Limited Information Maximum Likelihood: A Short Note 2024-11-06

jessekelighine.com
Jesse C. Chen 陳捷

Remark 1. Limited Information Maximum Likelihood (LIML) is a method first formally introduced by Anderson and Rubin (1949) to solve for Simultaneous Equations Model (SEM). In this note, we will focus on a specific application of SEM: Instrument Variable (IV) regression, and how LIML is used in this specific application.

Definition 1 (Instrument Variable Model)**.** Data is of the form $\{y_t, \mathbf{x}_t, \mathbf{z}_t, \mathbf{w}_t\}_{t=1}^T$ where y_t is a scalar outcome variable, $\mathbf{x}_t$ is a $N \times 1$ vector of potentially endogenous variable, $\mathbf{z}_t$ is a $K \times 1$ vector of IV, $\mathbf{w}_t$ is a $R \times 1$ vector of exogenous control variables. The model assumes the (matrix) form

$$\underset{(T\times 1)}{\mathbf{y}} = \underset{(T\times N)}{\mathbf{X}}\boldsymbol{\beta} + \underset{(T\times R)}{\mathbf{W}}\boldsymbol{\Psi} + \underset{(T\times 1)}{\mathbf{u}} \tag{1}$$

$$\underset{(T\times N)}{\mathbf{X}} = \underset{(T\times K)}{\mathbf{Z}}\boldsymbol{\Pi} + \underset{(T\times R)}{\mathbf{W}}\boldsymbol{\Phi} + \underset{(T\times N)}{\mathbf{v}} \tag{2}$$

$\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_T)^\top$, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_T)^\top$, $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_T)^\top$, $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_T)^\top$; $\mathbf{u} = (\mathbf{u}_1, ..., \mathbf{u}_T)^\top$ and $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_T)^\top$ are error terms; $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Phi}$ are parameters. Equation 1 is the so-called *structural equation* and Equation 2 is a linear projection, so-called *first stage*.

1 Simultaneous Equation Models

Definition 2 (Simultaneous Equations Model)**.** An SEM model is given in the following form

$$\underset{(T\times N)(N\times N)}{\mathbf{A}\ \ \boldsymbol{\Gamma}} = \underset{(T\times K)(K\times N)}{\mathbf{B}\ \ \boldsymbol{\Xi}} + \underset{(T\times N)}{\mathbf{U}} \tag{3}$$

where T is the number of observations, N is the number of potentially endogenous variables (N is also the number of model equations), $\mathbf{A}$ is the matrix of potentially endogenous variables, $\mathbf{B}$ is the matrix of exogenous variables, $\mathbf{U}$ is the matrix of errors, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Xi}$ are the parameters.

Remark 2. In addition, we require $\boldsymbol{\Gamma}$ to be invertible and write the reduced form as

$$\mathbf{A} = \mathbf{B}\boldsymbol{\Xi}\boldsymbol{\Gamma}^{-1} + \mathbf{U}\boldsymbol{\Gamma}^{-1} = \mathbf{B}\boldsymbol{\Omega} + \mathbf{V}.$$

Remark 3. For each endogenous variable/model equation $n = 1, ..., N$, we require that we can write

$$\underset{(T\times 1)}{\mathbf{A}_{\bullet n}} = \underset{(T\times(N-1))}{\mathbf{A}_{-\bullet n}}\ \underset{((N-1)\times 1)}{\boldsymbol{\gamma}_{\bullet n}} + \mathbf{B}\boldsymbol{\Xi}_{\bullet n} + \mathbf{U}_{\bullet n} \tag{4}$$

where

$$\boldsymbol{\gamma}_n = [-\boldsymbol{\Gamma}_{1n}, \cdots, -\boldsymbol{\Gamma}_{(n-1)n}, -\boldsymbol{\Gamma}_{(n+1)n}, \cdots, -\boldsymbol{\Gamma}_{nn}]^\top. \tag{5}$$

and $\mathbf{A}_{\bullet n}$, $\boldsymbol{\Xi}_{\bullet n}$, $\mathbf{U}_{\bullet n}$ denote the n-th column of the respective matrices and $\mathbf{A}_{-\bullet n}$ denotes the matrix $\mathbf{A}$ with n-th column removed. That is, we do not allow any endogenous variable to influence itself. This mean the diagonal of $\boldsymbol{\Gamma}$ are all ones. In our IV application, we will have this formulation automatically.

