INTEGRATED LIKELIHOOD APPROACH TO INFERENCE WITH MANY INSTRUMENTS

Michal Kolesár*
Cowles Foundation
Yale University

Version 1.5, October 9, 2013†

Abstract

I analyze a Gaussian linear instrumental variables model with a single endogenous regressor in which the number of instruments is large. I use an invariance property of the model and a Bernstein-von Mises type argument to construct an integrated likelihood which by design yields inference procedures that are valid under many instrument asymptotics and are asymptotically optimal under rotation invariance. I establish that this integrated likelihood coincides with the random-effects likelihood of Chamberlain and Imbens (2004), and that the maximum likelihood estimator of the parameter of interest coincides with the limited information maximum likelihood (liml) estimator.

Building on these results, I then relax the basic setup along two dimensions. First, I drop the assumption of Gaussianity. In this case, liml is no longer optimal, and I derive a new, more efficient estimator based on a minimum distance objective function that imposes a rank restriction on the matrix of second moments of the reduced-form coefficients. Second, I consider minimum distance estimation without imposing the rank restriction and I show that the resulting estimator corresponds to a version of the bias-corrected two-stage least squares estimator.

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

*Electronic correspondence: kolesarmi@googlemail.com. 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 in seminars at Harvard University and Yale University for helpful comments and suggestions.

†Version hash 6ab33ad, compiled October 9, 2013

1 Introduction

This paper provides a principled and unified way of doing inference in a linear instrumental variables model with homoscedastic errors in which the number of instruments is potentially large. The presence of a large number of instruments creates an incidental parameter problem (Neyman and Scott, 1948) because the number of first-stage coefficients corresponds to the number of instruments. To capture this problem in asymptotic approximations, I follow Kunitomo (1980), Morimune (1983), and Bekker (1994) and employ many instrument asymptotics that allow the number of instruments to increase in proportion with the sample size, thus allowing the number of incidental parameters in the model to diverge to infinity. I focus on the case in which each instrument is weak in the Staiger and Stock (1997) sense, but collectively the instruments have substantial predictive power, so that the concentration parameter grows at the same rate as the sample size. I allow the rate of growth of the instruments to be zero, in which case the asymptotics reduce to standard strong instrument asymptotics.

One possible way of dealing with the incidental parameter problem is to simply ignore it, and base inference on full likelihood of the model. This turns out to work for estimation, but not for testing or construction of confidence sets. In particular, the maximum likelihood estimator of the coefficient on the endogenous variable, β, known as the limited information maximum likelihood (LIML, Anderson and Rubin, 1949) estimator, remains consistent (Bekker, 1994) under many instrument asymptotics. Moreover, LIML is also efficient among estimators that are invariant to rotations of the instruments if the errors are Normal (Chioda and Jansson, 2009). However, the curvature of the likelihood is too large, and likelihood-based tests and confidence sets suffer from size-distortions.

In this paper, I address the incidental parameter problem directly. In particular, I show that if the errors are Normally distributed, an invariance property of the model and a Bernstein-von Mises type argument can be used to construct an integrated likelihood, which by design delivers inference procedures that are valid under many-instrument asymptotics, and asymptotically optimal under rotation invariance. I show that this likelihood coincides with the random-effects (RE) likelihood of Chamberlain and Imbens (2004), and that the maximum likelihood estimator of β coincides with LIML. Therefore, a simple and principled way of doing inference is to use LIML with standard errors based on the inverse Hessian of the RE likelihood, which I show has a simple closed form.

I derive this basic result in three steps. The first step is to orthogonalize the first stage coefficients so that the information matrix is block-diagonal in the new parametrization. This helps to separate the problem of inference about the parameter of interest β from that of inference about the nuisance parameters.

The second step is to appeal to the invariance principle to reduce the dimensionality of the model. I decompose the orthogonalized first-stage coefficients into a high-dimensional parameter

ω_n on the unit sphere which governs the direction of the coefficients a scalar parameter λ_n, proportional to the concentration parameter of Rothenberg (1984), that governs their norm. ω_n thus measures the relative strength of the individual instruments, while λ_n measures their collective strength. Under rotation invariance, the parameter ω_n drops out, so that the maximal invariant on the parameter space has fixed dimension even as the number of instruments increases to infinity. Imposing invariance is equivalent to assuming a uniform prior for ω_n, and the likelihood for the maximal invariant (invariant likelihood) is equivalent to an integrated likelihood which integrates ω_n out using this uniform prior.

Since the invariant model is locally asymptotically Normal (Chioda and Jansson, 2009), inference based on the invariant likelihood will be asymptotically efficient in the class of invariant procedures. Moreira (2009) shows that the maximum invariant likelihood estimator of β in the case when the reduced-form covariance matrix Ω is known coincides with LIMLK. I generalize this result along two dimensions. First, if Ω is not known, then the maximum invariant likelihood estimator coincides with LIML. This equivalence explains why LIML is a consistent and efficient invariant estimator despite being based on the concentrated likelihood which in general does not produce consistent estimators in incidental parameter problems. Second, constraining λ_n to equal to a particular value does not affect the maximum invariant likelihood estimate of β.

This result motivates the third step, to put prior a over λ_n in addition to a prior over ω_n and integrate the likelihood over both priors. This additional prior will not affect the maximum integrated likelihood estimator of β, which will still be LIML. If the prior is suitably chosen, the resulting integrated likelihood will yield simpler inference procedures than those based on the invariant likelihood which involve numerical optimization. Moreover, so long as the prior is not dogmatic, it will get dominated in large samples, so that imposing it will not affect asymptotic validity of inference about β either.

The prior I use is a scaled chi-square prior with an unknown scale parameter. This prior, together with a uniform prior on ω_n is equivalent to the random effects prior on the orthogonalized first-stage coefficients proposed by Chamberlain and Imbens (2004): a Normal prior with zero mean and unknown variance (which corresponds to the scale parameter). Therefore, my approach yields an integrated likelihood that is identical to the RE likelihood. Consequently, the random-effects quasi-maximum likelihood estimator of β coincides with LIML.

This analysis yields new insights into the sources of identification in the instrumental variables model, and I use these insights to relax two assumptions that underlie the basic setup. First, I drop the assumption that the errors are Normally distributed. I show that LIML is no longer efficient and derive a new, more efficient estimator. In particular, in the linear instrumental variables model, the coefficients on the instruments in the first-stage regression are proportional to the coefficients in the reduced-form outcome regression. This proportionality restriction implies a rank restriction on the matrix of second moments of the reduced form coefficients, which governs the distribution of the maximal invariant, and is the source of identification of β in the invariant model. I use this

rank restriction to construct a minimum distance objective function and show that the RE estimator of the model parameters minimizes this minimum distance objective function with respect to a particular weight matrix. This weight matrix is optimal if the errors in the instrumental variables model are Normally distributed, but not otherwise; using weights proportional the inverse of the asymptotic covariance matrix of the moment conditions yields a more efficient estimator. The validity of standard errors for LIML based on the Hessian of the RE likelihood also depends on the assumption of Normality; standard errors based on the conventional GMM/minimum distance formula are robust to non-normality.

Second, I derive an unrestricted minimum distance estimator that does not impose the rank restriction. I show that this estimator coincides with a version of the bias-corrected two-stage least squares estimator (Nagar, 1959; Donald and Newey, 2001), and derive its asymptotic variance without assuming proportionality of the reduced form coefficients. These results thus provide a way of doing inference that is robust, for example, to heterogeneity in the causal effect, as in Imbens and Angrist (1994). When the causal effect is heterogeneous, the reduced-form coefficients are no longer proportional, but the instrumental variables estimand β can be interpreted as a weighted average of the derivative of the effect of the endogenous variable on the outcome (Angrist, Graddy and Imbens, 2000).

The minimum distance objective function is also helpful in deriving a specification test that is robust to many instruments. A test of the rank restriction is equivalent to a test proposed by Cragg and Donald (1993), but with an adjusted critical value. The adjustment ensures that the test is valid under strong as well as many instrument asymptotics. In contrast, when the number of covariates is allowed to increase with the sample size, the size of the standard Sargan (1958) specification test converges to one.

This paper draws on two separate strands of literature. First is the literature on many instruments that builds on the on the work by Kunitomo (1980), Morimune (1983), Bekker (1994) and Chao and Swanson (2005). Like Anatolyev (2011), I relax the assumption that the dimension of covariates 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. An interesting new development is to employ shrinkage techniques to obtain more efficient estimators (see, for example, Belloni, Chen, Chernozhukov and Hansen, 2012, Gautier and Tsybakov, 2011, or Carrasco, 2012), although these results rely on an additional sparsity assumption on the first-stage coefficients. In contrast, I do not make any assumptions about the first-stage coefficients in this paper apart from assuming that collectively, the instruments are relevant. 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 (2010) and Bekker and Crudu (2012) also allow for

heteroscedasticity.

The second strand of literature is the literature on incidental parameters started by the seminal paper of Neyman and Scott (1948). Lancaster (2000) and Arellano (2003) discuss the incidental parameter 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 the invariance principle. Lancaster (2002) proposes to put a flat prior on the orthogonalized nuisance parameters, rather than the Normal prior with finite unknown variance used here. Cox and Reid (1987) suggest conditioning the likelihood on a maximum likelihood estimate of the orthogonalized incidental parameters. In the instrumental variables model, both proposals yield the concentrated limited information likelihood, and therefore don't deliver valid inference.

The remainder of this paper is organized as follows. Section 2 sets up the instrumental variables model, introduces the notation, and finds an orthogonal reparametrization. Section 3 reviews the limited information likelihood approach to inference. Section 4 uses invariance and Bernstein-von Mises arguments to construct the integrated likelihood and study its properties. Section 5 relaxes the Normality assumption and considers a minimum distance approach to inference. Section 6 considers minimum distance estimation without imposing proportionality of the reduced-form coefficients. Section 7 studies tests of overidentifying restrictions. Section 8 concludes. Proofs and derivations are collected in the Appendix. The online Supplementary 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 covariates to increase in proportion with the sample size. Second, I reduce the data to its sufficient statistics and find an orthogonal reparametrization of the first-stage coefficients.

2.1 Model and Assumptions

There is a sample of individuals $i = 1, \ldots, n$, whose outcomes y_i are determined by the structural equation

$$y_i = x_i\beta + w_i'\delta_n + \epsilon_i. \tag{1}$$

The parameter of interest is β, which governs the causal effect of x_i on the outcome y_i. δ_n is the coefficient from regressing $y_i - x_i\beta$ onto an ℓ_n-dimensional vector of covariates w_i (including an

intercept), so that w_i is by definition uncorrelated with the structural error ϵ_i. However, x_i may be endogenous in that it may be correlated with ϵ_i.

The assumption underlying identification of β in the linear instrumental variables model is that there is a k_n-dimensional vector of instruments z_i, correlated with x_i, but uncorrelated with the structural error:

Assumption LIV (Linear Instrumental Variables Model). $\mathbb{E}[z_i\epsilon_i] = 0$.

This assumption requires that **(i)** z_i only affects outcome through its effect on the endogenous variable—this is known as the exclusion restriction; and **(ii)** there is no heterogeneity in the effect of x_i on y_i. In Section 6, I relax this assumption to allow for such heterogeneity, and for certain violations of the exclusion restriction. If $k_n > 1$, then the model is overidentified in the sense that Assumption LIV is testable; I discuss tests of this assumption in Section 7.

It will be convenient to work with an orthogonalized version of the original instruments. To describe the orthogonalization, let W denote the $n \times \ell_n$ matrix of covariates 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'. Let $\tilde{Z} = Z - W(W'W)^{-1}W'Z$ denote the residuals from regressing Z onto W. Then the orthogonalized instruments $Z_\perp \in \mathbb{R}^{n\times k_n}$ are given by the rotation $Z_\perp = \tilde{Z}(R')^{-1}$, where the lower-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 $Z_\perp$ are orthogonal to each other as well as to the columns of W. This orthogonalization is sometimes called a standardizing transformation (see Phillips (1983) for discussion).

Let $Y = (y, x) \in \mathbb{R}^{n\times 2}$ with rows $Y_i' = (y_i, x_i)$ pool all endogenous variables in the model. Then the reduced-form regression of the endogenous variables onto $Z_\perp$ and W can be compactly written as

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

where $V \in \mathbb{R}^{n\times 2}$ with rows $v_i' = (v_{1i}, v_{2i})$ pools the reduced-form errors. Under Assumption LIV, the structural error is related to the reduced-form errors by $\epsilon_i = v_{1i} - v_{2i}\beta$, and $\pi_{1,n}$ is proportional to $\pi_{2,n}$, with the constant of proportionality given by β,

$$\pi_{1,n} = \pi_{2,n}\beta.$$

Throughout the paper, I assume that the reduced-form errors v_i are i.i.d. and conditionally homoscedastic:

$$\mathbb{E}[v_i \mid W, Z] = 0, \qquad \mathbb{E}[v_i v_i' \mid W, Z] = \Omega. \tag{3}$$

The consistency results in this paper will rely on the additional structure on second moments of the data that the conditional homoscedasticity provides. Recent papers by Hausman *et al.* (2012), Chao

et al. (2012), Bekker and Crudu (2012) and Kolesár (2012) propose to use jackknife type estimators that are consistent under many instruments even in the presence of heteroscedasticity. Those estimators are, however, less efficient under homoscedasticity than the estimators considered here.

In order to employ sufficiency and invariance arguments, I will also assume that the errors follow a bivariate Normally distribution:

Assumption N (Normality). $v_i \mid W, Z \sim \mathcal{N}_2(0, \Omega)$.

