EVALUATION OF ECONOMETRIC MODELS BY DECOMPOSITION AND AGGREGATION

Gregory C. Chow*

Econometric Research Program
Research Memorandum No. 235

Revised November 1979

[*] This paper is a revised version of a paper presented before the NSF Conference on Macroeconometric Models co-chaired by J. Kmenta and J. Ramsey held in Ann Arbor, Michigan, on Oct. 26-27, 1978. I would like to acknowledge financial support from the National Science Foundation through Grant No. SOC77-07677.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, New Jersey

EVALUATION OF ECONOMETRIC MODELS BY DECOMPOSITION AND AGGREGATION

Gregory C. Chow*

1. Introduction

Is an econometric model too large, or not large enough? This is the question which we hope to answer in this paper. This question is concerned with the choice of alternative models differing in size. An answer could conceivably come from *a priori* reasoning or theorizing, but the one suggested here is based purely on statistical inference using the information contained in a finite sample of n observations. We shall develop in Section 2 a statistical criterion for model selection, and derive in Section 3 explicit expressions based on this criterion for the selection of simultaneous-equation models of different sizes.

We will define the "size" of an econometric model designed to explain a given set of dependent variables by the number of functionally independent parameters in the model. By this definition, the size of a macroeconomic model consisting of a system of simultaneous stochastic equations will depend on whether it can be decomposed or aggregated. Given the same set of dependent variables to be explained, if a model is decomposable into submodels each capable of explaining a subset of dependent variables, then the number of parameters required to explain any subset would be smaller than in the case of a fully integrated system of simultaneous equations. If the model is not decomposable but block-recursive, the matrix of the Jacobian for transforming the random residuals into the dependent variables will be block-triangular and the covariance matrix of the residuals will be block-diagonal, both leading to a smaller number of parameters than in a completely interdependent system. By aggregating across equations, one is also likely to reduce the number of parameters required to explain any subset of dependent variables, thus reducing the size of the model for each subset. In Section 4, the statistical selection criterion will be applied to decide whether a model should be made block-triangular and whether certain dependent variables should be aggregated. Section 5 contains concluding remarks.

The basic viewpoint taken is that the better of two models is the one which, by the method of its construction, will on the average predict future observations better. Here better predictions could be defined by a smaller expected sum of squared deviations from the future observations. However, we will define better predictions by a smaller expected sum of the log-likelihood ratios of the true density of the future observations to the density specified by the model. Specifically, let $g(\cdot)$ be the true density of each of n independent future observations $(\tilde{y}_1,\ldots,\tilde{y}_n) = \tilde{Y}$, and let $f(\cdot|\theta)$ be the density specified by a possible model. Better predictions by $f(.|\theta)$ will be defined by a smaller expectation

$$(1.1) \qquad I_n[g;f(\cdot|\theta) = E \sum_{i=1}^{n} [\log g(\tilde{y}_i) - \log f(\tilde{y}_i|\theta)] \geq 0$$

where the expectation is evaluated by the true density $g(\cdot)$. The mean log-likelihood ratio $I[g;f(\cdot|\theta)] = E\log[g(y)/f(y|\theta)]$ is also called the mean information for discrimination between $g(y)$ and $f(y|\theta)$, as discussed in Kullback and Leibler (1951), Savage (1954, p. 50), and Kullback (1959, p. 5).

To illustrate the use of the mean log-likelihood $E(\log f(y|\theta)]$ as a measure of how well f approximates g , consider a univariate y having a normal distribution with mean μ and variance v . If the approximate distribution is normal with mean θ_1 and variance θ_2 , its mean log-likelihood is

$$(1.2) \qquad E[\log f(y|\theta)] = E\{-\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log\theta_2 - \tfrac{1}{2}(y-\theta_1)^2/\theta_2\}$$

$$= \tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log\theta_2 - \tfrac{1}{2}[v+(\theta_1-\mu)^2]/\theta_2 \quad .$$

For any given θ_2, the value of θ_1 which maximizes the mean log-likelihood (1.2) is the true mean μ itself. If $\theta_1 = \mu$, the value of θ_2 which maximizes (1.2) is the true v itself. This measure is better than the mean squared prediction error for judging the goodness of fit for f. If we let $\theta_1 = \mu$ but $\theta_2 = 5v$, the mean squared prediction error would be $E(y-\theta_1)^2 = v$, the smallest attainable. However, the model f may be a very poor approximation of g because θ_2 is very different from the true v.