Example 1 (IV as SEM)**.** We can rewrite the formulation in Definition 1 as an SEM:

$$[\mathbf{y}, \mathbf{X}] \begin{bmatrix} 1 & \mathbf{0} \\ -\boldsymbol{\beta} & \mathbf{I} \end{bmatrix} = [\mathbf{W}, \mathbf{Z}] \begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{\Phi} \\ \mathbf{0} & \boldsymbol{\Pi} \end{bmatrix} + [\mathbf{u}, \mathbf{v}]. \tag{6}$$

Notice that we put both $\mathbf{y}$ and $\mathbf{X}$ on the left-hand side and $\operatorname{diag}(\boldsymbol{\Gamma}) = \mathbf{1}$.

Remark 4. In the IV model, our main goal is to estimate the parameter $\boldsymbol{\beta}$. This means that in formulation (6), we only really care about the estimation of one of the model equations, namely the first column of $\boldsymbol{\Gamma}$ where $\boldsymbol{\beta}$ is present.

Remark 5. There are two general approaches to estimating SEM: Generalized Method of Moments (GMM) and Maximum Likelihood (ML). Specifically, in general SEM, we talk about Three Stage Least Square (3SLS) and Full Information Maximum Likelihood (FIML); and when we are talking about IV estimator, we usually talk about Two Stage Least Square (2SLS) and LIML.

To use 3SLS or FIML, one has to specify the entire structure of SEM, hence the name FIML. This is in contrast to what we have done in Example 1, where we really only specified the "second-stage" model, and for the "first-stage" we simply used the reduced form. Hence, we have the name LIML, where only some equations in the SEM are structural.

1.1 Two Stage Least Square

Definition 3 (Two Stage Least Square)**.** Suppose we have the SEM as in (6) and the first equation is the structure equation of interest. We proceed as follows to obtain estimate of $\boldsymbol{\beta}$, as defined in (5):

- (Stage 1) Regress $\mathbf{X}$ on $\mathbf{W}$ and $\mathbf{Z}$ and obtain the predicted values $\hat{\mathbf{X}}$.
- (Stage 2) Regress $\mathbf{y}$ on $\hat{\mathbf{X}}$, $\mathbf{W}$, and $\mathbf{Z}$ to obtain estimates of $\boldsymbol{\beta}$.

The explicit formula for 2SLS estimator is given by

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\Pi}} \end{bmatrix} = ([\mathbf{X}, \mathbf{Z}]^\top \mathbf{P} [\mathbf{X}, \mathbf{Z}])^{-1} [\mathbf{X}, \mathbf{Z}]^\top \mathbf{P} \mathbf{y} \tag{7}$$

where

$$\mathbf{P} = [\mathbf{W}, \mathbf{Z}]([\mathbf{W}, \mathbf{Z}]^\top [\mathbf{W}, \mathbf{Z}])^{-1} [\mathbf{W}, \mathbf{Z}]^\top.$$

Remark 6. Remember that 2SLS is not robust to heteroskedastic errors. That is, 2SLS estimator for variance is only for the case that $\mathbf{E}(\mathbf{u}_t^2 \mid \mathbf{w}_t, \mathbf{z}_t)$ does not depend on t. For heteroskedastic errors, GMM estimator are needed.

1.2 Limited Information Maximum Likelihood

Remark 7 (Excerpt from Hayashi (2001), p.538)**.** The advantage of the FIML estimator is that it allows you to exploit all the information afforded by the complete system of simultaneous equations. This, however, is also a weakness because, as is true with any other system or joint estimation procedure, the estimator is not consistent if any part of the system is misspecified. If you are confident that the equation in question is correctly specified but not so sure about the rest of the system, you may well prefer to employ single-equation methods such as 2SLS. The rest of this section derives the ML estimator called the LIML estimator, which is the ML counterpart of 2SLS.

Remark 8. Thus, in the end, there is nothing special about LIML, it can be viewed as an FIML estimator with some of the structure replaced by reduced form parameters. However, there is a lot to gain in terms of the actual computation of LIML, since the form of $\mathbf{\Gamma}$ and $\mathbf{\Xi}$ is very particular. For FIML, we have to resort to numerical methods to obtain the estimator, whereas with LIML, there is a closed form solution.

