Econometrica, Vol. 41, No. 6 (November, 1973)

THE THEORY OF PARAMETRIC IDENTIFICATION

BY ROGER BOWDEN

This paper sets out a general criterion for the identifiability of a statistical system, based on Kullback's information integral. It is shown that the general identification problem is equivalent to a maximisation problem, or where parameter restrictions are present, a problem in nonlinear programming. The relationship of this criterion to that based on the information matrix of the underlying distribution is also exhibited.

THE COLLECTION OF results on the identification problem in econometrics is by now assuming the proportions of an imposing edifice. It is, however, a little surprising to note that this structure has been growing upwards and, more recently, outwards, without a corresponding strengthening in the foundations. It is true that in the case of work on linear simultaneous equation systems (and this, with the work of Koopmans and Reiersol [3] and Chapters 1–4 of Fisher [1] in particular, has almost assumed the status of a "classical" line of development), these results are founded on a pretty secure rock; to wit, the identifiability of the reduced form in the absence of any singularities in the data matrices. Nevertheless, with the development of other wings on our edifice, it seems desirable to look to more basic things.

The recent paper by Thomas Rothenberg [5] provided a welcome attack on this subject. The identifiability of a parametric system is approached via the nonsingularity of R. A. Fisher's "information matrix" evaluated at the true value of the parameter. The present note generalizes this approach by providing a simple criterion for identifiability, which not only affords an approach to global identification, but also makes no assumptions about the regularity of the underlying distribution. Rothenberg's basic theorem emerges as a simple corollary to this result. The approach has a natural relationship with estimation theory and also provides a straightforward method for delineating the subspace of observationally equivalent parameters in the case of underidentification.

1. DEFINITIONS

We assume the usual distinction between a *model* and the set of *structures* which it comprises. Specifically, we shall mean by a model a probability distribution $F(x, \cdot)$ of known form, and by a structure that distribution for a given parameter θ.

We assume that the set of possible parameters θ is a subset Ω of R^n, and that for every θ in Ω, $F(x;\theta)$ is a proper distribution function.

DEFINITION 1: The parameter point θ_0 is *identifiable* if there exists no other $\theta \in \Omega$ such that with probability 1, $dF(x;\theta) = dF(x;\theta_0)$, where the measure is taken with respect to θ_0.

What this says is that if $F(x; \theta_0)$ is the true distribution, then no other θ should generate a distribution indistinguishable from $F(x; \theta_0)$ on the basis of sample observations. In other words, we assume from the start the existence of a true, underlying θ_0, and ask whether this can be distinguished on the basis of sample information. Such a definition is in accordance with conventional usage (see, e.g., Koopmans and Reiersol [**3**, p. 169]).

A particular case is the distribution F when it is absolutely continuous for all θ. Then $dF(x; \theta) = f(x; \theta)\, dx$, where $f = F'$ is the corresponding density function. In this case the above criterion reduces to the equality of $f(x; \theta)$ and $f(x; \theta_0)$, for all x. For the sake of convenience, Rothenberg uses this latter definition (but see the footnote on page 578 of his paper [**5**]). However, we shall stick to the more inclusive definition.

Notice that we have not invoked a prior definition of observational equivalence, but it is, of course, there all the same:

DEFINITION 2: θ_0 and θ_1 are *observationally equivalent* if, with probability 1, $dF(x; \theta_0) = dF(x; \theta_1)$, where the measure is taken with respect to θ_0. If θ_0 and θ_1 are not observationally equivalent, we shall find it convenient to speak of the distribution functions as "different."

Finally, note also that the definitions do not presuppose that *all* information on θ_0 is sample information; some (but not all!) can take the form of a priori information.