Having adopted better predictions, defined by the information measure (1.1), as the criterion for selecting alternative models, we would like to stress that the correct model, even if it is known, is not necessarily the one to be selected because it may contain too many unknown parameters. For example, let the true regression model be linear in X_1 and X_2 with coefficient vectors β_1 and β_2. If β_2 is fairly small, and if only a finite sample of n observations is available, the model linear in X_1 alone as estimated by the method of least squares may yield better predictions than the model linear in both X_1 and X_2. The reason is that the larger model, though correctly specified, may be more poorly estimated because of the larger number of parameters. Larger sampling errors in the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ may lead to larger prediction errors for future observations. Thus if two different models $f_1(\cdot|\theta_1)$ and $f_2(\cdot|\theta_2)$ are proposed for the prediction of the same dependent variable, one should not merely ask how well f_1 and f_2 would do when θ_1 and θ_2 can be consistently estimated by an infinite sample, but how well the estimated $f_1(\cdot|\hat{\theta}_1)$ and $f_2(\cdot|\hat{\theta}_2)$ based on a finite sample would do, on the average, allowing for the sampling distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$. To predict n future observations $(\tilde{y}_1,\ldots,\tilde{y}_n) = \tilde{Y}$, the model selection criterion is

$$(1.3)\qquad E_{\hat{\theta}} I_n(g;f(\cdot|\hat{\theta})) \equiv E_{\hat{\theta}}\{E_{\tilde{Y}} \sum_{i=1}^{n} [\log g(\tilde{y}_i) - \log f(\tilde{y}_i|\hat{\theta})]\} .$$

Akaike (1973; 1974) adopted the mean log-likelihood ratio (or information) of a future observation for model selection and proposed an estimate of the mean of $E_{\tilde{y}}[\log f(\tilde{y}|\hat{\theta})]$ over the sampling distribution of the maximum likelihood estimator $\hat{\theta}$. We will modify Akaike's derivation and correct an error in his estimate of $E_{\hat{\theta}}E_{\tilde{y}}[\log f(\tilde{y}|\hat{\theta})]$, thus proposing an alternative information criterion for model selection. We will also apply this criterion to the selection of simultaneous-equations models.

2. Derivation of An Information Criterion

We assume that the true density $g(\cdot)$ equals $f(\cdot|\theta^o)$ and that an approximate model results from imposing a set of r linear restrictions $H'\theta = -b$ on the parameters. The purpose of this section is to provide an estimate of $E_{\hat{\theta}}I_n[g;f(\cdot|\hat{\theta})]$ where $\hat{\theta}$ is the maximum likelihood estimator of θ subject to the restrictions $H'\theta = -b$ imposed by the approximate model. Our derivation consists of the following five steps.

First, assuming θ to be given, we will approximate $I_n[g;f(\cdot|\theta)]$ by a quadratic form in $\theta-\theta^o$. The mean information for discrimination between $g(\cdot)$ and $f(\cdot|\theta)$ using n future observations $(\tilde{y}_1,\ldots,\tilde{y}_n) = \tilde{Y}$ is

$$(2.1)\qquad I_n[g;f(\cdot|\theta)] = E_{\tilde{Y}}\sum_{i=1}^{n}[\log g(\tilde{y}_i) - \log f(\tilde{y}_i|\theta)]$$

$$= E_{\tilde{Y}}[\log L(\tilde{Y};\theta^o) - \log L(\tilde{Y};\theta)]$$

where $L(\tilde{Y};\theta)$ denotes the likelihood function. Expanding $\log L(\tilde{Y};\theta)$ in a second-order Taylor series about θ^o and substituting the result into (2.1), we obtain, after using the fact $E(\partial\log L(\tilde{Y};\theta^o)/\partial\theta) = 0$,

$$(2.2)\qquad I_n[g;f(\cdot|\theta)] = \tfrac{1}{2}(\theta-\theta^o)'J(\theta^+,\theta^o)(\theta-\theta^o)$$

where $\theta \leq \theta^+ \leq \theta^o$ and

$$J(\theta^{+},\theta^{o}) \equiv - E_{\tilde{Y}} \frac{\partial^2 \log L(\tilde{Y};\theta^{+})}{\partial\theta\partial\theta'} \quad ,$$

the parameter θ^{o} of $J(\theta^{+},\theta^{o})$ being used to define the distribution of $\tilde{Y}$. $J(\theta^{o};\theta^{o})$ is Fisher's information matrix.

Second, given the linear restrictions $H'\theta + b = 0$, we find the best approximate model by minimizing the information $n^{-1}I_n[g;f(\cdot|\theta)]$ with respect to θ subject to $H'\theta + b = 0$. Using (2.2) for I_n, we differentiate the Lagrangian expression (suppressing the arguments of J)

$$(2.3) \qquad \frac{1}{2n}(\theta-\theta^{o})'J(\theta-\theta^{o}) - \lambda'(H'\theta+b)$$

to yield

$$(2.4) \qquad \begin{bmatrix} n^{-1}J & -H \\ -H' & 0 \end{bmatrix} \begin{bmatrix} \theta^{*} \\ \lambda^{*} \end{bmatrix} = \begin{bmatrix} n^{-1}J\theta^{o} \\ b \end{bmatrix}$$

the solution of which is

$$(2.5) \qquad \theta^{*} = \theta^{o} + nJ^{-1}H\lambda^{*} \quad ; \quad \lambda^{*} = - n^{-1}(H'J^{-1}H)^{-1}(H'\theta^{o} + b) \ .$$

The vector θ^{*} can be considered the parameter of the approximate model and is called the pseudo-true parameter of the pseudo-true model in the language of Sawa (1978).