Example 2 (IV as SEM Continued)**.** For each sample, we have the SEM as

$$(y_t, \mathbf{x}_t)\begin{bmatrix} 1 & \mathbf{0} \\ -\boldsymbol{\beta} & \mathbf{I} \end{bmatrix} = (\mathbf{w}_t, \mathbf{z}_t)\begin{bmatrix} \mathbf{\Psi} & \mathbf{\Phi} \\ \mathbf{0} & \mathbf{\Pi} \end{bmatrix} + (\mathbf{u}_t, \mathbf{v}_t). \tag{8}$$

where $(\mathbf{u}_t, \mathbf{v}_t) \mid (\mathbf{w}_t, \mathbf{z}_t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$. The log-likelihood function is given by

$$\begin{aligned} \ell(\mathbf{\Gamma}, \mathbf{\Xi}, \mathbf{\Sigma}) = & -\frac{TN}{2}\ln(2\pi) - \frac{T}{2}\ln|\mathbf{\Sigma}| \\ & -\frac{1}{2}\sum_{t=1}^{T}[(y_t, \mathbf{x}_t)\mathbf{\Gamma} + (\mathbf{w}_t, \mathbf{z}_t)\mathbf{\Xi}]\mathbf{\Sigma}^{-1}[(y_t, \mathbf{x}_t)\mathbf{\Gamma} + (\mathbf{w}_t, \mathbf{z}_t)\mathbf{\Xi}]^\top. \end{aligned} \tag{9}$$

Definition 4 (Limited Information Maximum Likelihood)**.** Define the following projection and annihilator matrices:

$$\mathbf{P}_\mathbf{W} = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top \text{ and } \mathbf{M}_\mathbf{W} = \mathbf{I} - \mathbf{P}_\mathbf{W}.$$

Also, define $\mathbf{M} = \mathbf{I} - \mathbf{P}$. The LIML estimator for $\boldsymbol{\beta}$ is

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{\Pi}} \end{bmatrix} = ([\mathbf{X}, \mathbf{Z}]^\top(\mathbf{I} - k\mathbf{M})[\mathbf{X}, \mathbf{Z}])^{-1}[\mathbf{X}, \mathbf{Z}]^\top(\mathbf{I} - k\mathbf{M})\mathbf{y}. \tag{10}$$

where k is the smallest eigenvalue of

$$\left([\mathbf{y}, \mathbf{X}]^\top\mathbf{M}[\mathbf{y}, \mathbf{X}]\right)^{-1/2}[\mathbf{y}, \mathbf{X}]^\top\mathbf{M}_\mathbf{W}[\mathbf{y}, \mathbf{X}]\left([\mathbf{y}, \mathbf{X}]^\top\mathbf{M}[\mathbf{y}, \mathbf{X}]\right)^{-1/2}.$$

Remark 9. Some tedious algebra is needed to derive (10), even Hayashi (2001) did not bother to write down the derivation. Here I will present the outline of the algebra needed.

Sketch of Proof for Definition 4. (Davidson and MacKinnon (1993), pp.644–649) We derive the form (10) with the following steps:

1. We rewrite the log-likelihood function (9) as the concentrated log-likelihood function, i.e., we replace the $\mathbf{\Sigma}$ in the model with

$$\mathbf{\Sigma}(\mathbf{\Xi}, \mathbf{\Gamma}) = \frac{1}{T}(\mathbf{A}\mathbf{\Gamma} - \mathbf{B}\mathbf{\Xi})^\top(\mathbf{A}\mathbf{\Gamma} - \mathbf{B}\mathbf{\Xi}).$$

   This concentration is obtain by first order condition. Hence, we have the concentrated log-likelihood function as

$$\begin{aligned} \ell^\mathsf{C}(\mathbf{\Xi}, \mathbf{\Gamma}) &= -\frac{TN}{2}(\ln(2\pi) + 1) - \frac{T}{2}\ln\left|\frac{1}{T}(\mathbf{A}\mathbf{\Gamma} - \mathbf{B}\mathbf{\Xi})^\top(\mathbf{A}\mathbf{\Gamma} - \mathbf{B}\mathbf{\Xi})\right| \\ &= -\frac{TN}{2}(\ln(2\pi) + 1) - \frac{T}{2}\ln\left|\frac{1}{T}(\mathbf{A} - \mathbf{B}\mathbf{\Xi}\mathbf{\Gamma}^{-1})^\top(\mathbf{A} - \mathbf{B}\mathbf{\Xi}\mathbf{\Gamma}^{-1})\right| \end{aligned}$$

   where we used the fact that $|\mathbf{\Gamma}| = 1$. Thus, it is clear the objective now is to minimize the determinant.