2. A GENERAL IDENTIFIABILITY CRITERION

Suppose that the distribution functions $F(x; \theta)$ and $F(x; \theta_0)$ are absolutely continuous relative to one another. Then we define a function:

$$(1) \qquad H(\theta; \theta_0) = E \log \left[\frac{dF(x; \theta)}{dF(x; \theta_0)} \right], \qquad \text{given } \theta = \theta_0,$$

$$= \int_{-\infty}^{\infty} \log \left[\frac{dF(x; \theta)}{dF(x; \theta_0)} \right] dF(x; \theta_0).$$

Where the distribution functions are continuous,

$$H(\theta; \theta_0) = \int_{-\infty}^{\infty} \log \left[\frac{f(x; \theta)}{f(x; \theta_0)} \right] f(x; \theta_0)\, dx.$$

The function H has been defined by Kullback [**4**] as the information integral for discriminating $F(x; \theta_0)$ against $F(x; \theta_1)$ per observation from $F(x; \theta_0)$ (see also Wilks [**6**, Ch. 13]). The following result is by no means original (a version is proved, for instance, as Lemma 1 in Chapter 8 of Fraser [**2**]).

THEOREM: *If the distribution $F(x; \theta)$ is different from $F(x; \theta_0)$, and if $H(\theta_0; \theta_0)$ is finite, then $H(\theta; \theta_0) < 0$. Otherwise, $H(\theta; \theta_0) = 0$.*

PROOF: The proof uses the strict concavity of $\log(\cdot)$. Suppose that $F(x;\theta)$ is different from $F(x;\theta_0)$; then

$$\begin{aligned} H(\theta;\theta_0) &= E\log\left[\frac{dF(x;\theta)}{dF(x;\theta_0)}\right] \\ &< \log E\left[\frac{dF(x;\theta)}{dF(x;\theta_0)}\right] \qquad \text{(Jensen's inequality)} \\ &= \log\int_{-\infty}^{\infty}\left[\frac{dF(x;\theta)}{dF(x;\theta_0)}\right] dF(x;\theta_0) \\ &= \log 1 \\ &= 0. \end{aligned}$$

On the other hand, if $dF(x;\theta) = dF(x;\theta_0)$ with probability 1, then obviously

$$\int_{-\infty}^{\infty}\log\left[\frac{dF(x;\theta)}{dF(x;\theta_0)}\right] dF(x;\theta_0) = 0.$$

This completes the proof.

In view of Definitions 1 and 2, we can now establish the immediate corollary:

COROLLARY: *Suppose that for all $\theta \in \Omega$, the distribution functions $F(x;\theta)$ and $F(x;\theta_0)$ are absolutely continuous relative to one another. Then the parameter θ_0 is globally identified if and only if the equation*

$$H(\theta;\theta_0) = 0$$

has as solution in Ω only $\theta = \theta_0$.

It is locally identified if and only if θ_0 is the only solution in some open neighbourhood of θ_0.

Notice also that identifiability is closely connected with the maximum of $H(\theta;\theta_0)$. If this maximum is global and attained only at $\theta = \theta_0$, then θ_0 is globally identified. If the maximum is unique in a neighbourhood of $\theta = \theta_0$, then local identification holds.

Thus, for instance, a sufficient condition for θ_0 to be globally identified would be that $H(\theta;\theta_0)$ be strictly concave in Ω, and that Ω itself be convex.

3. IDENTIFIABILITY AND THE INFORMATION MATRIX

So far we have made no assumptions about the continuity of F as a function of θ. If we are willing to restrict F, we can now see how Rothenberg's Theorem 1 emerges as a simple corollary to the present approach.

Assume that F is absolutely continuous for all θ, and that, in a neighbourhood of θ_0, $f(x;\theta)$ and $\log f(x;\theta)$ are twice differentiable in θ, with derivatives continuous in x.

Under these conditions, one can easily show that $H(\theta;\theta_0)$ is continuous in the requisite neighbourhood. A sufficient condition for the existence of a unique local maximum is then that

$$H'(\theta_0;\theta_0) = 0$$

and

$$H''(\theta_0;\theta_0) < 0,$$

where the prime denotes $\partial/\partial\theta$. Initially we suppose θ a scalar.

