The Information Matrix Equality: proof, misspecification, and the quasi-maximum likelihood estimator

May 2014

Alecos Papadopoulos
Athens University of Economics and Business, Greece
Department of Economics
e-mail: papadopalex@aueb.gr

JEL CODES: C10, C13.

The Information Matrix Equality (IME), which is usually used in the context of maximum likelihood estimation, says that the expected value of the Hessian of the log-likelihood function equals the negative of the expected value of the outer product of its gradient. This simplifies the expression of the asymptotic variance-covariance matrix of the maximum likelihood estimator (MLE), and moreover permits us to obtain an estimate of it without the need to invert the Hessian, which may be problematic or, even with today's computing power, resource-intensive.

In this note we prove the IME, and we also show why misspecification invalidates it. Finally we discuss under which conditions we may nevertheless apply it, in the context of "Quasi"-maximum likelihood estimation.

The Information Matrix Equality

(The exposition uses Amemiya (1985), ch 1.3.2, but in a way that focuses on IME, while Amemiya does not even name it, and it uses it in the context of presenting the Cramer-Rao lower bound. But it is the one exposition I know that presents IME in its full generality).

Let $\mathbf{z}$ be an *m*-component column vector of random variables, *not necessarily independent and/or identically distributed*, and let $f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})$ be their joint density, where $\boldsymbol{\theta}$ is a *k*-component column vector of parameters in some parameter space $\boldsymbol{\Theta}$. Assume that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})\, d\mathbf{z} \quad = \quad \int \frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})\, d\mathbf{z} \ .$$

i.e. that we can interchange the order of differentiation and integration. A set of sufficient conditions for this are:

(i) The function $\frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})$ is continuous in $\mathbf{z}$ and in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is an open set

(ii) The integral $\int f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})\, d\mathbf{z}$ exists

(iii) $\int \left| \frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta}) \right| d\mathbf{z} < M < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

(Note that the integral is to be understood as a multiple integral, while the limits of integration are always $\mp\infty$).

Then the Information Matrix Equality holds i.e.

$$E \frac{\partial^2 \ln f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \quad = \quad - \ E \frac{\partial \ln f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$$

Proof. To ease notation we will write $f_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta}) \equiv f$.

Then, by standard rules of differentiation, we have

$$E\frac{\partial^2 \ln f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial \ln f}{\partial\boldsymbol{\theta}'}\right)\right] = E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{f}\frac{\partial f}{\partial\boldsymbol{\theta}'}\right)\right]$$

$$= E\left[-\frac{1}{f^2}\frac{\partial f}{\partial\boldsymbol{\theta}}\frac{\partial f}{\partial\boldsymbol{\theta}'} + \frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = -E\left[\left(\frac{1}{f}\frac{\partial f}{\partial\boldsymbol{\theta}}\right)\left(\frac{1}{f}\frac{\partial f}{\partial\boldsymbol{\theta}'}\right)\right] + E\left[\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$$

$$\Rightarrow E\frac{\partial^2 \ln f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = -E\left[\frac{\partial \ln f}{\partial\boldsymbol{\theta}}\frac{\partial \ln f}{\partial\boldsymbol{\theta}'}\right] + E\left[\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \qquad [1]$$

To obtain IME we need to show that the second term is zero. We have, by the so-called "Law of the Unconscious Statistician",

$$E\left[\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \int f\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}d\mathbf{z} = \int\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}d\mathbf{z}$$

Interchanging the order of integration and differentiation we have

$$\int\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}d\mathbf{z} = \int\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial f}{\partial\boldsymbol{\theta}'}\right)d\mathbf{z} = \frac{\partial}{\partial\boldsymbol{\theta}}\left(\int\frac{\partial f}{\partial\boldsymbol{\theta}'}d\mathbf{z}\right).$$

Interchanging it *again* we have

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\int\frac{\partial f}{\partial\boldsymbol{\theta}'}d\mathbf{z}\right) = \frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}'}\int f d\mathbf{z}\right)$$

But f is a joint density function in $\mathbf{z}$ so $\int f d\mathbf{z} = 1$. Therefore

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}'}\int f d\mathbf{z}\right) = \frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}'}(1)\right) = 0 \Rightarrow E\left[\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = 0 \text{ and so}$$

$$E\frac{\partial^2 \ln f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -E\left[\frac{\partial \ln f}{\partial \boldsymbol{\theta}}\frac{\partial \ln f}{\partial \boldsymbol{\theta}'}\right] \qquad [2]$$

which is the Information Matrix Equality. □

We see that the result is pretty general, subject to some regularity conditions on the joint probability density function. Usually these are satisfied, but not always (can you think of cases where condition (i) does not hold?). But apart from that it has wide applicability -in particular it does not require i.i.d. random variables in order to hold.

But more importantly, *it is not linked to any specific estimation procedure*. It is a theoretical result pertaining to the actual joint distribution that the random variables in question follow. In practice, we are usually faced with the task of estimation under uncertainty about what is the true distribution: this brings in the picture the *specified* joint density function also, alongside the true one.

Misspecification and the IME

Assume that our random variables have joint density f as before, but we specify that they have joint density g instead (the different letter just to avoid confusion: we do not identify what form the misspecification takes).

Then as before we will have

$$E_{\mathbf{z}}\frac{\partial^2 \ln g}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -E_{\mathbf{z}}\left[\frac{\partial \ln g}{\partial \boldsymbol{\theta}}\frac{\partial \ln g}{\partial \boldsymbol{\theta}'}\right] + E_{\mathbf{z}}\left[\frac{1}{g}\frac{\partial^2 g}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \qquad [3]$$

where we have used the subscript in the expected value operator in order to keep in sight the fact that the expected value is taken with respect to the *true* joint distribution of $\mathbf{z}$ (what we assume for our random variables does not affect their actual distribution! To take

the expected value with respect to the postulated joint distribution g would be equivalent to assume that g is after all the true joint density).