2. Notice that we have the expression

$$\mathbf{\Xi}\mathbf{\Gamma}^{-1} = \begin{bmatrix} \mathbf{\Psi} & \mathbf{\Phi} \\ \mathbf{0} & \mathbf{\Pi} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \beta & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi} + \mathbf{\Phi}\beta & \mathbf{\Phi} \\ \mathbf{\Pi}\beta & \mathbf{\Pi} \end{bmatrix}.$$

Note that the columns corresponding to the first stage regression does not have any restrictions. Hence, to minimize the determinant, we should simply use the Orfdinary Least Squares (OLS) estimators for that column. We can simply consider the model with those variables partialled out:

$$|(\mathbf{A}\mathbf{\Gamma} - \mathbf{B}\mathbf{\Xi})^\top \mathbf{M}_\mathbf{W}(\mathbf{A}\mathbf{\Gamma} - \mathbf{B}\mathbf{\Xi})|$$

Another way of understanding this is to invoke the Frisch-Waugh-Lovell theorem before writing the likelihood function.

3. Through some tedious algebra, the determinant can be reduced to

$$\begin{gathered}(1, -\boldsymbol{\beta}^\top)\mathbf{A}^\top\mathbf{M}_\mathbf{W}\mathbf{A}(1, -\boldsymbol{\beta}^\top)^\top \cdot \\ |(\mathbf{M}_\mathbf{W}\mathbf{X} - \mathbf{M}_\mathbf{W}\mathbf{Z}\mathbf{\Pi})^\top \mathbf{M}_{\mathbf{M}_\mathbf{W}\mathbf{A}(1,-\boldsymbol{\beta}^\top)^\top}(\mathbf{M}_\mathbf{W}\mathbf{X} - \mathbf{M}_\mathbf{W}\mathbf{Z}\mathbf{\Pi})|\end{gathered} \tag{11}$$

where $\mathbf{M}_{\mathbf{M}_\mathbf{W}\mathbf{A}(1,-\boldsymbol{\beta}^\top)^\top}$ denotes, as usual, the projection on to the orthogonal space of $\mathbf{M}_\mathbf{W}\mathbf{A}(1, -\boldsymbol{\beta}^\top)^\top$. Since $\mathbf{\Pi}$ only appears in the second term and it is in a quadratic form, we know the OLS estimator should be chosen to minimize the determinant. Hence, the second term, through some more tedious algebra, reduces to

$$|\mathbf{X}^\top\mathbf{M}_\mathbf{W}\mathbf{M}_{\mathbf{M}_\mathbf{W}\mathbf{A}(1,-\boldsymbol{\beta}^\top)^\top, \mathbf{M}_\mathbf{W}\mathbf{Z}}\mathbf{M}_\mathbf{W}\mathbf{X}| = \frac{|\mathbf{A}^\top\mathbf{M}\mathbf{A}|}{(1, -\boldsymbol{\beta}^\top)^\top\mathbf{A}^\top\mathbf{M}\mathbf{A}(1, -\boldsymbol{\beta}^\top)^\top}.$$

4. Plug the result back in (11) and we have that

$$\underbrace{\frac{(1, -\boldsymbol{\beta}^\top)\mathbf{A}^\top\mathbf{M}_\mathbf{W}\mathbf{A}(1, -\boldsymbol{\beta}^\top)^\top}{(1, -\boldsymbol{\beta}^\top)^\top\mathbf{A}^\top\mathbf{M}\mathbf{A}(1, -\boldsymbol{\beta}^\top)^\top}}_{\kappa}|\mathbf{A}^\top\mathbf{M}\mathbf{A}|.$$