This assumption has no effect on consistency results. Normality, however, does have an effect on asymptotic distributions and asymptotic efficiency properties of estimators. I relax this assumption in Section 5 when I discuss a minimum-distance approach to inference in this model.

Apart from being excluded from the structural equation, the instruments also have to be relevant in the sense that they have to be correlated with the endogenous variable. To measure the strength of identification, I follow Chamberlain (2007) and Andrews, Moreira and Stock (2008) and I use

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

This parameter is related to the concentration parameter of Rothenberg (1984), which is given by $\pi_{2,n}'\pi_{2,n}/(n\Omega_{22})$. Instead of dividing $\pi_{2,n}'\pi_{2,n}/n$ by the variance of v_{2i}, λ_n multiplies it by the (2,2) element of the precision matrix of (ϵ_i, v_{2i}), which is given by $a'\Omega^{-1}a$. Therefore, if there is no endogeneity problem, so that the correlation between ϵ_i and v_{2i} is zero, the two measures coincide. Otherwise, they are proportional to each other.

The goal is to construct inference procedures that work well even if the number of instruments k_n and the number of covariates ℓ_n is large relative to sample size. To capture the finite-sample behavior in these settings in asymptotic approximations, I follow Anatolyev (2011) and Kolesár, Chetty, Friedman, Glaeser and Imbens (2011) and allow for many-instrument asymptotics with both k_n and ℓ_n potentially growing in proportion to 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)** $\{(z_i, w_i, v_i) \in \mathbb{R}^{k_n} \times \mathbb{R}^{\ell_n} \times \mathbb{R}^2 \colon i = 1, \ldots, n; k_n + \ell_n < n\}_{n\geq 1}$ is a triangular array of i.i.d. random variables; **(iii)** (W, Z) is full column rank with probability one; and **(iv)** $\lambda_n \to \lambda$ 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 a large number of instruments is the result of interacting a few basic instruments with many covariates (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 covariates w_i, such as higher-level fixed effects in multilevel sampling; for example, if the set of instruments randomly

assigned within a school, we need to condition on school fixed effects. The remaining parts are standard. Part (ii) allows the distribution of the random variables to change with the sample size. To reflect this, I should index the random variables by n. I drop this index for ease of notation, and only use the subscript n for parameters which change with the sample size. Part (iii) normalizes the first-stage regressors to be full rank, so that the orthogonalized instruments $Z_\perp$ are uniquely defined. Finally, Part (iv) is the many-instruments equivalent of the relevance assumption. It is equivalent to assuming that the Rothemberg concentration parameter grows at the same rate as the sample size. By allowing $\alpha_k = \alpha_\ell = 0$, Assumption MI nests the standard strong instrument asymptotic sequence in which the number of instruments and covariates is fixed.

2.2 Sufficient statistics

Under Normality, the set of sufficient statistics is given by the least-squares estimator of the reduced form coefficients,

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

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

$$S = \hat{V}'\hat{V}/(n - k_n - \ell_n) \in \mathbb{R}^{2\times 2}, \qquad \hat{V} = Y - Z_\perp \hat{\Pi} - W\hat{\Psi}.$$

The virtue of working with the orthogonalized instruments is that now the rows of $\hat{\Pi}$ are mutually independent. Rather than working with the full set of sufficient statistics, I base inference on $\hat{\Pi}$ and S only[1] as in Moreira (2003) and Chamberlain and Imbens (2004). Since the distribution of $\hat{\Psi}$ is unrestricted, dropping it from the model does not result in loss of information. This step eliminates the potentially high-dimensional nuisance parameters $\psi_{1,n}$ and $\psi_{2,n}$, so that the model parameters are now given by the triplet $(\beta, \pi_{2,n}, \Omega)$.

It will be useful to define the following functions of the statistics $\hat{\Pi}$ and S:

$$T = \hat{\Pi}'\hat{\Pi}/n,$$

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

$$m_{\min} = \min \operatorname{eig}(S^{-1}T), \qquad m_{\max} = \max \operatorname{eig}(S^{-1}T).$$

The functions $Q_S(\beta,\Omega)$ and $Q_T(\beta,\Omega)$ of T will appear in several objective functions. The

[1]Formally, this requirement can be justified by requiring invariance to location shifts in $\hat{\Psi}$ in the sample space, and invariance to location shifts in $(\psi_{1,n}, \psi_{2,n})$ in the parameter space. Since the goal is to make inferences about β, the loss function will not depend on $(\psi_{1,n}, \psi_{2,n})$, and will therefore also be invariant to this transformation.

properties of $Q_S(\beta,\Omega)$ and $Q_T(\beta,\Omega)$ are discussed in Andrews, Moreira and Stock (2006).[2] The bigger eigenvalue, $m_{\max}$ will help to determine instrument relevance. On the other hand, the smaller eigenvalue $m_{\min}$ plays a key role in testing Assumption LIV.

2.3 Orthogonal parametrization

To help separate the problem of inference about β from that of the nuisance parameters, I rescale $\pi_{2,n}$ as

$$\eta_n = \pi_{2,n}\sqrt{a'\Omega^{-1}a/n},$$

so that the strength of identification is given by $\lambda_n = \eta_n'\eta_n$. The advantage of the (β,η_n,Ω) parametrization is that the parameter of interest β is information-orthogonal to the nuisance parameters (η_n,Ω) in the sense that the information matrix is block-diagonal.[3] In terms of this parametrization, the distribution of the statistics $\hat{\Pi}$ and S is given by

$$\text{vec}(\hat{\Pi}) \sim \mathcal{N}_{2k_n}\left((a'\Omega^{-1}a/n)^{-1/2}a\otimes\eta_n,\Omega\otimes I_{k_n}\right), \tag{5}$$

$$(n-k_n-\ell_n)S \sim \mathcal{W}_2(n-k_n-\ell_n,\Omega), \tag{6}$$

with $\hat{\Pi}$ independent of S, where $\mathcal{W}_2(n-k_n-\ell_n,\Omega)$ denotes a Wishart distribution with $n-k_n-\ell_n$ degrees of freedom, and scale matrix Ω.

3 Limited information likelihood

In this Section, I briefly review the failure of the model likelihood, called the limited information likelihood, to deliver asymptotically valid inference.[4]

The problem with likelihood-based inference under many-instrument asymptotics is that the dimension of the nuisance parameter η_n increases with the number of instruments to infinity. Therefore, the standard results about optimality of likelihood-based inference do not apply since they require the dimension of the parameter space to remain fixed.

The model likelihood based on the statistics $\hat{\Pi}$ and S is known as the limited-information likelihood after a seminal paper by Anderson and Rubin (1949). It turns out that maximizing it actually delivers a consistent estimator of β despite incidental parameter problem (Bekker, 1994).

[2]While the statistics $Q_S(\beta,\Omega)$ and $Q_T(\beta,\Omega)$ correspond to those in Andrews *et al.* (2006), my statistics S and T do not correspond to theirs.

[3]See Cox and Reid (1987) for a discussion of the consequences of orthogonal parametrization in problems with nuisance parameters.

[4]The derivation of the results in this section is given in the Supplementary Appendix.

This estimator, known as the limited information maximum likelihood (LIML) estimator, solves

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

Unfortunately, inference about β based on the limited information likelihood fails for two reasons.

First, the curvature of the likelihood is too big—the block of the information matrix corresponding to β is given by $\mathcal{I}_{\text{LI},11} = n \cdot b'\Omega b \cdot a'\Omega^{-1}a/\lambda_n$, while the asymptotic distribution of β under Assumptions LIV, N and MI is given by (see Bekker, 1994 and Kolesár *et al.*, 2011 for derivation)

$$\sqrt{n}\left(\hat{\beta}_{\text{LIML}} - \beta\right) \Rightarrow \mathcal{N}_1\left(0, \mathcal{V}_{\text{LIML},N}\right), \tag{8}$$

where

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

Unless $\alpha_k = \alpha_\ell = 0$ as the standard asymptotic sequence assumes, the (1,1) element of the inverse information matrix[5] $(\mathcal{I}_{\text{LI}}^{-1})_{11} = \mathcal{I}_{\text{LI},11}^{-1}$ will miss the correction factor in the parentheses. This correction factor can be substantial even when the ratio of instruments to sample size, α_k, is small if the normalized concentration parameter λ is small. The presence of many covariates, the case when $\alpha_\ell > 0$, has a negligible impact on the asymptotic variance unless α_ℓ is large. (On the other hand, their presence does have a big impact on tests for overidentifying restrictions as I discuss in Section 7.) As a result, confidence intervals for $\hat{\beta}_{\text{LIML}}$ based on $\mathcal{I}_{\text{LI},11}^{-1}$ will undercover.

Second, the maximum likelihood estimates of λ_n and Ω are inconsistent:

$$\hat{\lambda}_{\text{LIML}} = \frac{n-\ell_n}{n-k_n-\ell_n} m_{\max} \xrightarrow{p} \frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}(\lambda + \alpha_k), \tag{10a}$$

$$\hat{\Omega}_{\text{LIML}} = \frac{n-k_n-\ell_n}{n-\ell_n} S + \frac{n m_{\min}}{n-\ell_n}\left(S - \frac{\hat{a}_{\text{LIML}}\hat{a}'_{\text{LIML}}}{\hat{a}'_{\text{LIML}} S^{-1} \hat{a}_{\text{LIML}}}\right) \xrightarrow{p} \Omega - \frac{\alpha_k}{1-\alpha_\ell}\frac{aa'}{a'\Omega^{-1}a}. \tag{10b}$$

Consequently, a plug-in estimator of the inverse information matrix $\mathcal{I}_{\text{LI},11}(\hat{\beta}_{\text{LIML}}, \hat{\lambda}_{\text{LIML}}, \hat{\Omega}_{\text{LIML}})^{-1}$ will be asymptotically biased downward, so that the feasible confidence intervals that use the estimated information matrix will undercover even more than infeasible intervals that use the true information matrix $\mathcal{I}_{\text{LI},11}^{-1}$. Similarly, a plug-in estimator based on the correct asymptotic variance formula (9) will also be asymptotically downward biased. Bekker (1994) and Hansen *et al.* (2008) therefore suggest using different estimators λ and Ω that are consistent under MI when $\alpha_\ell = 0$. These asymptotic variance estimators, however, have to be modified again if we want to allow $\alpha_\ell > 0$ (Anatolyev, 2011; Kolesár *et al.*, 2011).

In the next section, I introduce an alternative (quasi-) likelihood approach which will by construction avoid these problems.

[5] Recall that the information matrix is block-diagonal under the (β, η_n, Ω) parametrization.

4 Equivalence between Integrated and Random Effects Likelihoods

This section derives the basic result of the paper that combining **(i)** an invariance argument; and **(ii)** a Bernstein-von Mises argument, we can construct an integrated likelihood that has a simple closed form and addresses the incidental parameter problem.

4.1 Using invariance to reduce the dimension of the parameter space

The idea behind using an invariance argument is that if we require inference to be invariant to suitably chosen group actions, the maximal invariant in the parameter space will preserve β, and it will have a fixed dimension even as the number of instruments grows. Therefore, the number of parameters in the invariant likelihood will be independent of the number of instruments, and, so long as the invariant likelihood is sufficiently smooth, inference based on it will be consistent and efficient among invariant procedures by standard likelihood-efficiency arguments.

I follow Andrews *et al.* (2006), Chamberlain (2007), Chioda and Jansson (2009), and Moreira (2009), and I consider transformations given by

$$\overline{m}_1(g,(\hat{\Pi},S)) = (g\hat{\Pi},S), \qquad \overline{m}_2(g,(\beta,\eta_n,\Omega)) = (\beta,g\eta_n,\Omega), \qquad g \in \mathcal{O}(k_n),$$

where $\mathcal{O}(k_n)$ is the group of $k_n \times k_n$ orthogonal matrices. Here $\overline{m}_1$ is the action on the sample space, and it rotates the direction of the instruments. Correspondingly, $\overline{m}_2$, the action on the parameter space, rotates the direction of the first-stage coefficients η_n, preserving the collective strength of the instruments as measured by $\lambda_n = \eta_n'\eta_n$. It is straightforward to show that the maximal invariants are given by $T = \hat{\Pi}'\hat{\Pi}/n$ and S on the sample space, and (β,λ_n,Ω) on the parameter space. The potentially high-dimensional vector of first-stage coefficients η_n has been reduced to a scalar.

Intuitively, imposing invariance means that estimation and inference should not depend on the choice of basis for the instruments—if we re-order the instruments, change their scale, or use a different orthogonalization procedure to construct $Z_\perp$—we should get the same point estimate and confidence intervals for β. If the decision rule (the rule used for constructing the point estimates and confidence intervals from the data) depends on the data only through the maximal invariants S and T, it will have this property.

A simple way to construct such invariant decision rules would be to use the likelihood based on S and T. Since the parameter space of the maximal invariant (S,T) is given by the maximal invariant on the parameter space, (β,λ_n,Ω), which has a fixed dimension irrespective of the number of instruments, the number of parameters in this likelihood, $\mathcal{L}_{\text{INV},n}(\beta,\lambda_n,\Omega;S,T)$, which I call the invariant likelihood, is now fixed, thus avoiding the nuisance parameter problem. Moreover, since this likelihood, is sufficiently smooth, so that it is locally asymptotically Normal under many-instrument asymptotics (Chioda and Jansson, 2009), inference based on it will be

asymptotically efficient among invariant procedures by standard arguments (see, for example, van der Vaart, 1998, Chapter 8).