Then we have

$$E_{\mathbf{z}}\left[\frac{1}{g}\frac{\partial^2 g}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \int f\frac{1}{g}\frac{\partial^2 g}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}d\mathbf{z} \qquad [4]$$

Previously, we moved on with the proof that IME holds because we had in the integrand $f\frac{1}{f}=1$. Here, due to the misspecification, we have $f\frac{1}{g}$, which does not simplify, and so we cannot take the derivative operator outside the integral (since it will then differentiate also f/g). So in general, under misspecification we will have $E_{\mathbf{z}}\left[\frac{1}{g}\frac{\partial^2 g}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\neq 0$ and IME will not hold. But this does not *totally* exclude the possibility that this expected value is after all zero.

The Quasi-maximum likelihood estimator

The QMLE (or "pseudo-MLE") can be seen as a generalization of the "standard" maximum likelihood estimation approach. The latter assumes a correct stochastic specification of the joint density of the random variables involved. With observational data, this should be the exception rather than the rule. So in the QMLE framework, we accept that most probably our specification only approximates the true distributions involved. This has an efficiency cost, but also it renders the Information Matrix Equality invalid as we just have seen -or does it?

Let a collection of n random variables that in reality have a joint density $f_{\mathbf{z}}(\mathbf{z})$, but for which we specify that their joint density is $g_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})$. Note carefully that $\boldsymbol{\theta}$ does *not*

appear in the true joint density function. Namely, we accept that the parametrization we postulate may have nothing to do with the true joint density. But, by maximizing the log-likelihood based on $f_{\mathbf{z}}(\mathbf{z})$, we are essentially estimating the minimizer $\boldsymbol{\theta}_{\arg\min}$ of the Kullback-Leibler Information Criterion (KLIC) for the "divergence/distance" between $g_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta})$ and $f_{\mathbf{z}}(\mathbf{z})$-see Kuan (2004) for more details. Note that the estimator $\hat{\boldsymbol{\theta}}_{QMLE}$ is a consistent estimator for $\boldsymbol{\theta}_{\arg\min}$ irrespective of the misspecification. But in general, here too the previous analysis applies and the IME will not hold, except if:

There exists $\boldsymbol{\theta}_0$ such that $g_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta}_0) = f_{\mathbf{z}}(\mathbf{z})\ \forall \mathbf{z}$, in which case

(i) $\boldsymbol{\theta}_0$ minimizes the KLIC, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\arg\min}$ (and so $\hat{\boldsymbol{\theta}}_{QMLE} \xrightarrow{p} \boldsymbol{\theta}_0$)

(ii) The Information Matrix Equality holds in relation to the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{QMLE}$

Remember that we do not consider the parameter vector $\boldsymbol{\theta}$ as part of the true distribution, which is liberating -we do not have to worry about the "true parameter values", because these parameters may not even exist in the true distribution. We only care whether there exists a $\boldsymbol{\theta}_0$ such that $g_{\mathbf{z}}(\mathbf{z},\boldsymbol{\theta}_0) = f_{\mathbf{z}}(\mathbf{z})\ \forall \mathbf{z}$ ... but is this feasible?

Well, in practice it appears not. What may happen is that we will specify a generalized version of the true distribution including a parameter vector, and so the true distribution will obtain for specific values of these parameters (even zero): for example the Exponential and Chi-square distributions are special cases of the Gamma distribution (family), the continuous Uniform distribution is a special case of the Beta distribution family, and in general all sorts of interrelations exist between many "named" distributions.

But methodologically, this "does not count" as misspecification. When the specified distribution *nests* the true distribution, most scholars will say that we have a "correctly

specified model", since the exact values of the parameters are naturally left for the data to point the finger at. But this is the only case (bar perhaps exotic examples), where the Information Matrix Equality holds under "some kind" of misspecification.

True Misspecification Example: Assume that the random variables are independent normal with identical mean but different variances, i.e. we have heteroskedasticity. Let's say we are interested in estimating the common mean only. We go on and specify an i.i.d. joint density, and execute maximum likelihood estimation. Our estimator is Quasi-MLE. Moreover, there is no parameter vector that will make the specified joint density equal to the true density. So the information matrix inequality does not hold, and we should not use the simplified expression for the asymptotic variance-covariance matrix of the estimator, because it will be wrong.

Conclusions

The Information Matrix Equality is an impressive and fairly general theoretical result. But it is a property of *one* joint density. In the context of statistical estimation, we have *two* joint densities, the true one that characterizes the data, and the one that we specify in order to execute statistical estimation and inference. In such a case, usually some form of misspecification will exist and the IME will not hold, especially with observational data. Only if we can validate the choice of a distributional family that nests the true joint distribution, so that we can let the data single out the specific variant of this family that characterizes them, it would be wise to use IME. In all other cases, it appears better to use the full formula for the asymptotic variance-covariance matrix of the maximum likelihood estimator, since chances are it will be a quasi-MLE.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard university press.

Kuan M.-C. (2004) Statistics: Concepts and Methods. 2nd edition (in Chinese) Taipei: Hua-Tai. Chapter 9 (in English) "The Quasi-Maximum Likelihood Method: Theory" downloadable from author's official webpage at http://homepage.ntu.edu.tw/~ckuan/pdf/et01/ch9.pdf