Third, if $\hat{\theta}^{*}$ is any estimate (not necessarily maximum likelihood) of θ^{*} satisfying $H'\hat{\theta}^{*} = -b$, we substitute $(\hat{\theta}^{*}-\theta^{*}) + (\theta^{*}-\theta^{o})$ for $(\theta-\theta^{o})$ in (2.2) to obtain the information measure for the estimated model

$$(2.6) \qquad I_n[g;f(\cdot|\hat{\theta}^{*})] = \frac{1}{2}(\hat{\theta}^{*}-\theta^{*})'J(\hat{\theta}^{*}-\theta^{*}) + \frac{1}{2}(\theta^{*}-\theta^{o})'J(\theta^{*}-\theta^{o})$$

where the cross-product $(\hat{\theta}^{*}-\theta^{*})'J(\theta^{*}-\theta^{o})$ has vanished on account of (2.5) and $H'\theta^{*} = H'\hat{\theta}^{*} = - b$. Parenthetically, the method of maximum likelihood is justfied as it chooses θ to minimize the information measure based on the sample $Y = (y_1,\ldots,y_n)$, i.e.,

$$\sum_{i=1}^{n} \log f(y_i|\theta^o) - \sum_{i=1}^{n} \log f(y_i|\theta) .$$

The first term of (2.6) captures the discrepancy between the estimated model and the best approximate model which is due to the sampling error in $\hat{\theta}^*$. The second term measures the discrepancy between the best approximate model and the true model which is the result of specification error. Reducing the number of restrictions on θ or increasing the number of linearly independent parameters will reduce the specification error while it raises the sampling error of the parameter estimate. The remaining problem is to estimate the expectation of (2.6) over the sampling distribution of $\hat{\theta}^*$.

Fourth, given a sample $Y = (y_1,\ldots,y_n)$ of n observations, we will estimate the second term of (2.6) as follows.

$$(2.7) \qquad \tfrac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o) = I_n[g;f(\cdot|\theta^*)] = E\log L(\tilde{Y};\theta^o) - E\log L(\tilde{Y};\theta^*)$$

$$\simeq E\log L(\tilde{Y};\theta^o) - \log L(Y;\theta^*)$$

where we have estimated $E\log L(\tilde{Y};\theta^*)$ by its sample analogue $\log L(Y;\theta^*)$. We will find the maximum likelihood estimate $\hat{\theta}^*$ and expand $\log L(Y;\theta^*)$ in a second-order Taylor series about $\hat{\theta}^*$. $\hat{\theta}^*$ is found by differentiating

$$(2.8) \qquad n^{-1}\log L(Y;\theta) + \lambda'(H'\theta+b)$$

to yield

$$(2.9) \qquad n^{-1}\frac{\partial \log L(Y;\hat{\theta}^*)}{\partial\theta} + H\hat{\lambda} = 0 .$$

Expanding $\log L(Y;\theta^*)$ in (2.7) about $\hat{\theta}^*$, we obtain

$$(2.10)\qquad \frac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o) \simeq E\log L(\tilde{Y};\theta^o)$$

$$- \log L(Y;\hat{\theta}^*) - \frac{1}{2}(\theta^*-\hat{\theta}^*)'\frac{\partial^2\log L(Y;\theta^*)}{\partial\theta\partial\theta'}(\theta^*-\hat{\theta}^*)$$

where we have observed $(\theta^*-\hat{\theta}^*)'[\partial\log L(Y;\hat{\theta}^*)/\partial\theta] = 0$ on account of (2.9) and $(\theta^*-\hat{\theta}^*)'H = 0$.

If $-\partial^2\log L(Y;\theta^*)/\partial\theta\partial\theta'$ in (2.10) is replaced by its expectation $J(\theta^*,\theta^o)$, and (2.10) is combined with (2.6), the result is

$$(2.11)\qquad I_n[g;f(\cdot|\hat{\theta}^*)] \simeq E\log L(\tilde{Y};\theta^o) - \log L(Y;\hat{\theta}^*) + (\hat{\theta}^*-\theta^*)'J(\theta^*,\theta^o)(\hat{\theta}^*-\theta^*) .$$

Since $E\log L(\tilde{Y};\theta^o)$, though unknown, is constant among alternative models obtained by specifying different sets of restrictions on θ , it can be ignored for the purpose of model selection.

Fifth, we arrive at a criterion for model selection by taking the expectation of $-I_n[g;f(\cdot|\hat{\theta}^*)]$ given by (2.11) over the sampling distribution of $\hat{\theta}^*$, (plus the above constant term), i.e.,

$$(2.12)\qquad - E_{\hat{\theta}^*}I_n[g;f(\cdot|\hat{\theta}^*)] + E\log L(\tilde{Y};\theta^o)$$

$$\simeq \log L(Y;\hat{\theta}^*) - \mathrm{tr}\{J(\theta^*,\theta^o)E(\hat{\theta}^*-\theta^*)(\hat{\theta}^*-\theta^*)'\}$$

The models will be ranked by (2.12),the one having the highest value to be selected. The remaining problem is to provide estimates of $J(\theta^*,\theta^o)$ and $E(\hat{\theta}^*-\theta^*)(\hat{\theta}^*-\theta^*)$.