We now investigate the form of these derivatives. Since $\int_{-\infty}^{\infty} f(x;\theta)\,dx = 1$ for any θ, it follows that

$$\int_{-\infty}^{\infty} f'(x;\theta_0)\,dx = 0 \tag{2}$$

and

$$\int_{-\infty}^{\infty} f''(x;\theta_0)\,dx = 0. \tag{3}$$

Differentiating (1), we obtain

$$\begin{aligned} H'(\theta;\theta_0) &= \int_{-\infty}^{\infty} \frac{f(x;\theta_0)}{f(x;\theta)}\cdot\frac{f'(x;\theta)}{f(x;\theta_0)}\cdot f(x;\theta_0)\,dx \\ &= E\,\frac{\partial}{\partial\theta}\log f(x;\theta), \qquad \text{given } \theta = \theta_0. \end{aligned} \tag{4}$$

We see immediately from (2) and (4) that

$$H'(\theta_0;\theta_0) = 0. \tag{5}$$

Also,

$$\begin{aligned} H''(\theta;\theta_0) &= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\left[\frac{f'(x;\theta)}{f(x;\theta)}\right] f(x;\theta_0)\,dx \\ &= \int_{-\infty}^{\infty} \left[\frac{f''(x;\theta)}{f(x;\theta)} - \left(\frac{f'(x;\theta)}{f(x;\theta)}\right)^2\right] f(x;\theta_0)\,dx \\ &= \int_{-\infty}^{\infty} \left[\frac{f''(x;\theta)}{f(x;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] f(x;\theta_0)\,dx. \end{aligned} \tag{6}$$

Applying (3) to the first of the two expectations on the right-hand side, we obtain

$$H''(\theta_0;\theta_0) = -E\left[\frac{\partial}{\partial\theta}\log f(x;\theta_0)\right]^2. \tag{7}$$

Extension to the case where θ is a vector of parameters is straightforward. H'' now becomes the classical information matrix, and the extension of (7) is

$$H''(\theta_0;\theta_0) = -\left[E\frac{\partial}{\partial\theta_i}\log f(x;\theta_0)\frac{\partial}{\partial\theta_j}\log f(x;\theta_0)\right]. \tag{7'}$$

We can now elucidate the relationship of the information matrix to identifiability.

(i) If $H''(\theta_0; \theta_0)$ is negative definite, it follows from standard maximization theory that $H(\theta; \theta_0)$ has a unique maximum at θ_0. That is, if the information matrix at θ_0 has full rank, then θ_0 is locally identified.

(ii) Suppose that in an open neighbourhood of θ_0, $H''(\theta; \theta_0)$ is regular (i.e., does not change rank). Then if θ_0 is locally identified, $H''(\theta_0; \theta_0)$ has full rank (i.e., is negative definite).

The proof of the second statement follows from the result (proved in the present case by expanding $H(\theta; \theta_0)$ in Taylor series about θ_0) whereby under the above regularity assumption, local uniqueness of θ_0 as a maximum of $H(\theta; \theta_0)$ implies $H''(\theta_0; \theta_0)$ is negative definite.[1]

Thus, the sufficiency of the criterion based on the information matrix requires no regularity assumption; but at least as far as present knowledge is concerned, necessity does seem to require such a restriction.

4. REMARKS

One can extend the analysis to allow for a priori restrictions connecting the elements of θ, or by allowing restrictions on the parameter space imposed by inequalities on elements of θ. The identification problem is then one of investigating the local or global maxima of $H(\theta; \theta_0)$ subject to those constraints (in other words, a problem in nonlinear programming).

As such, the approach is a very general one. Thus it provides a criterion which encompasses global as well as local criteria. In its general form, no assumption is made about the regularity of the distribution with respect to θ, a point which might turn out to be of some importance in practical situations.

Further, such a concept may prove to be of value in that it offers a quantitative measure of the "sharpness" of identification. Thus if $H(\theta; \theta_0)$ is changing only slowly in the vicinity of θ_0, the parameter is not sharply identified, in the sense that we know that a lot of sample information will be required to elucidate the true θ_0. This is reflected in the property that $H''(\theta_0 : \theta_0)$ will in these circumstances be close to zero, resulting in a low precision in estimation. (Recall, for instance, the asymptotic covariance matrix in maximum likelihood estimation.) Thus the criterion provides an integrated approach to identification and estimation theory.

Finally, suppose that θ_0 turns out to be unidentified. Solving the equations $H'(\theta; \theta_0) = 0$, together with the second-order conditions based on a negative semidefinite $H''(\theta; \theta_0)$, will then yield the subspace ω of observationally-equivalent parameters. Knowledge of ω may be information well worth having (as in, say, ratios of parameters), and the H function criterion is a very general way of delineating this subspace.

