W. K., Moreira, M. J. and Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, **74** (3), 715–752. 9

—, — and — (2008). Efficient two-sided nonsimilar invariant tests in iv regression with weak instruments. *Journal of Econometrics*, **146** (2), 241–254. 6

Angrist, J. D., Graddy, K. and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, **67** (3), 499–527. 3, 21

— and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, **90** (430), 431–442. 3, 21

— and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106** (4), 979–1014. 7

Arellano, M. (2003). *Panel Data Econometrics*. New York, NY: Oxford University Press. 4, 16

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, **62** (3), 657–681. 2, 4, 7, 8, 9, 15

— and Crudu, F. (2015). Jackknife instrumental variable estimation with heteroskedasticity. *Journal of Econometrics*, **185** (2), 332–342. 4

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. B. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80** (6), 2369–2429. 4

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, **170** (2), 383–398. 4

Cattaneo, M. D., Crump, R. K. and Jansson, M. (2012). Optimal inference for instrumental variables regression with non-gaussian errors. *Journal of Econometrics*, **167** (1), 1–15. 20

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, **18** (1), 5–46. 10, 19

— (2007). Decision theory applied to an instrumental variables model. *Econometrica*, **75** (3), 609–652. 4, 6, 9

— and Imbens, G. W. (2004). Random effects estimators with many instrumental variables. *Econometrica*, **72** (1), 295–306. 1, 3, 8, 11, 12

— and Moreira, M. J. (2009). Decision theory applied to a linear panel data model. *Econometrica*, **77** (1), 107–133. 4

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R. and Woutersen, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics*, **178** (1), 15–21, Annals Issue: Misspecification Test Methods in Econometrics. 4

— and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, **73** (5), 1673–1692. 4

—, —, Hausman, J. A., Newey, W. K. and Woutersen, T. (2012). Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory*, **12** (1), 42–86. 4, 29

Chioda, L. and Jansson, M. (2009). Optimal invariant inference when the number of instruments is large. *Econometric Theory*, **25** (3), 793–805. 4, 9, 12

Cragg, J. G. and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, **9** (2), 222–240. 3, 26

Dobbie, W. and Song, J. (2015). Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review*, **105** (3), 1272–1311. 20

Donald, S. G. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, **69** (5), 1161–1191. 3, 21, 22

Eaton, M. L. (1989). *Group invariance applications in statistics*, *Regional conference series in Probability and Statistics*, vol. 1. Hayward, California: Institute of Mathematical Statistics. 9, 31, 32

Gautier, E. and Tsybakov, A. B. (2014). High-dimensional instrumental variables regression and confidence sets, arXiv:1105.2454. 4

Goldberger, A. S. and Olkin, I. (1971). A minimum-distance interpretation of limited-information estimation. *Econometrica*, **39** (3), 635–639. 15

Hahn, J. (2002). Optimal inference with many instruments. *Econometric Theory*, **18** (1), 140–168. 4, 19

Hansen, C. B., Hausman, J. A. and Newey, W. K. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, **26** (4), 398–422. 4, 18, 29

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, **3** (2), 211–255. 4

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** (2), 467–475. 21

Kolesár, M. (2013). Estimation in instrumental variables models with heterogeneous treatment effects, working paper, Princeton University. 21

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E. L. and Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, **33** (4), 474–484. 7, 9, 21, 22

Kunitomo, N. (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *Journal of the American Statistical Association*, **75** (371), 693–700. 2, 4

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, **95** (2), 391–413. 4

— (2002). Orthogonal parameters and panel data. *Review of Economic Studies*, **69** (3), 647–666. 13

Lee, Y. and Okui, R. (2012). Hahn-Hausman test as a specification test. *Journal of Econometrics*, **167** (1), 133–139. 26

Magnus, J. R. and Neudecker, H. (1980). The elimination matrix: Some lemmas and applications. *SIAM Journal on Algebraic and Discrete Methods*, **1** (4), 422–449. 14, 28, 35, 37

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, **71** (4), 1027–1048. 8

— (2009). A maximum likelihood method for the incidental parameter problem. *The Annals of Statistics*, **37** (6A), 3660–3696. 4, 9, 11, 31

Morimune, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica*, **51** (3), 821–841. 2, 4

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, **27** (4), 575–595. 3, 21

Newey, W. K. and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, *Chapter 36*, New York, NY: Elsevier, pp. 2111–2245. 15, 18, 37

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16** (1), 1–32. 2, 4

Peiser, A. M. (1943). Asymptotic formulas for significance levels of certain distributions. *The Annals of Mathematical Statistics*, **14** (1), 56–62. 39

Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 1, *Chapter 8*, North Holland, pp. 449–516. 5

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, *Chapter 15*, New York, NY: North Holland, pp. 881–935. 7

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, **26** (3), 393–415. 4, 26

Sims, C. A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of Econometrics*, **95** (2), 443–462. 4

Stroock, D. W. (1999). *A Concise Introduction to the Theory of Integration*. Boston, MA: Birkhäuser, 3rd edn. 33

van der Ploeg, J. and Bekker, P. A. (1995). *Efficiency Bounds for Instrumental Variable Estimators Under Group-Asymptotics*. SOM Research Report 95B24, University of Groningen. 19

van Hasselt, M. (2010). Many instruments asymptotic approximations under nonnormal error distributions. *Econometric Theory*, **26** (02), 633–645. 4, 18

Wong, C. S. and Wang, T. (1992). Moments for elliptically contoured random matrices. *Sankhya: The Indian Journal of Statistics, Series B*, **54** (3), 265–277. 19