To find the distribution of the maximum likelihood estimate $\hat{\theta}^*$ subject to the restrictions $H\theta^* = - b$, we follow the work of Silvey (1959). Expanding $\partial\log L(Y;\hat{\theta}^*)/\partial\theta$ in (2.9) about θ^* , we get

$$(2.13)\qquad n^{-1}\frac{\partial\log L(Y;\theta^*)}{\partial\theta} + \left[n^{-1}\frac{\partial^2\log L(Y;\theta^*)}{\partial\theta\partial\theta'} + o(1)\right][\hat{\theta}^*-\theta^*] + H\hat{\lambda} = 0 .$$

Since θ^* is obtained by minimizing $n^{-1}I_n[g;f(\cdot|\theta)]$ as we did in (2.3), or alternatively by maximizing $n^{-1}E\log L(Y;\theta)$, subject to $H\theta + b = 0$, we have

$$(2.14) \qquad n^{-1}\frac{\partial E\log L(Y;\theta^*)}{\partial\theta} + H\lambda^* = 0 .$$

Subtraction of (2.14) from (2.13) yields

$$(2.15) \qquad \begin{bmatrix} -n^{-1}\frac{\partial^2\log L(Y;\theta^*)}{\partial\theta\partial\theta'} + o(1) & -H \\ -H' & 0 \end{bmatrix} \begin{bmatrix} \hat{\theta}^*-\theta^* \\ \hat{\lambda}-\lambda^* \end{bmatrix} = \begin{bmatrix} n^{-1}\frac{\partial\log L(Y;\theta^*)}{\partial\theta} - n^{-1}\frac{\partial E\log L(Y;\theta^*)}{\partial\theta} \\ 0 \end{bmatrix}$$

Abbreviating $L(Y;\theta^*)$ by L^*, we observe that the asymptotic distribution of $n^{-1}\left[\partial\log L^*/\partial\theta - \partial E\log L^*/\partial\theta\right]$ is normal by the central limit theorem and its mean is zero by the law of large numbers. The covariance matrix of $n^{-\frac{1}{2}}[\partial\log L^*/\partial\theta - \partial E\log L^*/\partial\theta]$ is

$$(2.16) \qquad V_{\theta^*} = n^{-1}\left[E\frac{\partial\log L^*}{\partial\theta}\cdot\frac{\partial\log L^*}{\partial\theta'} - \frac{\partial E\log L^*}{\partial\theta}\cdot\frac{\partial E\log L^*}{\partial\theta'}\right] .$$

As n increases, the sample mean $-n^{-1}\partial^2\log L^*/\partial\theta\partial\theta'$ approaches its expectation $n^{-1}J(\theta^*,\theta^o)$. Therefore, the solution of (2.15) yields an asymptotic distribution for $n^{\frac{1}{2}}(\hat{\theta}^*-\theta^*)$ and $n^{\frac{1}{2}}(\hat{\lambda}-\lambda^*)$ which is normal with mean 0 and covariance matrix

$$(2.17) \qquad \begin{bmatrix} P_{\theta^*}V_{\theta^*}P_{\theta^*} & P_{\theta^*}V_{\theta^*}Q_{\theta^*} \\ Q'_{\theta^*}V_{\theta^*}P_{\theta^*} & Q'_{\theta^*}V_{\theta^*}Q_{\theta^*} \end{bmatrix}$$

where

$$(2.18)\qquad \begin{bmatrix} P_{\theta^*} & Q_{\theta^*} \\ Q'_{\theta^*} & R_{\theta^*} \end{bmatrix} = \begin{bmatrix} n^{-1}J(\theta^*,\theta^o) & -H \\ -H' & 0 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} nJ^{-1} - nJ^{-1}H(H'J^{-1}H)^{-1}H'J^{-1} & -J^{-1}H(H'J^{-1}H)^{-1} \\ -(H'J^{-1}H)^{-1}H'J^{-1} & -n^{-1}(H'J^{-1}H)^{-1} \end{bmatrix} .$$

This result was given by Silvey (1959, Lemma 1, p. 394).

In the important special case when the restrictions consist entirely of zero restrictions on a subset of parameters, we write $\theta^* = (\theta_1^*\ 0)$, $H' = [0\ I]$, and

$$(2.19)\qquad J(\theta^*,\theta^o) = \begin{bmatrix} J_{11}(\theta^*,\theta^o) & J_{12}(\theta^*,\theta^o) \\ J_{21}(\theta^*,\theta^o) & J_{22}(\theta^*,\theta^o) \end{bmatrix}$$