Moreira (2009) shows that if Ω is known, the maximum invariant likelihood estimator for β coincides with LIMLK, which is indeed asymptotically efficient among invariant estimators. The next proposition generalizes this result:

Proposition 1. *The* MLE *based on the invariant likelihood* $\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T)$ *is given by* $\hat{\beta}_{\text{LIML}}$. *This result also holds if* λ_n *is fixed at an arbitrary value.*

The first part of the proposition generalizes Moreira's result to the case when Ω is not known, and shows that the maximal invariant likelihood estimator then coincides with LIML. Since LIML is efficient among regular invariant estimators, this result confirms that maximizing the invariant likelihood indeed produces an efficient invariant estimator. Furthermore, this result also explains why the limited-information likelihood produces an estimator that is robust to many instruments: it is because LIML happens to coincide with the maximum invariant likelihood estimator.

The second part of the proposition shows that constraining λ_n to be equal to a particular value does not affect the maximum invariant likelihood estimate. This result is similar to that in Chamberlain (2007) who shows that the Bayes rule under a particular loss function and prior for β does not depend on the prior for λ_n. Since the information matrix of the invariant likelihood is block-diagonal between λ_n and β, the maximum likelihood estimate of β when λ_n is given should vary only slowly with λ_n (see Cox and Reid, 1987, Section 2.2). The proposition shows that the dependence is even more limited: the estimate does not vary with λ_n at all.

There is an alternative way of building the invariant likelihood that will allow me to use this result to build a connection between it and the random-effects likelihood of Chamberlain and Imbens (2004). The argument is similar to that in Chamberlain and Moreira (2009), who relate invariant likelihood to a correlated random effects likelihood in a dynamic panel data model. In particular, imposing invariance is equivalent to assuming a particular prior distribution for the model parameters, induced by the Haar measure on $\mathcal{O}(k_n)$, called the invariant prior distribution (Eaton, 1989). Since the group $\mathcal{O}(k_n)$ is compact, this prior is unique. Consider a polar decomposition of the first stage coefficients, $\eta_n = \omega_n \lambda_n^{1/2}$, where

$$\omega_n = \eta_n / \|\eta_n\|, \qquad \lambda_n = \|\eta_n\|^2.$$

The potentially high-dimensional nuisance parameter ω_n is a point on the unit sphere that measures the direction of η_n; it can be thought of as measuring the relative strength of the individual instruments. Under this decomposition, the invariant prior is given by the uniform distribution over the unit sphere $\mathbb{S}^{k_n-1}$ in $\mathbb{R}^{k_n}$, the parameter space for the parameter ω_n. Furthermore, the invariant likelihood is equivalent to the integrated (marginal) likelihood that uses this invariant

prior as a prior distribution. Denoting the prior by $F_{\omega_n}(\cdot)$, this relationship can be written as

$$\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T) = \int_{\mathbb{S}^{k_n-1}} \mathcal{L}_{\text{LI},n}(\beta, \lambda_n, \omega_n, \Omega; \hat{\Pi}, S) \, \mathrm{d}F_{\omega_n}(\omega_n), \tag{11}$$

where $\mathcal{L}_{\text{LI},n}$ is the limited information likelihood, the likelihood for the statistics $\hat{\Pi}$ and S.

4.2 Integrated likelihood and random effects likelihood

One disadvantage of the invariant likelihood is that due to the presence of Bessel functions in the likelihood expression, estimates of λ_n and Ω are not available in closed form and have to be computed by maximizing the invariant likelihood numerically. This makes construction of likelihood-based confidence intervals for β difficult, since these estimates are needed for evaluating the Hessian. Therefore, although the inverse Hessian evaluated at maximum likelihood estimates is a consistent estimator of the asymptotic variance of $\hat{\beta}_{\text{LIML}}$, getting Hessian-based standard error estimates involves numerical optimization.

This motivates an introduction of a prior over λ_n, in addition to the uniform prior over ω_n. If this additional prior is appropriately chosen, integrating the limited information likelihood over both priors will yield an integrated likelihood that is more convenient to work with than the invariant likelihood. Since by Proposition 2, constraining λ_n does not affect the maximum invariant likelihood estimator for β, introducing a prior for λ_n will not affect it either: it will still be given by $\hat{\beta}_{\text{LIML}}$. Moreover, so long as this low-dimensional prior is not dogmatic, the Bernstein-von Mises theorem should apply, and the prior should get dominated in large samples. Therefore inference based on the integrated likelihood should agree with inference based on the invariant likelihood in large samples.

The family of priors I consider is a scaled chi-square family with an unknown scale parameter $\lambda > 0$:

$$\lambda_n \sim \frac{\lambda}{k_n} \chi^2(k_n). \tag{12}$$

The hyperparameter λ in this prior 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. Combined with the uniform prior over ω_n, these two priors are equivalent to a single Normal prior over the scaled first-stage coefficients η_n,

$$\eta_n \sim \mathcal{N}(0, \lambda / k_n \cdot I_{k_n}). \tag{13}$$

This Normal prior is the random-effects prior proposed in Chamberlain and Imbens (2004)—it says that we should treat the the first-stage coefficients as random with a zero-mean Normal distribution. Therefore, the integrated likelihood obtained after integrating the limited information likelihood over the invariant prior on ω_n and the chi-square prior on λ_n coincides with the RE likelihood that integrates the limited information likelihood over a single Normal prior (13). The

RE likelihood, unlike the invariant likelihood, has a simple closed form (see the Supplementary Appendix for derivation):[6]

$$\begin{aligned}\mathcal{L}_{\text{RE},n}(\beta,\lambda,\Omega) &= \int_{\mathbb{R}^{k_n}} \mathcal{L}_{\text{LI},n}(\beta,\eta_n,\omega_n,\Omega;\hat{\Pi},S)\,\mathrm{d}F_{\eta_n|\lambda}(\eta_n \mid \lambda) \\ &= \int_{\mathbb{R}}\int_{\mathbb{S}^{k_n-1}} \mathcal{L}_{\text{LI},n}(\beta,\lambda_n,\omega_n,\Omega;\hat{\Pi},S)\,\mathrm{d}F_{\omega_n}(\omega_n)\,\mathrm{d}F_{\lambda_n|\lambda}(\lambda_n \mid \lambda) \\ &= \left(1+\frac{n}{k_n}\lambda\right)^{-k_n/2} |\Omega|^{-(n-\ell_n)/2} e^{-\frac{1}{2}\operatorname{tr}(\Omega^{-1}((n-k_n-\ell_n)S+nT))+\frac{n}{2}\frac{\lambda}{k_n/n+\lambda}Q_{\mathcal{T}}(\beta,\Omega)}.\end{aligned} \tag{14}$$

This equivalence shows that there are two ways of thinking about the RE assumption (13) that the first-stage coefficients η_n are Normally distributed with zero mean and unknown variance that is to be estimated from the data. The first, which was the motivation in Chamberlain and Imbens (2004), is to view it as a modeling tool: since the prior has zero mean, it captures the idea that the individual instruments may not be very relevant, and it reduces the original high-dimensional model to a smooth model in which the parameter space stays 5-dimensional even as $\ell_n \to \infty$ and $k_n \to \infty$. Hence, if the RE assumption holds, so that the first-stage coefficients are actually generated according to (13), inference based on the RE likelihood will have the usual asymptotic optimality properties—maximum likelihood estimators, Wald, LM and LR test will be asymptotically efficient, and the inverse Hessian will be a consistent estimator for the asymptotic variance.

The second way of thinking about the RE prior (13) is to view it as arising from two priors. The uniform prior over ω_n can be motivated by invariance arguments. Moreover, Chamberlain (2007) shows this prior is least favorable, so that it can also be motivated by finite-sample minimax considerations. The prior on λ_n is used to make inference more convenient and will not matter asymptotically. Therefore, even if the first-stage coefficients are not actually drawn according to (13)—in particular, even if they are viewed as fixed—inference based on the RE likelihood will stay asymptotically valid, and it will be asymptotically optimal among invariant procedures.

Proposition 2. *Consider the model* (1)–(3).

(i) *Suppose that $m_{\max} > k_n/n$. Then the maximum likelihood estimators based on the* RE *likelihood* (14) *are given by:*

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

[6]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.

(ii) *Under Assumptions LIV, N, and MI,* $(\hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}}) \xrightarrow{p} (\lambda, \Omega)$.

Part (i) of Proposition 2 formalizes the claim that the estimator of β remains unchanged under the additional chi-square prior for λ_n. Part (ii) of Proposition 2 shows that, unlike estimators based on the limited information likelihood given in Equation (10), the RE estimators of λ and Ω are consistent under many instrument asymptotics. The assumption that $m_{\max} \geq k_n/n$ makes sure that the constraint $\lambda \geq 0$ does not bind when maximizing the likelihood. It will hold in large samples if Assumption MI (iv) holds.

Proposition 3. *Consider the model* (1)–(3).

(i) *The (1,1) element of the inverse Hessian of the* RE *likelihood* (14)*, evaluated at* $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$*, is given by:*

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

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

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

This result 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 $\hat{\beta}_{\text{RE}}$.

The key condition underlying Proposition 3 is that the extra prior on λ_n is not dogmatic. For example, Lancaster (2002) suggests to deal with incidental parameters in panel data models by first orthogonalizing them, and then integrating them out with respect to a suitable uniform prior. In the instrumental variables model the parameter space for the orthogonalized parameters η_n is $\mathbb{R}^k$, so that a "uniform prior" corresponds to a flat prior on $\mathbb{R}^k$, which in turn corresponds to a uniform prior on ω_n, and an improper prior on λ_n, obtained by taking the limit as $\lambda \to \infty$ of the chi-square prior (12). The integrated likelihood based on this prior corresponds to the limit of the RE likelihood (14) as $\lambda \to \infty$:

$$\lim_{\lambda \to \infty} \mathcal{L}_{\text{RE},n}(\beta, \lambda, \Omega) = |\Omega|^{-(n-\ell_n)/2} e^{-\frac{1}{2}\operatorname{tr}(\Omega^{-1}((n-k_n-\ell_n)S + nT)) + \frac{n}{2} Q_{\mathcal{T}}(\beta, \Omega)}.$$

This objective function coincides with the concentrated limited information likelihood that concentrates η_n out, and therefore does not produce valid confidence intervals, since the prior on λ puts all its mass far away from regions near zero. On the other hand, this dogmatic prior on λ_n does not affect the consistency of the maximum integrated likelihood estimator of β as the second part of Proposition 1 predicts.

5 Efficient minimum distance estimation under non-Normal errors

Identification in the invariant model comes from restrictions on the expectation of the invariant statistics S and T imposed by the exclusion restriction. The Normality assumption on the errors plays no role. Therefore, another way to construct invariant decision rules is to use a minimum distance estimator that minimizes distance between the maximal invariants S and T and their expected values. If the weight function weighs the restrictions efficiently, this approach will still deliver estimators and inference procedures that are efficient among invariant rules. 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.

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 weighs the restrictions efficiently under Normality, but not otherwise. I then use this equivalence result to construct minimum-distance based standard errors for LIML that are valid under non-Normality. Finally, I derive an efficient minimum distance estimator when the Normality assumption is dropped.

To simplify the expressions in this section, let D_2 denote the duplication matrix, L_2 the elimination matrix and N_2 the symmetrizer matrix. The duplication matrix transforms the vech operator into a vec operator[7], and the elimination operator performs the reverse operation, so that $D_d \operatorname{vech}(A) = \operatorname{vec}(A)$, and $L_d \operatorname{vec}(A) = \operatorname{vech}(A)$, where $A \in \mathbb{R}^{d\times d}$. The symmetrizer matrix has the property that $N_d \operatorname{vec}(A) = (1/2)\operatorname{vec}(A + A')$. Other properties of these matrices are given in Appendix A.

5.1 Random effects and minimum distance

The reduced form (2)–(3) of the instrumental variables model without any further assumptions implies

$$\mathbb{E}[S] = \Omega, \tag{15a}$$

$$\mathbb{E}[T - (k_n/n)S] = \Xi_n, \qquad \text{where} \quad \Xi_n = \frac{1}{n}\begin{pmatrix}\pi_{1,n} & \pi_{2,n}\end{pmatrix}\begin{pmatrix}\pi_{1,n} & \pi_{2,n}\end{pmatrix}'. \tag{15b}$$

Since the parameters Ω, $\pi_{1,n}$ and $\pi_{2,n}$ are unrestricted, these two expectations are unrestricted. Under Assumption LIV, however, the second-stage coefficients $\pi_{1,n}$ are restricted to be proportional to the first stage coefficients: $\pi_{1,n} = \pi_{2,n}\beta$. This restriction leads to a rank restriction on the matrix

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

of second moments of the reduced form coefficients, Ξ_n, namely that $\Xi_n = \Xi_{22,n} aa'$, where

$$\Xi_{22,n} = \pi_{2,n}' \pi_{2,n} / n = \lambda_n / (a' \Omega^{-1} a).$$

This restriction can be used to build a minimum distance objective function

$$\mathcal{Q}_n(\beta, \Xi_{22,n}; \hat{W}_n) = \operatorname{vech}\left(T - (k_n/n)S - \Xi_{22,n} aa'\right)' \hat{W}_n \operatorname{vech}\left(T - (k_n/n)S - \Xi_{22,n} aa'\right), \tag{16}$$

where $\hat{W}_n \in \mathbb{R}^{3\times 3}$ is some weight matrix. Since the nuisance parameter Ω only appears in the first moment condition (15a), which is unrestricted, minimizing an objective function (16) that only uses the second moment condition with respect to an efficient weight matrix will yield an estimator of β that has the same asymptotic variance as the efficient minimum distance that uses both of them (Chamberlain, 1982, Section 3.2).