We conclude with an example which illustrates the criterion in operation.

[1] An alternative proof is to note that $H(\theta; \theta_0)$ is certainly locally concave (albeit, perhaps not strictly so); thus any critical point will of necessity be a maximum. Now critical points are zeros of $H'(\theta;\theta_0)$; and H'' can be regarded as the Jacobian of H'. The required necessary condition then follows from the necessary condition for an isolated zero of H' (see the Appendix to Chapter 5 of [**1**]).

EXAMPLE: Suppose $\boldsymbol{y} = (y_1, \ldots, y_T)'$ is distributed normally, with mean $X\boldsymbol{\beta}_0$ and nonsingular covariance matrix Ω_0. We denote by θ the set of coefficients $\boldsymbol{\beta}$ and Ω.

$$
\begin{aligned}
(8) \qquad H(\theta;\theta_0) &= \tfrac{1}{2}\log\left(\frac{\det \Omega_0}{\det \Omega}\right) + \tfrac{1}{2}E(y - X\boldsymbol{\beta}_0)\Omega_0^{-1}(y - X\boldsymbol{\beta}_0) \\
&\quad - \tfrac{1}{2}E(\boldsymbol{y} - X\boldsymbol{\beta})'\Omega^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}); \boldsymbol{\beta}_0, \Omega_0 \\
&= \tfrac{1}{2}\log\left(\frac{\det \Omega_0}{\det \Omega}\right) \\
&\quad + \tfrac{1}{2}(T - \operatorname{tr}\Omega^{-1}\Omega_0 - (\beta - \beta_0)'X'\Omega^{-1}X(\beta - \beta_0)).
\end{aligned}
$$

Now

$$
(9) \qquad \frac{\partial H}{\partial \boldsymbol{\beta}} = -(X'\Omega^{-1}X)(\boldsymbol{\beta} - \boldsymbol{\beta}_0).
$$

This can equal zero for $\beta \neq \beta_0$ only if X has incomplete rank. Thus if X has full rank, the $\boldsymbol{\beta}$ of any observationally equivalent parameter set must be equal to $\boldsymbol{\beta}_0$, so that the last term of (8) vanishes. To avoid prolixity we merely assume that the resulting equation in Ω has only one solution, to wit $\Omega = \Omega_0$.

Suppose now that X has not full rank. Then $\boldsymbol{\beta}_0$ is not identified. For if $\boldsymbol{\beta} - \boldsymbol{\beta}_0$ belongs to the null space (i.e., if $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'$ belongs to the row kernel) of $X'\Omega^{-1}X$, then the last term of (8) vanishes, so that certainly the pair (β, Ω_0) is observationally equivalent to (β_0, Ω_0).

It is obvious from (9) that the only observationally equivalent $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ must belong to the null space of $X'\Omega^{-1}X$, so that the last term in (8) is in fact always zero for such $\boldsymbol{\beta}$. The remaining equation is in Ω alone, and it is thus clear from comparison with the full rank case, that Ω is still identified. Thus if X has incomplete rank, $\boldsymbol{\beta}_0$ is not identified, but Ω_0 is. The observationally-equivalent subspace ω is $\{\beta_0 + \text{null space } [X'\Omega_0^{-1}X]\}$.

University of Auckland

Manuscript received December 1971; revision received September 1972.

REFERENCES

[1] FISHER, F. M.: *The Identification Problem in Econometrics*. New York: McGraw-Hill, 1966.
[2] FRASER, D. A. S.: *The Structure of Inference*. New York: John Wiley, 1968.
[3] KOOPMANS, T. C., AND REIERSOL, O.: "The Identification of Structural Characteristics," *Annals of Mathematical Statistics*, 21 (1950), 165–181.
[4] KULLBACK, S.: *Information Theory and Statistics*. New York: John Wiley, 1959.
[5] ROTHENBERG, THOMAS J.: "Identification in Parametric Models," *Econometrica*, 39 (1971), 577–592.
[6] WILKS, SAMUEL S.: *Mathematical Statistics*. New York: John Wiley, 1962.