The matrix P_{θ^*} from (2.18) becomes

$$(2.20)\qquad P_{\theta^*} = \begin{bmatrix} nJ_{11}^{-1}(\theta^*,\theta^o) & 0 \\ 0 & 0 \end{bmatrix}$$

and the covariance matrix of $(\hat{\theta}_1^* - \theta_1^*)$ from (2.17) becomes

$$(2.21)\qquad J_{11}^{-1}(\theta^*,\theta^o)\left[E\frac{\partial \log L^*}{\partial \theta_1} \cdot \frac{\partial \log L^*}{\partial \theta_1'}\right]J_{11}^{-1}(\theta^*,\theta^o)$$

since $\partial E\log L^*/\partial\theta_1 = 0$ as θ_1^* is obtained by maximizing (differentiating) $E\log L(Y;\theta_1,0)$ with respect to θ_1 . Combining (2.21) with (2.12), we have the following model selection criterion in this case:

$$(2.22)\qquad \log L(Y;\hat{\theta}^*) - \mathrm{tr}\{E\left[\frac{\partial \log L^*}{\partial \theta_1} \cdot \frac{\partial \log L^*}{\partial \theta_1'}\right]J_{11}^{-1}(\theta^*,\theta^o)\} .$$

Akaike (1973) was incorrect in claiming that $J_{11}^{-1}(\theta^*,\theta^o)$ is the asymptotic covariance matrix of $\hat{\theta}_1^*$, as we have shown in (2.21). If this claim were valid, the trace term in (2.12) would become k, the number of unknown parameters in θ_1, and (2.12) would become Akaike's information criterion which selects the model having the largest value for the maximum log-likelihood minus the number of parameters to be estimated. The claim is incorrect because only when the model is correctly specified, i.e., when $\theta^* = \theta^o$, do we have $J_{11}^{-1}(\theta^*,\theta^*)$ as the asymptotic covariance matrix of $\hat{\theta}^*$. In order to apply our criterion (2.12) to simultaneous-equation models, we have to estimate $J(\theta^*,\theta^o)$ and $E(\hat{\theta}^*-\theta^*)(\hat{\theta}^*-\theta^*)$ as given by (2.17), or by (2.21) in the special case of zero restrictions on θ^*. This is our task in the next section.

3. Estimation of the Information Criterion for Simultaneous-Equations Models

In this section, we will provide an estimate of the information criterion for the selection of linear simultaneous-equations models, while leaving a discussion of its econometric applications to the following section. Let the true model be

$$(3.1) \qquad Y\Gamma^o + XB^o = U \qquad\qquad EU'U = n\Sigma^o \equiv nR^{o^{-1}}$$

where Y is an $n\times g$ matrix of endogenous variables, X is an $n\times k$ matrix of exogenous variables, and selected elements of Γ^o and B^o are zero because of the identification restrictions. Let the approximate model be

$$(3.2) \qquad Y\Gamma^* + XB^* = U^* \qquad\qquad EU^{*\prime}U^* = n\Sigma^* \equiv nR^{*^{-1}}$$

where the elements of Γ^*, B^* and Σ^* are subject to additional linear restrictions. The elements of these "pseudo-true" parameters are obtained by a constrained maximization of

$$(3.3)\qquad \mathrm{ElcgL}(Y;\Gamma,B,R) = \frac{ng}{2}\log(2\pi) + \frac{n}{2}\log|R| + n\log|\Gamma| - \frac{1}{2}\,\mathrm{tr}\{R\cdot E(Y\Gamma+XB)'(Y\Gamma+XB)\}$$

where the expectation E is evaluated by assuming that Y is generated by the true model.

To evaluate the matrix V_{θ^*} of (2.16), we need the derivatives of $\log \mathcal{L}(Y;\Gamma,B,R)$ evaluated at Γ^*, B^* and Σ^* minus their expectations. The derivatives are, with $U^* = Y\Gamma^* + XB^* = (u_1^* \ldots u_g^*)$,

$$(3.4)\qquad \frac{\partial \log L^*}{\partial B} = -\, X'U^*R^*$$

$$(3.5)\qquad \frac{\partial \log L^*}{\partial \Gamma} = -\, Y'U^*R^* + n(\Gamma^{*\prime})^{-1}$$

$$(3.6a)\qquad \frac{\partial \log L^*}{\partial r_{ij}} = n\sigma_{ij}^* - u_i^{*\prime} u_j^*$$

$$(3.6b)\qquad \frac{\partial \log L^*}{\partial r_{ii}} = \frac{1}{2}(n\sigma_{ii}^* - u_i^{*\prime} u_i^*) \;.$$

Defining the true reduced-form to be

$$(3.7)\qquad Y = -\, XB^{o}\Gamma^{o^{-1}} + U\Gamma^{o^{-1}} \equiv X\Pi^{o} + V \qquad EV'V = n\Omega^{o} \equiv n\Gamma^{o\prime^{-1}}\Sigma^{o}\Gamma^{o^{-1}} \;,$$

we can write

$$(3.8)\qquad U^* = Y\Gamma^* + XB^* = X(\Pi^{o}\Gamma^* + B^*) + V\Gamma^* = D + V\Gamma^*$$

where

$$(3.9)\qquad D = EU^* = X(\Pi^{o}\Gamma^* + B^*) \;.$$

Therefore, the derivatives given by (3.4) and (3.5) minus their expectations are

$$(3.10) \qquad \frac{\partial \log L^*}{\partial B} - E\frac{\partial \log L^*}{\partial B} = X'V\Gamma^* R^*$$

$$(3.11) \qquad \frac{\partial \log L^*}{\partial \Gamma} - E\frac{\partial \log L^*}{\partial \Gamma} = - \Pi^{o\prime} X'V\Gamma^* R^* - V'DR^* - V'V\Gamma^* R^* + n\Omega^o \Gamma^* R^* .$$

The expectations of (3.6a) and (3.6b) are zero.