In the random effects model, the identification of the model coefficients is based on the same restriction. The only difference is that the parameter $\Xi_{22,n} = \lambda_n / (a' \Omega^{-1} a)$ is replaced by its expectation under the chi-square prior (12) on λ_n, $\Xi_{22} = \lambda / (a' \Omega^{-1} a)$. There should therefore exist a weight matrix such that the random effects estimator of (β, Ξ_{22}) is asymptotically equivalent to a minimum distance estimator with respect to this weight matrix. The next proposition shows that if the weight matrix is chosen carefully, the minimum distance and random effects estimators are in fact *identical*.

Proposition 4. *Suppose that* $\operatorname{tr}(S^{-1}T) \geq 2k_n/n$. *Then the minimum distance estimator based on the objective function* (16) *with respect to the weight matrix* $\hat{W}_{\text{RE}} = D_2'(S^{-1} \otimes S^{-1}) D_2$ *is given by* $(\hat{\beta}_{\text{RE}}, \hat{\Xi}_{22,\text{RE}})$.

The condition that $\operatorname{tr}(S^{-1}T) > 2k_n/n$ makes sure that the objective function is not minimized at a boundary. It will hold in large samples if $\lambda > 0$. The next Proposition shows that if the errors are Normally distributed, then the random effects weight matrix $\hat{W}_{\text{RE}}$ weighs the moment condition (15b) efficiently under many-instrument asymptotics.

Proposition 5. *Consider the model* (1)–(3)*, and suppose that Assumptions LIV, N, and MI hold. Consider a minimum distance estimator based on the objective function* (16)*. Suppose that, for some constants* $\bar{c} > 0, c \geq 0$, *the weight matrix satisfies*

$$\hat{W}_n \xrightarrow{p} \bar{c} D_2' \left[\Omega \otimes \Omega + c\Omega \otimes (aa') + c(aa') \otimes \Omega\right]^{-1} D_2.$$

Then the minimum distance estimator for $(\beta, \Xi_{22}, \Omega)$ *is optimal among the class of minimum distance estimators.*

When $\alpha_k > 0$, then setting $c = \Xi_{22}/\tau$, and $\bar{c} = 2/\tau$, where $\tau = \alpha_k(1 - \alpha_\ell)/(1 - \alpha_k - \alpha_\ell)$ in the limit weight in Proposition 5 yields the inverse of the asymptotic covariance of the moment condition (15b). The proposition shows that it is possible to misspecify Ξ_{22} in the optimal weight

without affecting the asymptotic distribution of the minimum distance estimator.[8] In particular, the weight matrix $\hat{W}_{\text{RE}}$ satisfies the condition in Proposition 5 with $c = 0$. When $\alpha_k = 0$ (standard asymptotics case), then the variance of the moment condition is reduced-rank since one of the three moment conditions is redundant, and in this case any positive definite weight matrix will yield an asymptotically optimal estimator.

The result that the random effects estimators can be obtained by minimizing a minimum distance objective function with respect to an efficient weight matrix is similar to Goldberger and Olkin (1971), who consider a minimum distance objective function based on the proportionality restriction that the exclusion restriction imposes on the expectation of $\hat{\Pi}$

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

Goldberger and Olkin (1971) show that this objective function is minimized at $\hat{\beta}_{\text{LIML}}$. The weight matrix $S^{-1} \otimes I_{k_n}$ consistently estimates the inverse of the asymptotic variance of $\text{vec}(\hat{\Pi})$ under standard asymptotics.

5.2 Minimum distance estimation under non-Normal errors

The efficiency result in Proposition 5 as well expression for asymptotic distribution of $\hat{\beta}_{\text{LIML}}$ given in (9) depend on the Normality assumption N. This sensitivity to the assumption of Normality is similar to the result in panel-data models in which identification is based on covariance restrictions; there the weight matrix used by the maximum likelihood estimator is only optimal under Normality (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 (15b). The moment condition depends on the data through the three-dimensional statistic $\text{vech}(T - (k_n/n)S)$, which can be written as a quadratic form

$$T - (k_n/n)S = n(Z_\perp \pi_{2,n}a' + V)'H(Z_\perp \pi_{2,n}a' + V),$$

where

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

We need to impose some regularity conditions on the components H, $Z_\perp \pi_{2,n}a'$, and V of the quadratic form:

Assumption RC (Regularity conditions). (i) The reduced-form errors v_i are iid with finite fourth moments; **(ii)** For some $\delta, \mu \in \mathbb{R}$, $d'd/n \to \delta$ and $n^{-1}\pi_{2,n}'Z_\perp' d \to \mu$ where $d = \text{diag}(H) \in \mathbb{R}^n$; and

[8] 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).

(iii) For some constant $C \in \mathbb{R}$, $\sup_n \sup_{1\le i\le n} \|(Z_\perp)_i' \pi_{2,n}\| < C$ and $\sup_n \sup_{1\le i\le n} \sum_{s=1}^n (|(Z_\perp Z_\perp')_{is}| + |(W(W'W)^{-1}W')_{is}|) < C$.

Part (i) relaxes the Normality assumption on the errors. Part (ii) ensures that all terms in the asymptotic covariance matrix are well-defined. Part (iii) implies that a Lindeberg-type condition holds.

Lemma 1. *Consider the model* (1)–(3)*. Then, under Assumptions LIV, MI, and RC:*

(i)

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

where

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

(ii) *Let $M = I_n - Z_\perp Z_\perp' - W(W'W)^{-1}W'$, and let $\hat{V} = MY$ with rows $\hat{v}_i$ denote estimates of the reduced-form errors. If the errors v_i have finite eighth moments and $\alpha_k > 0$, then*

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

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. 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 $\hat{\Delta}_j$ 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, and replacing Ψ_3 and Ψ_4 by $\hat{\Psi}_3$ and $\hat{\Psi}_4$.

Since $\hat{\beta}_{\text{LIML}}$ is a distance estimator, its asymptotic variance under Assumptions LIV, MI and RC is given by the (1,1) element of the matrix

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

where $W = D_2'(\Omega^{-1} \otimes \Omega^{-1})D_2 = \operatorname{plim} \hat{W}_{\text{RE}}$. This element evaluates as

$$\mathcal{V}_{\text{LIML}} = \mathcal{V}_{\text{LIML},N} + \frac{2\mu}{\Xi_{22}^2}\mathbb{E}[v_{2\backslash\epsilon}\epsilon^2] + \frac{\delta}{\Xi_{22}^2}\mathbb{E}[\epsilon^2 v_{2\backslash\epsilon}^2 - |\Omega|],$$

where $v_{2\backslash\epsilon} = v_2 - b'\Omega e_2 \cdot (b'\Omega b)^{-1}\epsilon$ is the part of the first-stage error that is uncorrelated with $\epsilon = v_1 - v_2\beta$, the error in the structural equation. 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 excessive kurtosis. Anatolyev (2011) 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 this variance can easily be computed by replacing Δ by $\hat{\Delta}$ and replacing a and Ω in the expressions for G and W by $\hat{a}_{\text{RE}}$ and $\hat{\Omega}_{\text{RE}}$, and plugging the estimates $\hat{G}$, $\hat{W}$, and $\hat{\Omega}$ into the expression (19).

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

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

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 $\mathcal{V}_{\text{EMD}}$ is given by the (1,1) element of the matrix $(G'\Delta^{-1}G)^{-1}$, where G is the derivative of the moment condition

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

The expression evaluates as

$$\mathcal{V}_{\text{EMD}} = \mathcal{V}_{\text{LIML}} - \frac{1}{\Xi_{22}^2 (b'\Omega b)^2} \frac{\left(\mu\mathbb{E}[\epsilon^3] + \delta\mathbb{E}[v_{2\backslash\epsilon}\epsilon^3]\right)^2}{2\tau + \delta\mathbb{E}[\epsilon^4/(b'\Omega b)^2 - 3]}. \tag{20}$$

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.

The second term in (19) represents the efficiency gain relative to LIML. This efficiency gain is zero in two important cases. First, if the structural error has zero skewness, and when

the dependence between ϵ and v_2 is linear so that $\mathbb{E}[v_{2\backslash\epsilon}\epsilon^3] = 0$. This is the case when the reduced-form errors are Normal, but it also holds more generally. For example, Anderson *et al.* (2010) show that when the errors belong to the family of elliptically contoured distributions,[9] then LIML is efficient in the class of estimators that depend on the data only through smooth functions of S and T. Indeed, for elliptically contoured distributions, $\Psi_3 = 0$, so that $\mathbb{E}[\epsilon^3] = 0$ and Ψ_4 is proportional to $\text{vec}(\Omega)\,\text{vec}(\Omega)' + 2N_2\Omega \otimes \Omega$ (Wong and Wang, 1992), which implies $\mathbb{E}[v_{2\backslash\epsilon}\epsilon^3] = (b' \otimes b')\Psi_4(b \otimes (v_2 - (b'\Omega e_2)/(b'\Omega b)b)) = 0$, so that the RE weight matrix remains efficient for β (although it is not efficient for $\Xi_{22,n}$).

Second, when $\delta = 0$, which by the Cauchy-Schwarz inequality, $\mu^2 \leq \delta\Xi_{22}$, implies $\mu = 0$. The term δ measures how balanced the design is. By definition of the matrix H, $0 \leq \delta < \tau$. If the diagonal elements of the projection matrices $(Z_\perp Z_\perp')_{ii}$ and $(W(W'W)^{-1}W)_{ii}$ (called the leverage of i) are equal to k_n/n and ℓ_n/n, respectively, then $\delta = 0$. This situation arises, for example, when the instruments are group indicators, and the group sizes are all equal. Therefore, for relatively balanced designs in which no observations have a large influence on the reduced-form estimates as measured by their leverage, the term δ, and the efficiency gain, will be small. Consequently, the efficiency loss of $\hat{\beta}_{\text{LIML}}$ relative to the optimal minimum distance estimator will be small unless the design is unbalanced, *and* distribution of the reduced-form errors displays substantial skewness or kurtosis.

6 Minimum distance estimation without rank restriction

Assumption LIV imposes a proportionality restriction on the reduced form (2) in that $\pi_{1,n} = \pi_{2,n}\beta$. This proportionality restriction implies that the matrix Ξ_n is reduced rank. In particular, it implies that there are two sources of information for estimating β,

$$\Xi_{11,n} = \Xi_{12,n}\beta, \qquad \text{and} \tag{21a}$$

$$\Xi_{12,n} = \Xi_{22,n}\beta. \tag{21b}$$

The minimum distance objective function (16) weighs both sources of identification. In this section, I consider estimation without imposing the rank restriction, and 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 does not use Equation (21a) to estimate β.

6.1 Motivation for relaxing the rank restriction

If Assumption LIV fails, then it is no longer the case that $\pi_{1,n}$ is proportional to $\pi_{2,n}$, and the matrix Ξ_n is no longer reduced-rank. However, there are two important cases in which

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

the ratio $\Xi_{12,n}/\Xi_{22,n}$, which is the estimand of the two-stage least squares estimator under standard asymptotics (Kolesár, 2012), is still of interest. Consequently, estimators that only use Equation (21b) to identify β will be robust to these two failures of Assumption LIV.

The first case arises when the effect of x_i on y_i is heterogeneous, as in Imbens and Angrist (1994), so that the true structural model is $y_i = h_1(x_i, \epsilon_i)$, where ϵ_i is a vector of individual unobserved heterogeneity, and h_1 is some unknown function. For simplicity, suppose there are no covariates w_i beyond a constant. Suppose that **(i)** Exclusion restriction holds: $\epsilon_i \perp\!\!\!\perp z_i$, so that the instrument only affects the outcome through its effect on x_i; and **(ii)** Monotonicity holds: $x_i = h_2(z_i, u_i)$, and for any pair (z_1, z_0), $\mathbb{P}(h_2(z_1, u_i) \geq h_2(z_0, u_i))$ equals either zero or one. Then $\Xi_{12,n}/\Xi_{22,n}$ can be written as a particular weighted average of average partial derivatives $\beta(z) = \mathbb{E}[\partial h_1(h_2(z, u_i), \epsilon_i)/\partial x]$ (see Angrist and Imbens (1995) and Angrist *et al.* (2000) for details). On the other hand, the ratio $\Xi_{11,n}/\Xi_{12,n}$ may be outside of the convex hull of the average partial derivatives (Kolesár, 2012).

The second case arises when the exclusion restriction fails and the instrument has a direct effect on the outcome. In this case, the error in the structural equation has the form $\epsilon_i = z'_{\perp,i}\gamma_n + \nu_i$, where $\mathbb{E}[z_{\perp,i}\nu_i] = 0$ and γ_n measures the strength of the direct effect. Consequently, the coefficient $\pi_{1,n}$ in the reduced-form regression of the outcome on instruments is given by $\pi_{1,n} = \pi_{2,n}\beta + \gamma_n$. Without any restrictions on γ_n, the parameter β is no longer identified. However, Kolesár *et al.* (2011) show that if the direct effects are orthogonal to the effects of the instruments on the endogenous variable in the sense that

$$\pi'_{2,n}\gamma_n/n \to 0, \tag{22}$$

then β can still be consistently estimated. In particular, under this condition $\Xi_{12,n}/\Xi_{22,n} = \beta + \gamma'_n\pi_{2,n}/\pi'_{2,n}\pi_{2,n} \to \beta$. In contrast, $\Xi_{11,n}/\Xi_{12,n} \to \beta$ only if direct effects disappear asymptotically so that $\gamma'_n\gamma_n/n \to 0$.