Notice that $\boldsymbol{\beta}$ only appears in κ and $|\mathbf{A}^\top\mathbf{M}\mathbf{A}|$ is determined. Therefore, minimizing the determinant boils down to minimizing κ. Note tat κ is something like a Rayleigh quotient, thus, the minimum value achievable is the smallest eigenvalue of the matrix

$$\Big([\mathbf{y}, \mathbf{X}]^\top\mathbf{M}[\mathbf{y}, \mathbf{X}]\Big)^{-1/2}[\mathbf{y}, \mathbf{X}]^\top\mathbf{M}_\mathbf{W}[\mathbf{y}, \mathbf{X}]\Big([\mathbf{y}, \mathbf{X}]^\top\mathbf{M}[\mathbf{y}, \mathbf{X}]\Big)^{-1/2}.$$

Notice that the eigenvalue is at least one.

5. Lastly, to obtain the estimator for $\boldsymbol{\beta}$, we simply take the first order condition of κ with respect to $(1, -\boldsymbol{\beta}^\top)$ and plug in the minimum eigenvalue:

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}^\top(\mathbf{M}_\mathbf{W} - k\mathbf{M})\mathbf{X}]^{-1}\mathbf{X}^\top(\mathbf{M}_\mathbf{W} - k\mathbf{M})\mathbf{y}.$$

Written together with the estimator for $\mathbf{\Pi}$ and we have

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{\Pi}} \end{bmatrix} = ([\mathbf{X}, \mathbf{Z}]^\top(\mathbf{I} - k\mathbf{M})[\mathbf{X}, \mathbf{Z}])^{-1}[\mathbf{X}, \mathbf{Z}]^\top(\mathbf{I} - k\mathbf{M})\mathbf{y}.$$

#

Remark 10 (Excerpt from Hayashi (2001), p.541)**.** If we do not necessarily require k to be as just defined, the estimator (10) is called a *k-class estimator*. The LIML estimator is thus a k-class estimator with k defined above. Inspection of the 2SLS formula in terms of data matrices (7) immediately shows that the 2SLS estimator is a k-class estimator with $k = 1$, and the OLS estimator is a k-class estimator with $k = 0$. It follows that LIML and 2SLS are numerically the same when the equation is just identified (so that $k = 1$).

Remark 11. Since LIML is a special case of LIML, the asymptotic properties follows from standard assumptions on m-estimators.

Remark 12 (Excerpt from Davidson and MacKinnon (1993), p.541)**.** It can be shown that k-class estimators are consistent whenever k tends to 1 asymptotically at rate $o(n^{-1/2})$.

Remark 13 (Excerpt from Hayashi (2001), p.542)**.** Since LIML and 2SLS share the same asymptotic distribution, you cannot prefer one over the other on asymptotic grounds. For any finite sample, LIML has the invariance property, while 2SLS is not invariant. Furthermore, the conclusion one could draw from the large literature on finite-sample properties (see, e.g., Judge et al., 1985, Section 15.4) is that LIML should be preferred to 2SLS.

Remark 14. Under weak IV asymptotics, developed by Staiger and Stock (1997), the LIML estimator is not consistent, and is different from 2SLS asymptotically.

Acronyms

2SLS	Two Stage Least Square. 2, 5
3SLS	Three Stage Least Square. 2
FIML	Full Information Maximum Likelihood. 2, 3
GMM	Generalized Method of Moments. 2
IV	Instrument Variable. 1–3, 5
LIML	Limited Information Maximum Likelihood. 1–3, 5
ML	Maximum Likelihood. 2
OLS	Orfdinary Least Squares. 4, 5
SEM	Simultaneous Equations Model. 1–3

References

Anderson, T. W. and Herman Rubin (1949). "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations". In: *The Annals of Mathematical Statistics* 20.1, pp. 46–63. DOI: `10.1214/aoms/1177730090`. URL: `https://doi.org/10.1214/aoms/1177730090`.

Davidson, Russell and James G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press. ISBN: 0-19-506011-3.

Hayashi, Fumio (2001). *Econometrics*. Princeton University Press. ISBN: 9780691010182.

Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regression with Weak Instruments". In: *Econometrica* 65.3, pp. 557–586. ISSN: 00129682, 14680262. URL: `http://www.jstor.org/stable/2171753` (visited on 08/13/2024).