Since only the unknown elements of B^* and Γ^* are of concern, we denote by β_i and γ_i respectively the column vectors consisting of only the unknown elements in the i-th columns of B^* and Γ^*. Similarly, X_i and Y_i denote the matrices composed of those columns of X and Y which are associated respectively with the unknown coefficients in β_i and γ_i. Also, we will denote $X\Pi^o$ by $\tilde{Y}^o$, $X\Pi^* \equiv - XB^*\Gamma^{*-1}$ by $\tilde{Y}^*$, and $\Gamma^{*\prime}\Omega^o\Gamma^*$ by W for convenience. Using these notations together with (3.10) and (3.11), we derive the required components of nV_{θ^*} as

$$(3.12) \qquad \mathrm{Cov}\left| \frac{\partial \log L^*}{\partial \beta_i} \quad \frac{\partial \log L^*}{\partial \beta_j'} \right| = X_i'X_j(r_i^{*\prime}Wr_j^*)$$

$$(3.13) \qquad \mathrm{Cov}\left| \frac{\partial \log L^*}{\partial \gamma_i} \quad \frac{\partial \log L^*}{\partial \gamma_j'} \right| = \tilde{Y}_i^{o\prime}\tilde{Y}_j^o(r_i^{*\prime}Wr_j^*) + \Omega_i^{o\prime}\Gamma^* r_j^* r_i^{*\prime} D'\tilde{Y}_j^o + \tilde{Y}_i^{o\prime} Dr_j^* r_i^{*\prime}\Gamma^*\Omega_j^o$$

$$+ n\Omega_{ij}^o r_{ij}^* + n\Omega_i^{o\prime}\Gamma^* r_j^* r_i^{*\prime}\Gamma^{*\prime}\Omega_j^o$$

$$(3.14) \qquad \mathrm{Cov}\left| \frac{\partial \log L^*}{\partial \gamma_i} \quad \frac{\partial \log L^*}{\partial \beta_j'} \right| = \tilde{Y}_i^{o\prime}X_j(r_i^{*\prime}Wr_j) + \Omega_i^{o\prime}\Gamma^+ r_j^* r_i^{*\prime} D'X_j$$

where Ω_i^o denotes a matrix composed of only those columns of $\Omega^o = (\omega_{ij}^o)$ which are associated with the unknown elements of γ_i^*, and Ω_{ij}^o denotes a matrix extracted from Ω^o whose rows correspond to the unknown elements of

γ_i^* and whose columns correspond to the unknown elements of γ_j^* . The proof of (3.13) has utilized the relation, for $V = (v_1 \ldots v_g)$,

$$E(v_1'v_2)(v_3'v_4) = n^2\omega_{12}^o\omega_{34}^o + n\omega_{13}^o\omega_{24}^o + n\omega_{14}^o\omega_{23}^o$$

which implies

$$EV'V\Gamma^*r_i^*r_j^{*\prime}\Gamma^{*\prime}V'V = n^2\Omega^o\Gamma^*r_i^*r_j^{*\prime}\Gamma^{*\prime}\Omega^o + n\Omega^o\Gamma^*r_j^*r_i^{*\prime}\Gamma^*\Omega^o + n\Omega_{ij}^o(r_i^{*\prime}\Gamma^{*\prime}\Omega^o\Gamma^*r_j^*) \ .$$

By contrast, the elements of $J(\theta^*,\theta^o)$ as derived from differentiating (3.4) and (3.5) are

$$(3.15) \qquad -E\frac{\partial \log L^*}{\partial\beta_i\partial\beta_j'} = X_i'X_jr_{ij}^*$$

$$(3.16) \qquad -E\frac{\partial \log L^*}{\partial\gamma_i\partial\gamma_j'} = \tilde{Y}_i^{o\prime}\tilde{Y}_j^o r_{ij}^* + n\Omega_{ij}^o r_{ij}^* + n\gamma^{j(i)}\gamma^{i(j)\prime}$$

$$(3.17) \qquad -E\frac{\partial \log L^*}{\partial\gamma_i\partial\beta_j'} = \tilde{Y}_i^{o\prime}X_jr_{ij}^*$$

where $\gamma^{i(j)}$ denotes a column vector consisting of those elements of the i-th row of $(\Gamma^*)^{-1}$ which correspond to the unknown elements of γ_j^* . Note that when $(\Gamma^*,B^*,R^*) = (\Gamma^o,B^o,R^o)$, i.e., when the approximate model coincides with the true model, (3.12), (3.13) and (3.14) will reduce to (3.15), (3.16) and (3.17) respectively, as $r_i^{*\prime}Wr_j^* = r_i^{o\prime}\Sigma^o r_j^o$ will become $r_{ij}^o = r_{ij}^*$ and $D = 0$.