To explain the motivation behind the condition (22), consider an example from Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan (2011). Chetty *et al.* (2011) are interested in estimating the effect of early childhood achievement, as measured by kindergarten test scores, on subsequent outcomes, using data from the Tennessee STAR experiment. For concreteness, take the outcome of interest to be first-grade scores. In the STAR experiment, children and teachers were randomly assigned to kindergarten classrooms, generating an exogenous variation in kindergarten test scores. Assuming that teachers only affect subsequent outcomes through their effect on test scores, we should therefore be able to use kindergarten teacher indicators as instruments for kindergarten test scores. However, since classes mostly stay together in subsequent years, the instrument also affects outcomes directly: the kindergarten teacher indicator coincides with kindergarten classroom indicator, which also has an effect on outcomes through the first-grade teacher. We cannot partial out the effect of first-grade teachers since their assignment is perfectly correlated

with kindergarten teacher assignment. However, in the STAR experiment, the first-grade teachers are randomly assigned. Hence, the first-grade teacher effects γ_n are orthogonal to the kindergarten teacher effects, $\pi_{2,n}$ so that the condition (22) is satisfied.

6.2 Unrestricted minimum distance estimation

To relax rank restriction on Ξ_n, the matrix of second moments of the reduced-form vectors $\pi_{2,n}$ and $\pi_{1,n}$, parametrize it as

$$\Xi_n = \Xi(\Xi_{11,n}, \Xi_{22,n}, \beta_n) = \begin{pmatrix} \Xi_{11,n} & \Xi_{22,n}\beta_n \\ \Xi_{22,n}\beta_n & \Xi_{22,n} \end{pmatrix},$$

so that now β_n is defined simply as the ratio $\Xi_{12,n}/\Xi_{22,n}$. This parametrization leads to the objective function

$$\begin{aligned} &\mathcal{Q}_n(\beta, \Xi_{11}, \Xi_{22}; \hat{W}_n) = \\ &\qquad \operatorname{vech}\left(T - (k_n/n)S - \Xi(\Xi_{11}, \Xi_{22}, \beta)\right)' \hat{W}_n \operatorname{vech}\left(T - (k_n/n)S - \Xi(\Xi_{11}, \Xi_{22}, \beta)\right), \end{aligned} \quad (23)$$

where $\hat{W}_n \in \mathbb{R}^{3\times 3}$ is some weight matrix. If we restrict $\Xi_{11,n}$ to equal to $\Xi_{22,n}\beta^2$, then minimizing this objective function is equivalent to minimizing the original objective function (16). 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.*, 2011), 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 introducing the MBTSLS estimator in Kolesár *et al.* (2011) was to modify the Donald and Newey BTSLS 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 6. *Consider the reduced form* (2)–(3)*. Suppose that Assumptions N and MI (i)–(iii) holds*

and that $\Xi_n \to \Xi$*, where* Ξ *is some positive semi-definite matrix with* $\Xi_{22} > 0$*. Then*

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

where

$$V_{\text{UMD}} = \frac{b'\Omega b}{\Xi_{22}}\left(1 + \frac{\tau}{a'\Omega^{-1}a \cdot \Xi_{22}}\right) + \frac{\Omega_{22}|\Xi|}{\Xi_{22}^3} + \frac{2\tau(\Omega_{12} - \beta\Omega_{22})^2}{\Xi_{22}^2}, \tag{24}$$

and $\tau = \alpha_k(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell)$*.*

The asymptotic distribution of $\hat{\beta}_{\text{UMD}}$ follows from a central limit theorem for $T - (k_n/n)S$, and the delta method. It is possible to relax the Normality assumption and generalize Lemma 1; I focus on the Normal case here for simplicity. A consistent plug-in estimator of the asymptotic variance can easily be constructed using the fact that $T - (k_n/n)S \xrightarrow{p} \Xi$, and $S \xrightarrow{p} \Omega$.

The asymptotic variance consists of three components. The first term coincides with the asymptotic variance of LIML given in Equation (8). The second component, $\Omega_{22}|\Xi|/\Xi_{22}^3$ represents the increase in asymptotic variance due to the failure of rank restriction; when Assumption LIV holds, $|\Xi| = 0$, and this term drops out. The last term represents the asymptotic efficiency loss relative to LIML when Assumption LIV holds; the price for the for the extra robustness is that the estimator does not use the information contained in (21a) when the rank restriction holds. This price is zero under the standard asymptotics when $\tau = 0$, and also when there is no endogeneity, since in this case $0 = \mathbb{E}[\epsilon v_{2i}] = \mathbb{E}[b'Vv_{2i}] = \Omega_{12} - \beta\Omega_{22}$.

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 be positive semi-definite,[10] which is equivalent to the constraint $\Xi_{11,n} \geq \beta_n^2\Xi_{22,n}$. If the random-effects weight matrix is used, then the resulting estimator will be a mixture between $\hat{\beta}_{\text{LIML}}$ and $\hat{\beta}_{\text{UMD}}$: 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 $\hat{\Xi}_{11} = \hat{\beta}^2\hat{\Xi}_{22}$, and the estimator equals $\hat{\beta}_{\text{LIML}}$. When Ξ is full rank, then the constrain won't bind in large samples, and the estimator will be asymptotically equivalent to $\hat{\beta}_{\text{UMD}}$. However, when Ξ 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 about β complicated. I provide additional details on how to do inference using this estimator in the Supplementary Appendix.

7 Tests of overidentifying restrictions

Assumption LIV imposes a proportionality restriction on the reduced form (2) that $\pi_{1,n} = \pi_{2,n}\beta$. If Assumption LIV does not hold, the reduced-form coefficients are unrestricted. A variety of tests

[10] Since Ξ is a matrix of second moments of $\pi_{2,n}$ and $\pi_{1,n}$, it has to be positive semi-definite.

of this restriction that work under the standard asymptotics that hold k_n and ℓ_n fixed have been proposed in the literature. First, I will discuss the robustness of three such tests to the presence of many instruments and many covariates. I will then relate these tests to a test based on the minimum distance objective function.

The most popular test, due to Sargan (1958), is based on the observation that the nR^2 from regressing the estimated residuals in the structural equation (1) on the instruments and covariates is asymptotically distributed according to χ^2_{k-1} under Assumption LIV and standard asymptotics that hold the number of instruments and covariates fixed, so that $k_n = k, \ell_n = \ell$. If LIML is used to estimate β and δ_n, the estimated residuals can be written as $(I - W'(W'W)^{-1}W)Y\hat{\beta}_{\text{LIML}}$, and consequently, the R^2 is given by

$$\hat{J}_{\text{S}} = \frac{\hat{b}'_{\text{LIML}} T \hat{b}_{\text{LIML}}}{\hat{b}'_{\text{LIML}}(T - (k_n/n)S)\hat{b}_{\text{LIML}}} = \frac{m_{\min}}{1 - k_n/n - \ell_n/n + m_{\min}}.$$

The Sargan test therefore rejects if $n\hat{J}_{\text{S}}$ is greater than $q^{\chi^2_{k-1}}_{1-\text{ns}}$, the $1-\text{ns}$ quantile of a χ^2_{k-1} distribution where ns denotes the desired nominal size.

A closely related alternative is the generalized likelihood ratio test based on the limited information likelihood of Anderson and Rubin (1949). The test statistic is given by $n\hat{J}_{\text{AR}}$, where $\hat{J}_{\text{AR}} = \log(nm_{\min}/(n - k_n - \ell_n) + 1)$. It is also asymptotically distributed according to χ^2_{k-1} under the null and standard asymptotics.

Third, Cragg and Donald (1993) suggest a test based on the minimum distance objective function (17). They show that the minimum of the objective function is given by

$$\hat{J}_{\text{CD}} = \frac{\hat{b}'_{\text{LIML}} T \hat{b}_{\text{LIML}}}{\hat{b}'_{\text{LIML}} S \hat{b}_{\text{LIML}}} = m_{\min}.$$

Compared to the Sargan test statistic, $\hat{J}_{\text{CD}}$ replaces $T - (k_n/n)S$ by S in the denominator. Cragg and Donald (1993) also show that $nm_{\min} \Rightarrow \chi^2_{k-1}$ under standard asymptotics.

All three tests are equivalent in the sense that they all reject for large values of $m_{\min}$. Therefore, the only difference between them in finite samples is how well the chi-squared approximation controls size in each case. While under standard asymptotics their asymptotic distributions coincide and therefore do not provide any guidance as to which test has the best size control, allowing for $\alpha_k, \alpha_\ell > 0$ reverses this conclusion:

Lemma 2. *Consider the model* (1)–(3)*. Then, under Assumptions LIV, N and MI:*

$$n^{1/2}\left(\hat{J}_{\text{S}} - \tfrac{\alpha_k}{1-\alpha_\ell}\right) \Rightarrow \mathcal{N}\left(0, \tfrac{2\alpha_k(1-\alpha_k-\alpha_\ell)}{(1-\alpha_\ell)^3}\right), \tag{25}$$

$$n^{1/2}\left(\hat{J}_{\text{AR}} - \log\left(\tfrac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right)\right) \Rightarrow \mathcal{N}\left(0, 2\tau/(1-\alpha_\ell)^2\right), \tag{26}$$

$$n^{1/2}\left(\hat{J}_{\text{CD}} - \alpha_k\right) \Rightarrow \mathcal{N}(0, 2\tau), \tag{27}$$

where $\tau = \frac{\alpha_k(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell}$. *Moreover, if* $\alpha_k > 0$,

$$\mathbb{P}\left(n\hat{J}_{\text{S}} \geq q^{\chi^2_{k_n-1}}_{1-ns}\right) \to \begin{cases} \Phi(\Phi^{-1}(ns)/\sqrt{1-\alpha_k}) & \text{if } \alpha_\ell = 0, \\ 1 & \text{otherwise.} \end{cases}$$

$$\mathbb{P}\left(n\hat{J}_{\text{AR}} \geq q^{\chi^2_{k_n-1}}_{1-ns}\right) \to 1,$$

$$\mathbb{P}\left(n\hat{J}_{\text{CD}} \geq q^{\chi^2_{k_n-1}}_{1-ns}\right) \to \Phi\left(\Phi^{-1}(ns)\sqrt{(1-\alpha_k-\alpha_\ell)/(1-\alpha_\ell)}\right),$$

where $\Phi(\cdot)$ *is the cdf of a standard Normal distribution.*

Anatolyev and Gospodinov (2011) and Anatolyev (2011) derive the results for the Sargan test. The results for the Anderson-Rubin overidentification test and the Cragg-Donald test are new.

When $\alpha_k > 0$ and $\alpha_\ell = 0$, 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 of the Sargan test to match the asymptotic size with the nominal size—instead of using the $q^{\chi^2_{k_n-1}}_{1-\text{ns}}$ critical value, they suggest using $q^{\chi^2_{k_n-1}}_{1-\Phi(\sqrt{1-\alpha_k}\Phi^{-1}(\text{ns}))}$. As the Lemma demonstrates, the problem with this solution is that it breaks down when $\alpha_\ell > 0$: in this case, the size distortion of the test gets *worse* as the sample size increases. Furthermore, it is no longer possible to adjust the critical value to correct the asymptotic size because the test statistic is centered at the wrong value—$\alpha_k/(1-\alpha_\ell)$ rather than $\mathbb{E}[\chi^2_{k_n-1}/n] \approx \alpha_k$. Similar conclusions apply to the Anderson-Rubin overidentification test.

The Cragg-Donald test is also size-distorted, although the size distortion is rather small. With $\alpha_k = \alpha_\ell = 0.1$ for example, the asymptotic size of the test with nominal size 0.05 is given by 0.06. Moreover, we can apply the Anatolyev and Gospodinov (2011) adjustment to the critical value to correct the size distortion. In particular, comparing $nm_{\min}$ against the $1-\Phi(\sqrt{(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell)}\Phi^{-1}(\text{ns}))$ quantile of the $\chi^2_{k_n-1}$ distribution will yield a critical value that will control size under standard as well as many-instrument asymptotics.

An alternative to size-correcting existing tests of overidentification to make them robust to the presence of many instruments is to make use of the invariant model. In this model, a test of Assumption LIV is equivalent to testing whether 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 (23) minimized subject to the restriction that $|\Xi|$ is reduced rank with its value when it is minimized subject to $|\Xi|$ being positive definite. When $\hat{W}_{\text{RE}}$ is used as a weight

matrix, the test statistic is given by

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

The test is equivalent to the Sargan, Anderson-Rubin and Cragg-Donald tests of overidentification in the sense that all tests reject for large values of $m_{\min}$. It follows from (27) that in large samples, the $n\hat{J}_{\text{MD}}$ will be distributed as a mixture between a χ^2_1 distribution scaled by 2τ, and a degenerate distribution with point mass on 0. Moreover, it follows from the previous discussion that a test that rejects whenever $n\hat{J}_{\text{MD}}/2\tau \geq q^{\chi^2_1}_{1-\text{ns}}$ will be asymptotically equivalent to the size-corrected Cragg-Donald test. This result suggests that the preferred test for overidentifying restrictions is given by the size-corrected Cragg-Donald test.

8 Conclusion

In this paper, I outlined an integrated likelihood approach to inference in the instrumental variables model when the number of instruments is large. This approach addresses the incidental parameter problem that the large number of instruments create. It it principled and unified, as it explicitly uses an invariance argument to deal with the incidental parameters and it is based on a well-motivated and well-behaved objective function. I show that this integrated likelihood coincides with the random effects likelihood of Chamberlain and Imbens (2004), and that the maximum likelihood estimator of β coincides with LIML. Moreover, maximizing this integrated likelihood is equivalent to minimizing a minimum distance objective function that imposes a rank restriction on the matrix of second moments of the reduced-form coefficients. I use this equivalence to show that when the reduced-form errors are not Normal, a minimum distance estimator with respect to an efficient weight matrix is more efficient than LIML. Finally, I show that when the rank restriction is relaxed, the resulting minimum distance estimator corresponds to a version of the bias-corrected two-stage least squares estimator.