We next derive the expectations involving the derivatives of $\log L$ with respect to r_{ij} . Using (3.6) we obtain by straightforward manipulations, with $\Gamma^{*\prime}\Omega^o\Gamma^* = W = (w_{ij})$,

$$(3.18a) \qquad E\left|\frac{\partial \log L^*}{\partial r_{ij}} \cdot \frac{\partial \log L^*}{\partial r_{k\ell}}\right| = n[\sigma_{ik}^*w_{j\ell} + \sigma_{jk}^*w_{i\ell} + (\sigma_{i\ell}^*-w_{i\ell})w_{jk} + (\sigma_{j\ell}^*-w_{j\ell})w_{ik}]$$

$$(3.18b) \qquad E\left|\frac{\partial \log L^*}{\partial r_{ii}} \cdot \frac{\partial \log L^*}{\partial r_{k\ell}}\right| = n[\sigma_{ik}^*w_{i\ell} + (\sigma_{i\ell}^*-w_{i\ell})w_{ik}]$$

$$(3.18c)\qquad E\left[\frac{\partial \log L^*}{\partial r_{ii}}\cdot\frac{\partial \log L^*}{\partial r_{kk}}\right] = \frac{n}{2}\,[\sigma^*_{ik}w_{ik} + (\sigma^*_{ik}-w_{ik})w_{ik}]$$

and the corresponding expressions

$$(3.19a)\qquad -E\left[\frac{\partial^2 \log L^*}{\partial r_{ij}\partial r_{k\ell}}\right] = n[\sigma^*_{ik}\sigma^*_{j\ell}+\sigma^*_{jk}\sigma^*_{i\ell}]$$

$$(3.19b)\qquad -E\left[\frac{\partial^2 \log L^*}{\partial r_{ii}\partial r_{k\ell}}\right] = n\sigma^*_{ik}\sigma^*_{i\ell}$$

$$(3.19c)\qquad -E\left[\frac{\partial^2 \log L^*}{\partial r_{ii}\partial r_{kk}}\right] = \frac{n}{2}\,\sigma^{*\,2}_{ik} \quad .$$

Again, when the approximate model coincides with the true model, we have $W = \Sigma^o = \Sigma^* = (\sigma^*_{ij})$, and (3.18) will be identical with (3.19).

As can be seen by differentiating (3.4) and (3.5), the expectations of $\partial^2\log L^*/\partial\beta_i\partial r_{k\ell}$ and $\partial^2\log L^*/\partial\gamma_i\partial r_{k\ell}$ are zero. Therefore, letting α denote a column vector composed of the unknown elements of $\beta_1,\ldots,\beta_g$, $\gamma_1,\ldots,\gamma_g$, r denote a column vector consisting of $r_{11},\ldots,r_{1g}$, $r_{22},\ldots,r_{2g}$, $r_{31},\ldots,r_{gg}$, and θ' denote $(\alpha'\ r')$, we can write

$$(3.20)\qquad nV_{\theta^*} = \mathrm{Cov}\left[\frac{\partial \log L^*}{\partial\theta}\right] \equiv \begin{bmatrix} \mathrm{Cov}\left(\frac{\partial \log L^*}{\partial\alpha}\right) & 0 \\ 0 & \mathrm{Cov}\left(\frac{\partial \log L^*}{\partial r}\right)\end{bmatrix}$$

where the elements of $\mathrm{Cov}(\partial\log L^*/\partial\alpha)$ and $\mathrm{Cov}(\partial\log L^*/\partial r)$ are given by (3.12)-(3.14) and (3.18) respectively. These matrices, together with the elements of $J(\theta^*,\theta^o)$ given by (3.15)-(3.17) and (3.19), provide an explicit expression for the asymptotic covariance matrix of $\hat\theta^*$ through (2.17) and also for the adjustment factor $\mathrm{tr}\{J(\theta^*,\theta^o)\mathrm{Cov}(\hat\theta^*)\}$ used in our model selection criterion (2.12). In actual applications, the parameters of the models (3.1) and (3.2) required to evaluate $J(\theta^*,\theta^o)$ and $Cov(\hat{\theta}^*)$ are unknown, but can be estimated by the method of maximum likelihood.

In the important special case when B^* , Γ^* and Σ^* are obtained by additional zero restrictions on the parameters of the model (3.1), Σ^* being block-diagonal, our model selection criterion becomes (2.22) with an adjustment factor equal to

$$(3.21)\qquad tr\{Cov\frac{\partial \log L^*}{\partial\theta_1})J_{11}^{-1}(\theta^*,\theta^o)\} = tr\{Cov(\frac{\partial \log L^*}{\partial\alpha})[-E\frac{\partial^2 \log L^*}{\partial\alpha\partial\alpha'}]^{-1}\}$$

$$+ tr\{Cov(\frac{\partial \log L^*}{\partial r})[-E\frac{\partial^2 \log L^*}{\partial r\partial r'}]^{-1}\}$$

where the four matrices on the right-hand side are given by (3.12) - (3.19).