Appendix A Definitions and identities

First I state a couple of simple identifies that are used throughout the appendix. Then, in Appendix B I state and prove some auxiliary Lemmata that are helpful for proving the main results. The propositions and theorems stated in the text are derived in Appendix C.

Let $e_1 = (1,0)'$, and $e_2 = (0,1)'$. For any symmetric matrix $\Omega \in \mathbb{R}^{2\times 2}$, and vectors $a = (\beta, 1)'$ and $b = (1, -\beta)'$, $\beta \in \mathbb{R}$:

$$Q_{\mathcal{S}}(\beta,\Omega) + Q_{\mathcal{T}}(\beta,\Omega) = \operatorname{tr}(\Omega^{-1}T), \tag{28a}$$

$$|\Omega|\, a\Omega^{-1}a = b'\Omega b, \tag{28b}$$

$$|nT + (n - k_n - \ell_n S)| = \left(n^2 m_{\max} m_{\min} + (n-k_n-\ell_n)^2 + n(n-k_n-\ell_n)\operatorname{tr}(S^{-1}T)\right)|S|\,. \tag{28c}$$

All equalities follow from simple algebra. Secondly, I use the following properties of the Kronecker product:

$$a \otimes b' = b' \otimes a = ab', \qquad \operatorname{vec}(ACB) = (B' \otimes A)\operatorname{vec}(C), \tag{29}$$

for some vectors $a, b \in \mathbb{R}^d$, and conformable matrices A, B, C.

Denote the duplication, elimination, and commutation matrices by D_d, L_d and K_d (see Magnus and Neudecker (1980) for definitions of these matrices). Let $N_d = (I_{d^2} + K_{dd})/2$ be the symmetrizer matrix. Then for arbitrary matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$ (Magnus and Neudecker, 1979, 1980),

$$K_{m1} = K_{1m} = I_m, \qquad (B \otimes A)K_{qn} = K_{pm}(A \otimes B), \tag{30a}$$

$$K_d D_d = D_d, \qquad D_d L_d N_d = N_d. \tag{30b}$$

Appendix B Auxiliary Lemmata

Lemma 3. *Suppose $P \sim \mathcal{W}_d(\nu_n, V, M_n)$, a $d \times d$-non-central Wishart distribution with ν_n degrees of freedom, scale matrix V, and non-centrality parameter M_n.*[11] *Then:*

(i) *[Magnus and Neudecker, 1979, Theorem 4.4] The mean and variance of P are given by:*

$$\mathbb{E}[P] = \nu_n V + M_n, \qquad \operatorname{var}(\operatorname{vec}(P)) = 2N_d\left[\nu_n(V\otimes V) + V \otimes M_n + M_n \otimes V\right],$$

where N_d is the symmetrizer matrix.

(ii) *Suppose $M_n/n \to M$, and $\nu_n/n = \alpha + o(n^{-1/2})$ where $\alpha < 1$. Then, as $n \to \infty$*

$$\sqrt{n}\operatorname{vec}\left(P/n - \mathbb{E}[P/n]\right) \Rightarrow \mathcal{N}_{d^2}\left(0, 2N_d[\alpha(V\otimes V) + V\otimes M + M \otimes V]\right).$$

[11] Hence, if $X_i \sim \mathcal{N}_d(\mu_i, V)$, then $\sum_{i=1}^{\nu_n} X_i X_i' \sim \mathcal{W}_d(\nu_n, V, \sum_{i=1}^{\nu_n}\mu_i\mu_i')$

Proof. To prove part (ii), decompose P as $P = \sum_{i=1}^{\nu_n} X_i X_i'$, where $X_i \sim \mathcal{N}_d(\mu_i, V)$ such that $M_n = \sum_i \mu_i \mu_i'$. Suppose first that $\alpha > 0$. Then it follows by the Central Limit Theorem that:

$$\nu_n^{-1/2} \operatorname{vec}(P - \mathbb{E}[P]) \Rightarrow \mathcal{N}\left(0, 2N_d\left[(V \otimes V) + V \otimes M/\alpha + M \otimes V/\alpha\right]\right),$$

which implies the result. If $\alpha = 0$, then

$$\operatorname{var} \operatorname{vec}\left(n^{-1/2} \textstyle\sum_i (X_i - \mu_i)(X_i - \mu_i)'\right) \to 0,$$

so that $\operatorname{vec}\left(n^{-1/2} \sum_i (X_i - \mu_i)(X_i - \mu_i)' - \nu_n V\right) = o_p(1)$. Therefore, we have:

$$\begin{aligned}
\sqrt{n} \operatorname{vec}(P/n - \mathbb{E}[P/n]) &= n^{-1/2} \operatorname{vec}\left(\textstyle\sum_i (X_i - \mu_i)(X_i - \mu_i)' - \nu_n V + \sum_i (X_i \mu_i' + \mu_i X_i' - 2\mu_i \mu_i')\right) \\
&= n^{-1/2} \textstyle\sum_i \operatorname{vec}\left(X_i \mu_i' + \mu_i X_i' - 2\mu_i \mu_i'\right) + o_p(1) \\
&= n^{-1/2} \textstyle\sum_i \left((X_i - \mu_i) \otimes \mu_i + \mu_i \otimes (X_i - \mu_i)\right) + o_p(1).
\end{aligned} \tag{31}$$

Now,

$$\begin{aligned}
&\mathbb{E}[(X_i - \mu_i) \otimes \mu_i + \mu_i \otimes (X_i - \mu_i)]^2 \\
&\quad = V \otimes \mu_i \mu_i' + \mathbb{E}[(X_i - \mu_i)\mu_i' \otimes \mu_i (X_i - \mu_i)'] + [\mu_i (X_i - \mu_i)' \otimes (X_i - \mu_i)\mu_i'] + \mu_i \mu_i' \otimes V \\
&\quad = (I + K_{dd})\left(V \otimes \mu_i \mu_i' + \mu_i \mu_i' \otimes V\right),
\end{aligned}$$

where the last line uses the identity $ab' \otimes ba' = a \otimes (bb' \otimes a) = K_{dd}(bb' \otimes aa')$ for any vectors $a, b \in \mathbb{R}^d$ that follows from Equations (29) and (30a). Hence

$$\sum_i \left((X_i - \mu_i) \otimes \mu_i + \mu_i \otimes (X_i - \mu_i)\right) \sim \mathcal{N}_{d^2}\left(0, (I + K_{dd})\left(V \otimes M_n + M_n \otimes V\right),\right)$$

which, combined with (31), yields the result. □

Corollary 1. *Consider the model* (1)–(3) *and suppose Assumptions LIV, N and MI hold. Then:*

$$\begin{aligned}
\sqrt{n} \operatorname{vec}(S - \Omega) &\Rightarrow \mathcal{N}_4\left(0, \frac{1}{1 - \alpha_k - \alpha_\ell} 2N_2(\Omega \otimes \Omega)\right) \\
\sqrt{n} \operatorname{vec}\left(T - \alpha_k \Omega - \frac{\lambda_n}{a'\Omega^{-1}a} aa'\right) &\Rightarrow \mathcal{N}_4\left(2N_2 \Phi\right)
\end{aligned}$$

where $\Phi = \alpha_k \Omega \otimes \Omega + \frac{\lambda}{a'\Omega^{-1}a} \Omega \otimes (aa') + \frac{\lambda}{a'\Omega^{-1}a} (aa') \otimes \Omega$, *and* N_2 *is the symmetrizer matrix.*

Proof. The result follows from Lemma 3 (ii). □

Lemma 4. *Consider an invertible matrix* $V \in \mathbb{R}^{d \times d}$, *a vector* $m \in \mathbb{R}^d$ *and a constant* c. *Then:*

$$\begin{aligned}
\left(V \otimes V + c(mm') \otimes (mm')\right)^{-1} &= V^{-1} \otimes V^{-1} - \frac{c(V^{-1}mm'V^{-1}) \otimes (V^{-1}mm'V^{-1})}{1 + c(m'V^{-1}m)^2}, \\
(D_d'(V \otimes V + c(mm') \otimes (mm'))D_d)^{-1} &= L_d N_d \left(V \otimes V + c(mm') \otimes (mm')\right)^{-1} N_d L_d', \\
(L_d N_d'(V \otimes V + c(mm') \otimes (mm'))L_d')^{-1} &= D_d' \left(V \otimes V + c(mm') \otimes (mm')\right)^{-1} D_d.
\end{aligned}$$

Proof. The first identity can be checked by direct calculation. The second identity follows from Lemma 4.4 in Magnus and Neudecker (1980). □

Lemma 5. *Consider the quadratic form $Q = (V + M)'P(M + V)$, where $P \in \mathbb{R}^{n\times n}$ is symmetric with $\operatorname{tr}(P^2) = r$, $V, M \in \mathbb{R}^{n\times g}$, the rows $v_i \sim [0, \Omega]$ of V are iid with finite fourth moments, and M is non-random.*

(i) *The variance of Q is given by:*

$$\begin{aligned}\operatorname{var}(\operatorname{vec}(Q)) &= (I_g + K_{gg})(M'PM \otimes \Omega + \Omega \otimes M'PM + (r - d'd)\Omega \otimes \Omega) \\ &+ d'd\left[\mathbb{E}(vv') \otimes (vv') - \operatorname{vec}(\Omega)\operatorname{vec}(\Omega)'\right] + \mathbb{E}[vv' \otimes (\overline{m}v' + v\overline{m}')] + \mathbb{E}[(\overline{m}v' + v\overline{m}') \otimes vv'],\end{aligned}$$

where $\overline{m} = M'P\operatorname{diag}(P)$ and $d = \operatorname{diag}(P)$.

(ii) *Suppose in addition that for some constant D,*

(a) $\sup_{i\geq 1}\|m_i\| < D < \infty$;

(b) $M'PM/n \xrightarrow{p} \Lambda$;

(c) $\overline{m}/n \to \mu$;

(d) $r/n \to \tau_r$ *and* $\operatorname{tr}(P) = \tau_P + o(n^{-1/2})$;

(e) $d'd/n \to \delta$; *and*

(f) $\sup_n \sup_{1\leq i\leq n} \sum_{s=1}^n |p_{si}| < D < \infty$

Then:

$$n^{-1/2}\operatorname{vec}(Q - M'PM - \operatorname{tr}(P)\Omega) \Rightarrow \mathcal{N}(0, \operatorname{plim}(\operatorname{var}(\operatorname{vec}(Q))/n)).$$

Proof. Proof of part (i) follows by a tedious but straightforward calculation, and it is given in the Supplementary Appendix. Proof of part (ii) follows from part (i) and Theorem 1 in van Hasselt (2010). □

Appendix C Proofs

Proof of Proposition 1. The density of T is proportional to (Moreira, 2009, Theorem 4.1):

$$f_T(T \mid \beta, \lambda_n, \Omega) \propto e^{-\frac{n}{2}\left(\lambda_n + \operatorname{tr}(\Omega^{-1}T)\right)} |\Omega|^{-k_n/2} \left(n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta)}\right)^{-(k_n-2)/2} I_{(k_n-2)/2}\left(n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta)}\right), \tag{32}$$

where $I_\nu(\cdot)$ is modified Bessel function of the first kind of order ν. Using the integral representation of the Bessel function (Abramowitz and Stegun, 1965, Equation 9.6.18, p. 376),

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

and $\bar{\Gamma}$ is the gamma function. The density (32) can therefore be written (up to a constant) as:

$$f_T(T \mid \beta, \lambda_n, \Omega) \propto e^{-\frac{n}{2}\left(\lambda_n + \operatorname{tr}(\Omega^{-1}T)\right)} |\Omega|^{-k_n/2} G_{k_n}\left(n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta)}\right).$$

Combining this expression with the density for S, which, by Equation (6), is given by

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

yields the invariant likelihood

$$\log \mathcal{L}_{\text{INV},n}(\beta,\lambda_n,\Omega;S,T) \propto -\frac{1}{2}\left((n-\ell_n)\log|\Omega| + \operatorname{tr}(\Omega^{-1}\tilde{S}) + n\lambda_n - 2\log G_{k_n}(n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta,\Omega)})\right),$$

where $\tilde{S} = (n-k_n-\ell_n)S + nT$. Dropping the k_n index from the G function to avoid clutter, the derivative with respect to Ω is given by:

$$\frac{\partial \log \mathcal{L}_{\text{INV},n}}{\partial \Omega} = \frac{1}{2}\left[\Omega^{-1}\tilde{S}\Omega^{-1} - (n-\ell_n)\Omega^{-1} - \frac{G'(\cdot)}{G(\cdot)}\frac{n\lambda_n^{1/2}}{Q_{\mathcal{T}}(\beta,\Omega)^{1/2}}\left(\Omega^{-1}T\Omega^{-1} - \frac{Q_{\mathcal{S}}(\beta,\Omega)}{b'\Omega b}bb'\right)\right], \tag{33}$$

where the derivative $\partial Q_{\mathcal{T}}(\beta,\Omega)/\partial\Omega$, given by the expression in parentheses, is computed using the identity (28a). 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 function, it follows from the expression for the invariant likelihood that:

$$\hat{\beta}_{\lambda_n} = \operatorname*{argmax}_{\beta} Q_{\mathcal{T}}(\beta,\hat{\Omega}_{\lambda_n}) = \operatorname*{argmin}_{\beta} Q_{\mathcal{S}}(\beta,\hat{\Omega}_{\lambda_n}). \tag{34}$$

Secondly, the derivative (33) evaluated at $(\hat{\beta}_{\lambda_n},\hat{\Omega}_{\lambda_n})$ has to be equal to zero. Pre-multiplying and post-multiplying Equation (33) by $\hat{b}'_{\lambda_n}\hat{\Omega}_{\lambda_n}$ and $\hat{\Omega}_{\lambda_n}\hat{b}_{\lambda_n}$ therefore yields:

$$(n-\ell_n)\hat{b}'_{\lambda_n}\hat{\Omega}_{\lambda_n}\hat{b}_{\lambda_n} = \hat{b}'_{\lambda_n}\tilde{S}\hat{b}_{\lambda_n} \tag{35}$$

This implies:

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

as required. By similar arguments, Equations (34) and (35) must also hold when the likelihood is maximized over λ_n as well, so that $\hat{\beta}_{\text{INV}} = \hat{\beta}_{\text{LIML}}$. □

Proof of Proposition 2. It follows from Equation (14) that the log-likelihood, parametrized in terms of (ψ,λ,Ω) where $\psi = \Omega^{-1}a$, can be written as:

$$\log \mathcal{L}_{\text{RE},n}(\psi,\lambda,\Omega) = -\frac{1}{2}\left((n-\ell_n)\log|\Omega| + k_n\log\left(1+\frac{n}{k_n}\lambda\right) + \operatorname{tr}(\Omega^{-1}\tilde{S}) - \frac{n\lambda}{k_n/n+\lambda}Q_{\mathcal{S}}(\psi,\Omega)\right),$$

where $Q_{\mathcal{S}}(\psi,\Omega) = \psi'T\psi/(\psi'\Omega\psi)$, and $\tilde{S} = nT + (n-k_n-\ell_n)S$. The derivative with respect to λ is given by

$$\frac{\partial}{\partial\lambda}\log \mathcal{L}_{\text{RE},n}(\psi,\lambda,\Omega) = -\frac{1}{2}\frac{k_n}{k_n/n+\lambda}\left(1 - \frac{Q_{\mathcal{S}}(\psi,\Omega)}{k_n/n+\lambda}\right).$$

Now suppose that

$$Q_{\mathcal{S}}(\psi,\Omega) > k_n/n. \tag{36}$$

Then the ML estimator of λ with ψ and Ω given is given by:

$$\hat{\lambda}_{\psi,\Omega} = Q_{\mathcal{S}}(\psi,\Omega) - k_n/n.$$

Therefore, the likelihood with λ concentrated out is given by:

$$\log \mathcal{L}_{\text{RE},n}(\psi, \hat{\lambda}_{\psi,\Omega}, \Omega) \propto -\frac{1}{2}\left((n-\ell_n)\log|\Omega| + k_n \log\left(Q_{\mathcal{S}}(\psi,\Omega)\right) + \text{tr}(\Omega^{-1}\tilde{S}) - nQ_{\mathcal{S}}(\psi,\Omega)\right).$$

The derivative with respect to Ω is given by:

$$\frac{\partial}{\partial \Omega}\log \mathcal{L}_{\text{RE},n}(\psi, \hat{\lambda}_{\psi,\Omega}, \Omega) = \Omega^{-1}\tilde{S}\Omega^{-1} - (n-\ell_n)\Omega^{-1} + \frac{k_n - nQ_{\mathcal{S}}(\psi,\Omega)}{\psi'\Omega\psi}\psi\psi'. \tag{37}$$

Setting the derivative to zero, and pre-multiplying it by $\hat{\Omega}_\psi$ and $\psi'\hat{\Omega}_\psi$, and post-multiplying it by $\hat{\Omega}_\psi\psi$ yields:

$$\psi' S\psi = \psi'\hat{\Omega}_\psi\psi, \qquad \text{and} \qquad \frac{1}{(n-k_n-\ell_n+nQ_{\mathcal{S}}(\psi,S))}\tilde{S}\psi = \Omega\psi, \tag{38}$$

where $Q_{\mathcal{S}}(\psi, S) = Q_{\mathcal{S}}(\psi, \hat{\Omega}_\psi)$. Plugging these expressions back into (37) yields:

$$(n-\ell_n)\hat{\Omega}_\psi = \tilde{S} + \frac{k_n - nQ_{\mathcal{S}}(\psi,S)}{\psi' S\psi}\frac{1}{(n-k_n-\ell_n+nQ_{\mathcal{S}}(\psi,S))^2}\tilde{S}\psi\psi'\tilde{S}. \tag{39}$$

Hence:

$$|\hat{\Omega}_\psi| = \frac{1}{(n-\ell_n)}\frac{|\tilde{S}|}{n-k_n-\ell_n+nQ_{\mathcal{S}}(\psi,S)} \qquad \text{tr}(\hat{\Omega}_\psi^{-1}\tilde{S}) = 2(n-\ell_n) - k_n + nQ_{\mathcal{S}}(\psi,S)$$

Therefore, the likelihood with both λ and Ω concentrated out is given by:

$$\log \mathcal{L}_{\text{RE},n}(\psi, \hat{\lambda}_\psi, \hat{\Omega}_\psi) \propto \frac{1}{2}\left((n-\ell_n)\log(n-k_n-\ell_n+nQ_{\mathcal{S}}(\psi,S)) - k_n\log\left(Q_{\mathcal{S}}(\psi,S)\right)\right)$$

This expression is increasing in $Q_{\mathcal{S}}$ if $Q_{\mathcal{S}} > k_n/n$. The maximum is obtained at $Q_{\mathcal{S}}(\hat{\psi}_{\text{RE}}, S) = m_{\max}$, Equation (36) holds, and $\hat{\lambda}_{\text{RE}} = m_{\max} - k_n/n$.

The estimator $\hat{\psi}_{\text{RE}}$ is given by the eigenvector that corresponds to the $m_{\max}$, the larger eigenvalue of $S^{-1}T$. Therefore, $S^{-1}T\hat{\psi}_{\text{RE}} = m_{\max}\hat{\psi}_{\text{RE}}$. Secondly, since $Q_{\mathcal{S}}(\hat{\psi}_{\text{RE}}, S) = Q_{\mathcal{T}}(\hat{\beta}_{\text{LIML}}, S)$, it follows that $\hat{\psi}_{\text{RE}} = S^{-1}\hat{a}_{\text{LIML}}$. Combining these two observations yields $\tilde{S}\hat{\psi}_{\text{RE}} = S(n-k_n-\ell_n+nm_{\max})\psi = (n-k_n-\ell_n+nm_{\max})\hat{a}_{\text{LIML}}$, and $\hat{\psi}'_{\text{RE}}S\hat{\psi}_{\text{RE}} = \hat{a}'_{\text{LIML}}S\hat{a}_{\text{LIML}}$. Plugging these result into Equations (38) and (39) yields:

$$\hat{a}_{\text{RE}} = \hat{\Omega}_{\text{RE}}\hat{\psi}_{\text{RE}} = \hat{a}_{\text{LIML}} \qquad \hat{\Omega}_{\text{RE}} = \frac{1}{n-\ell_n}\left(\tilde{S} - \frac{\hat{\lambda}_{\text{RE}}}{\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}}}\hat{a}_{\text{RE}}\hat{a}'_{\text{RE}}\right)$$

Next, to prove consistency of $\hat{\lambda}_{\text{RE}}$, note that by continuity of the trace operator, and Corollary 1

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

The consistency of $\hat{\Omega}_{\text{RE}}$ follows by consistency of $\hat{\lambda}_{\text{RE}}$ and $\hat{\beta}_{\text{RE}}$, Corollary 1, and Slutsky's Theorem. □

Proof of Proposition 3. To avoid clutter, I write $(\hat{\beta}, \hat{\lambda}, \hat{\Omega})$ and $\hat{Q}_{\mathcal{S}}$ in place of $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$ and $Q_{\mathcal{S}}(\hat{\beta}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$.

The score equations based on the random-effects likelihood (14) are given by:

$$\mathcal{S}_\beta(\beta,\lambda,\Omega) = \frac{n\lambda}{k_n/n+\lambda}\frac{e_2'\left(T - Q_\mathcal{S}(\beta,\Omega)\Omega\right)b}{b'\Omega b}, \tag{40a}$$

$$\mathcal{S}_\lambda(\beta,\lambda,\Omega) = -\frac{1}{2}\frac{k_n}{k_n/n+\lambda}\left(1 - \frac{Q_\mathcal{T}(\beta,\Omega)}{k_n/n+\lambda}\right), \tag{40b}$$

$$\mathcal{S}_\Omega(\beta,\lambda,\Omega) = \frac{1}{2}D'\operatorname{vec}\left[\Omega^{-1}\tilde{S}\Omega^{-1} - (n-\ell_n)\Omega^{-1} - \frac{n\lambda}{k_n/n+\lambda}\left(\Omega^{-1}T\Omega^{-1} - \frac{Q_\mathcal{S}}{b'\Omega b}bb'\right)\right], \tag{40c}$$

where $\tilde{S} = (n-k_n-\ell_n)S + nT$. The Hessian, evaluated at ML estimates, is given by:

$$H_{\text{RE}}(\hat\beta,\hat\lambda,\hat\Omega) = \begin{pmatrix} \frac{n\hat\lambda}{(k_n/n+\hat\lambda)\hat b'\hat\Omega\hat b}(\hat Q_\mathcal{S}\hat\Omega_{22} - T_{22}) & 0 & \hat H_{1,3:5} \\ 0 & -\frac{1}{2}\frac{k_n}{(k_n/n+\hat\lambda)^2} & \hat H_{2,3:5} \\ \hat H_{1,3:5}' & \hat H_{2,3:5}' & \hat H_{3:6,3:5} \end{pmatrix},$$

where

$$H_{1,3:5} = -\frac{1}{2}\frac{\hat c(n-\ell_n)}{\hat b'\hat\Omega\hat b}\left(2\frac{e_2'\hat\Omega\hat b}{\hat b'\hat\Omega\hat b}\hat b\otimes\hat b - \hat b\otimes e_2 - e_2\otimes b\right)' D,$$

$$\hat H_{2,3:5} = -\frac{1}{2}\frac{k_n}{(k_n/n+\hat\lambda)^2}\left(\frac{\hat Q_\mathcal{S}}{\hat b\hat\Omega\hat b}\hat b\otimes\hat b - \operatorname{vec}(\hat\Omega^{-1}T\hat\Omega^{-1})\right)' D,$$

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

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

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

Applying Lemma 4 yields:

$$\hat H_{3:5,3:5}^{-1} = -\frac{2}{n-\ell_n}LN\left[\left(\hat\Omega + \frac{\hat c}{1-\hat c}\frac{\hat\Omega\hat b\hat b'\hat\Omega}{\hat b'\hat\Omega\hat b}\right)\otimes\left(\hat\Omega + \frac{\hat c}{1-\hat c}\frac{\hat\Omega\hat b\hat b'\hat\Omega}{\hat b'\hat\Omega\hat b}\right) + \frac{\hat c^2 - 2\hat c}{(1-\hat c)^2}\frac{\hat\Omega\hat b\hat b'\hat\Omega\otimes\hat\Omega\hat b\hat b'\hat\Omega}{(\hat b'\hat\Omega\hat b)^2}\right]NL',$$

where L is the elimination matrix and N is the symmetrizer matrix. It follows that

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

Finally, since $\hat H_{2,3:6}\hat H_{3:6,3:6}^{-1}\hat H_{1,3:6}' = 0$, Equation (41) combined with the expression above yields

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

which yields the result.

Now, consider its probability limit. We have:

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

since $m_{\min} \xrightarrow{p} \alpha_k$. 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

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

which completes the proof. □

Proof of Proposition 4. The objective function evaluates as:

$$\mathcal{Q}_n(\beta, \Xi_{22}; \hat{W}_{\text{RE}}) = \text{tr}((TS^{-1} - (k_n/n)I_2)^2) - 2\Xi_{22}a'S^{-1}TS^{-1}a + 2(k_n/n)\Xi_{22}a'S^{-1}a + \Xi_{22}^2(a'S^{-1}a)^2. \tag{42}$$

Setting derivative wrt Ξ_{22} to zero yields

$$\hat{\Xi}_{22}(\beta) = \frac{Q_{\mathcal{T}}(\beta, S) - k_n/n}{a'S^{-1}a}. \tag{43}$$

Therefore, the objective function with Ξ_{22} concentrated out is given by

$$\mathcal{Q}_n(\beta, \hat{\Xi}_{22}(\beta)) = \text{tr}((TS^{-1} - (k_n/n)I_2)^2) - (Q_{\mathcal{T}}(\beta, S) - k_n/n)^2,$$

which is maximized at $\max_\beta Q_{\mathcal{T}}(\beta, S)$, since by the identity (28a), $\text{tr}(S^{-1}T) > 2k_n/n$ implies $Q_{\mathcal{T}}(\beta, S) - k_n/n > k_n/nQ_{\mathcal{S}}(\beta, S)$, and $\min_\beta Q_{\mathcal{S}}(\beta, S) = \min_\beta Q_{\mathcal{T}}(\beta, S) = m_{\min}$. Hence, $\hat{\beta}_{\text{MD}} = \hat{\beta}_{\text{LIML}}$. 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 first line follows from the definition of $\hat{\Omega}_{\text{RE}}$ and $\hat{\lambda}_{\text{RE}}$ given in Proposition 2, the second line follows by the Woodbury identity, the third line follows from Equation (28b), and the fourth line follows from Equation (28c). Therefore, we have that

$$\hat{\Xi}_{22,\text{RE}} = \frac{\hat{\lambda}_{\text{RE}}}{\hat{a}'_{\text{RE}}\hat{\Omega}^{-1}_{\text{RE}}\hat{a}_{\text{RE}}} = \frac{m_{\max} - k_n/n}{\hat{a}'_{\text{RE}}S^{-1}\hat{a}_{\text{RE}}}, \tag{45}$$

which, by Equation (43) equals $\hat{\Xi}_{22,\text{MD}}$, as asserted. □

Proof of Proposition 5. Let $W_{\bar{c},t}$, $t = c/\Xi_{22}$, denote the probability limit of $\hat{W}_n$ given in the statement of the Proposition. This limit weight can be written as

$$W_{\bar{c},t} = \bar{c}D_2'\Phi_t^{-1}D_2, \qquad \text{where} \qquad \Phi_t = \Omega \otimes \Omega + \Omega \otimes tmm' + tmm' \otimes \Omega,$$

and $m = \Xi_{22}^{1/2} a$. By Lemma 4, 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 1, the asymptotic variance of the moment condition

$$\operatorname{vech}(T - (k_n/n)S - \Xi_{22,n} aa') \tag{46}$$

is given by

$$\Delta = 2L_2 N_2 \left[\tau \Omega \otimes \Omega + \Omega \otimes (mm') + (mm') \otimes \Omega \right] L_2'. \tag{47}$$

If $\alpha_k > 0$, then Δ is invertible, and a necessary and sufficient condition for optimality is that for some matrix C_t (Newey and McFadden, 1994, Section 5.2)

$$G' W_{\bar{c},t} = C_t G' \Delta^{-1}, \tag{48}$$

where G is the derivative of the moment condition (46), given by:

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

To prove the Proposition for the case $\alpha_k > 0$, we therefore need to find C_t such that (48) holds. We have:

$$W_{\bar{c},t} G = -\bar{c} D_2' \Phi_t^{-1} (m \otimes M + M \otimes m),$$
$$\Delta^{-1} G = -\frac{1}{2\tau} D_2' \Phi_{1/\tau}^{-1} (m \otimes M + M \otimes m),$$

since $\Delta^{-1} = \frac{1}{2\tau} D_2' \Phi_{1/\tau}^{-1} D_2$ by Lemma 4. After some algebra, it therefore follows that the equality (48) holds with

$$C_t = \frac{2\bar{c}}{1+t\lambda} \left((\tau + \lambda) I_2 + \frac{1 - \tau t}{(1+2t\lambda)} m M'^{-1} \otimes M' \Omega^{-1} m \right).$$

If $\tau = 0$, then the asymptotic variance Δ given in Equation (47) is degenerate, since one of the three moment conditions given in Equation (46) is asymptotically redundant: the first moment condition equals 2β times the second minus β^2 times the third. In this case, any positive definite weight matrix will be optimal, and in particular W_t is optimal. □

Proof of Lemma 1. Part (i) of the Lemma follows from Lemma 5. Next, it it follows from Lemma A.5 in Anatolyev (2011) that

$$\sum_i \hat{v}_i \otimes \hat{v}_i \otimes \hat{v}_i' = \sum_{ij} (M_{ij})^3 \mathbb{E}[v_i \otimes v_i \otimes v_i'] + O_p(1),$$
$$\sum_i \hat{v}_i \hat{v}_i' \otimes \hat{v}_i \hat{v}_i' = \sum_{ij} (M)_{ij}^4 \mathbb{E}[v_i v_i' \otimes v_i v_i']$$
$$+ \left[\sum_i M_{ii}^2 - \sum_{ij} M_{ij}^4 \right] ((I_4 + K_{2,2}) \Omega \otimes \Omega + \operatorname{vec}(\Omega) \operatorname{vec}(\Omega)') + O_p(1).$$

Part (ii) then follows. □

Proof of Proposition 6. It follows from Lemma 3 and the fact that T and S are uncorrelated that

$$\sqrt{n} \operatorname{vec}(T - (k_n/n)S - \Omega) \Rightarrow \mathcal{N}_4(0, 2N_2 (\tau \Omega \otimes \Omega + \Xi \otimes \Omega + \Omega \otimes \Xi))$$

Since $\hat{\beta}_{\text{UMD}} = g(\text{vec}(T - (k_n/n)S))$, where $g(T) = T_3/T_4$, it follows by 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

$$\begin{aligned} V_{\text{UMD}} &= \frac{2}{\Xi_{22}^2}(e_2 \otimes b)' N_2 \left(\tau\Omega \otimes \Omega + \Xi \otimes \Omega + \Omega \otimes \Xi\right)(e_2 \otimes b) \\ &= \frac{1}{\Xi_{22}^2}(e_2 \otimes b + b \otimes e_2)' \left(\tau\Omega \otimes \Omega + \Xi \otimes \Omega + \Omega \otimes \Xi\right)(e_2 \otimes b). \end{aligned}$$

Expanding this expression using the identity $|\Omega| = b'\Omega b \Omega_{22} - (b'\Omega e_2)^2$ and the identity (28b) yields the result. □

Proof of Lemma 2. We have:

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

where the first line follows from the identity $m_{\min} = Q_{\mathcal{S}}(\hat{\beta}_{\text{LIML}}, S)$, the second line follows by algebra, the third line follows from Equation (29), and the fourth line follows from $\hat{\beta}_{\text{LIML}} - \beta = O_p(n^{-1/2})$. Using Corollary 1, consistency of $\hat{\beta}_{\text{LIML}}$, the continuous mapping theorem and Equation (30a), we obtain:

$$\sqrt{n}(\hat{b}_{\text{LIML}} \otimes \hat{b}_{\text{LIML}})' \,\text{vec}\left(T - \alpha_k S - \lambda_n aa'/(a'\Omega^{-1}a)\right) \Rightarrow \mathcal{N}\left(0, 2\tau(b'\Omega b)^2\right). \tag{49}$$

Combining these results, we get:

$$\sqrt{n}(m_{\min} - \alpha_k) \Rightarrow \mathcal{N}\left(0, 2\tau\right).$$

The results for $\hat{J}_{\text{S}}$ and $\hat{J}_{\text{AR}}$ follow by the Delta method. To prove the remainder of the Lemma, I use the approximation from Peiser (1943) (see also (Anatolyev and Gospodinov, 2011)) that as $k \to \infty$,

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

Therefore,

$$\begin{aligned} \mathbb{P}\left(n\hat{J}_{\text{CD}} \geq q_{1-\text{ns}}^{\chi^2_{k_n - 1}}\right) &= \mathbb{P}\left(\sqrt{n}\hat{J}_{\text{CD}} \geq k_n/\sqrt{n} + \Phi^{-1}(1 - \text{ns})\sqrt{2\alpha_k} + o(1)\right) \\ &= \mathbb{P}\left(\sqrt{n}(\hat{J}_{\text{CD}} - \alpha_k)/\sqrt{2\tau} \geq \Phi^{-1}(1 - \text{ns})\sqrt{\alpha_k/\tau} + o(1)\right) \\ &= \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \Phi^{-1}(1 - \text{ns})\sqrt{\alpha_k/\tau} + o(1)\right) = 1 - \Phi\left(\Phi^{-1}(1 - \text{ns})\sqrt{\alpha_k/\tau}\right) + o(1) \\ &\to \Phi\left(\Phi^{-1}(\text{ns})\sqrt{\alpha_k/\tau}\right). \end{aligned}$$

Secondly,

$$\mathbb{P}\left(n\hat{J}_{\mathrm{s}} \geq q_{\mathrm{ns}}^{\chi^2_{k_n-1}}\right) = \mathbb{P}\left(\sqrt{n}\hat{J}_{\mathrm{s}} \geq k_n/\sqrt{n} + \Phi^{-1}(1-\mathrm{ns})\sqrt{2\alpha_k} + o(1)\right)$$
$$= \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \left(\frac{(1-\alpha_\ell)^3}{2\alpha_k(1-\alpha_k-\alpha_\ell)}\right)^{1/2}\left(-\sqrt{n}\frac{\alpha_k\alpha_\ell}{1-\alpha_\ell} + \Phi^{-1}(1-\mathrm{ns})\sqrt{2\alpha_k}\right) + o(1)\right).$$

Now, if $\alpha_k > 0$, then the right-hand side converges to $-\infty$, so that the rejection probability converges to one. If $\alpha_k = 0$, then

$$\mathbb{P}\left(n\hat{J}_{\mathrm{s}} \geq q_{\mathrm{ns}}^{\chi^2_{k_n-1}}\right) = \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \frac{\Phi^{-1}(1-\mathrm{ns})}{(1-\alpha_k)^{1/2}} + o(1)\right) \to \Phi\left(\frac{\Phi^{-1}(\mathrm{ns})}{(1-\alpha_k)^{1/2}}\right).$$

Thirdly,

$$\mathbb{P}\left(n\hat{J}_{\mathrm{AR}} \geq q_{\mathrm{ns}}^{\chi^2_{k_n-1}}\right) = \mathbb{P}\left(\sqrt{n}\hat{J}_{\mathrm{AR}} \geq k_n/\sqrt{n} + \Phi^{-1}(1-\mathrm{ns})\sqrt{2\alpha_k} + o(1)\right)$$
$$= \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \frac{1-\alpha_\ell}{\sqrt{2\tau}}\left(\sqrt{n}\left[\alpha_k - \log\left(\frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right)\right] + \Phi^{-1}(1-\mathrm{ns})\sqrt{2\alpha_k} + o(1)\right)\right).$$

Now, since $\alpha_k < -\log(1-\alpha_k)$,

$$\alpha_k - \log\left(\frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right) < \log\left(\frac{1}{1-\alpha_k}\right) - \log\left(\frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right) = \log\left(\frac{1-\alpha_k-\alpha_\ell}{(1-\alpha_k)(1-\alpha_\ell)}\right) < \log(1) = 0,$$

so that the right-hand side of the expression converges to $-\infty$, and the rejection probability converges to 1. □

References

Abramowitz, M. and Stegun, I. A. (eds.) (1965). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. New York: Dover. 30

Anatolyev, S. (2011). Instrumental variables estimation and inference in the presence of many exogenous regressors. 4, 7, 10, 20, 26, 35

— and Gospodinov, N. (2011). Specification testing in models with many instruments. *Econometric Theory*, **27** (2), 427–441. 26, 36

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, 20, 21

— 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. 2, 9, 25

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, 11

—, — and — (2008). Efficient two-sided nonsimilar invariant tests in IV regression with weak instruments. *Journal of Econometrics*, **146** (2), 241–254. 7

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. 4, 22

— 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. 22

— 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*. Oxford University Press. 5, 18

Bekker, P. A. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, **62** (3), 657–681. 2, 4, 7, 9, 10

— and Crudu, F. (2012). Symmetric Jackknife Instrumental Variable Estimation. 4, 7

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*, (forthcoming). 4

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

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, **18** (1), 5–46. 17

— (2007). Decision theory applied to an instrumental variables model. *Econometrica*, **75** (3), 609–652. 4, 7, 11, 12, 14

— and Imbens, G. W. (2004). Random Effects Estimators with Many Instrumental Variables. *Econometrica*, **72** (1), 295–306. 1, 2, 3, 8, 12, 13, 14, 27

— and Moreira, M. J. (2009). Decision Theory Applied to a Linear Panel Data Model. *Econometrica*, **77** (1), 107–133. 5, 12

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R. and Woutersen, T. (2010). Testing Overidentifying Restrictions with Many Instruments and Heteroscedasticity. 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, 6

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? *Quarterly Journal of Economics*, **126** (4), 1593–1660. 22

Chioda, L. and Jansson, M. (2009). Optimal Invariant Inference When the Number of Instruments Is Large. *Econometric Theory*, **25** (3), 793–805. 2, 3, 4, 11

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, **49** (1), 1–39. 5, 9, 12

Cragg, J. G. and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, **9** (2), 222–240. 4, 25

Donald, S. G. and Newey, W. K. (2001). Choosing the Number of Instruments. *Econometrica*, **69** (5), 1161–1191. 4, 21, 23

Eaton, M. L. (1989). *Group invariance applications in statistics*, *Regional conference series in Probability and Statistics*, vol. 1. Hayward, California: Institute of Mathematical Statistics. 12

Gautier, E. and Tsybakov, A. B. (2011). High-dimensional instrumental variables regression and confidence sets. 4

Goldberger, A. S. and Olkin, I. (1971). A minimum-distance interpretation of limited-information estimation. *Econometrica*, **39** (3), 635–639. 18

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

Hansen, C. B., Hausman, J. A. and Newey, W. K. (2008). Estimation With Many Instrumental Variables. *Journal of Business and Economic Statistics*, **26** (4), 398–422. 4, 10, 20

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, 6

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** (2), 467–475. 4, 22

Kolesár, M. (2012). Estimation in instrumental variables models with heterogeneous treatment effects. 7, 22

Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. and Imbens, G. W. (2011). Identification and Inference with Many Invalid Instruments. 7, 10, 22, 23

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. 5

— (2002). Orthogonal Parameters and Panel Data. *Review of Economic Studies*, **69** (3), 647–666. 5, 15

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: Some properties and applications. *The Annals of Statistics*, **7** (2), 381–394. 28

— and — (1980). The elimination matrix: Some lemmas and applications. *SIAM Journal on Algebraic and Discrete Methods*, **1** (4), 422–449. 28, 30

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. 3, 4, 5, 11, 12, 30

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. 4, 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*, Elsevier, pp. 2111–2245. 18, 20, 35

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16** (1), 1–32. 2, 5

Peiser, A. M. (1943). Asymptotic formulas for significance levels of certain distributions. *The Annals of Mathematical Statistics*, **14** (1), 56–62. 36

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*, Elsevier, vol. 1, pp. 449–516. 6

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*, Elsevier, pp. 881–935. 3, 7

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, **26** (3), 393–415. 4, 25

Sims, C. A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of Econometrics*, **95** (2), 443–462. 5

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65** (3), 557–586. 2

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press. 12

van Hasselt, M. (2010). Many Instruments Asymptotic Approximations Under Nonnormal Error Distributions. *Econometric Theory*, **26** (02), 633–645. 4, 20, 30

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. 21