To appreciate the result (3.21), consider the special case $\Gamma^o = \Gamma^* = I$ and $X_i = X$ for $(i = 1,\ldots,g)$, which is a model of g linear regressions. If the approximate model has k_1 explanatory variables, (3.21) is reduced to

$$(3.22)\qquad k_1 \sum_{i=1}^{g} (r_i^{*\prime}\Sigma^o r_i^*)/r_{ii}^* + tr\{Cov\left(\frac{\partial \log L^*}{\partial r}\right)\left[-E\frac{\partial^2 \log L^*}{\partial r\partial r'}\right]^{-1}\} .$$

For the case of a multiple regression model, with $g = 1$, (3.22) is further reduced to

$$(3.23)\qquad k_1\frac{\sigma_{11}^o}{\sigma_{11}^*} + \frac{\sigma_{11}^o}{\sigma_{11}^*}\left(2 - \frac{\sigma_{11}^o}{\sigma_{11}^*}\right)$$

which is identical with the result of Sawa (1978, Theorem 3.2, p. 1280). When the approximate regression model coincides with the true model, $\sigma_{11}^o = \sigma_{11}^*$; the adjustment constant (3.23) becomes $k_1 + 1$ or the number of parameters, as in Akaike's formula. In general, $\sigma_{11}^* > \sigma_{11}^o$ when the approximate model differs from the true model, and the adjustment factor will be smaller than the number of parameters. For example, let the true model have 8 parameters (7 coefficients plus σ^o) and the approximate model have only 6 coefficients, and let $\sigma^o_{11} = .9\sigma^*_{11}$. The adjustment constant for the approximate model is 5.4 + .99 = 6.39, smaller than 7 or the number of parameters. The difference between the two trace terms to be subtracted from the respective maximum likelihood functions is 8 - 6.39 = 1.61, as compared with 8 - 7 = 1 by Akaike's formula. Thus the rule (3.23) favors the small model more than Akaike's rule does. This example suggests that, when the model already contains many parameters, our information criterion is quite strict in allowing the addition of one more parameter. As Sawa (1978, p. 1283) has shown, the information criterion based on (3.23) is equivalent to a t test for an additional coefficient using a critical value which can be larger than 2 when k_1 is large and n is small.

4. Should a Macromodel be Decomposed or Aggregated?

If one accepts the view that the "true" economic world is a very large and interdependent system of simultaneous stochastic equations, as many economists tend to accept, one is faced with the almost insurmountable problem of estimating very large systems of simultaneous equations. After making significant contributions to the identification and estimation of simultaneous equations, T. C. Koopmans (1950) asked, "When is an equation system complete for statistical purposes?" He gave very strict statistical conditions which would permit one to specify certain variables as exogenous and/or predetermined for the purpose of explaining the remaining endogenous variables, thus reducing the size of the model for the latter variables. One wonders when, if ever, these strict conditions stated by Koopmans will be met. T. C. Liu (1955, 1960), being convinced that the "true" world is a completely interdependent system of simultaneous equations, questioned how one could ever estimate the true parameters even if the sample were infinite; the necessary conditions for identification would not be met since each equation contains too many variables. Franklin Fisher (1961), coming to the rescue, argued that if the coefficients of the dependent variables in each structural equation, though numerous, are mostly very small, then treating them as zero in order to satisfy the identification condition will only lead to very small inconsistencies in the estimation of the remaining parameters. On the other extreme, Herman Wold (1953) argued that the world is recursive anyway and there is no great statistical difficulty in estimating its parameters.

While we grant that the true economic model might very well be a very large and completely interdependent system of simultaneous equations, an econometrician might wish to estimate not the true model but only an approximate model because the sample is finite. One realizes that the conditions stated by Koopmans for defining the exogenous and/or predetermined variables are never met, that the coefficients of many endogenous and exogenous variables in a structural equation are not zero as Liu has pointed out, and that the true model for quarterly economic time series is not strictly recursive in the sense of Wold. However, one might not wish to raise the question of F. M. Fisher, whether by making certain assumptions necessary for identification, the remaining parameters in a true model can be almost consistently estimated. One is seldom in a position to estimate the parameters of the true model because the number of available observations is often smaller than the number of its parameters. One is mainly interested in the parameters θ^* of the approximate models because they are the models relevant for practical purposes. To illustrate, let the true model be

$$- y_{1t} + \theta_1 y_{2t} + \theta_2 x_{1t} + \theta_3 x_{2t} + \dots + \theta_{100} x_{99,t} = u_{1t}$$

$$\theta_{101} y_{1t} - y_{2t} - \theta_{102} x_{1t} + \theta_{103} x_{2t} + \dots + \theta_{200} x_{99,t} = u_{2t}$$

where all parameters are small except θ_1, θ_2, θ_{101} and θ_{103} . This model is unidentifiable. Fisher points out that if θ_3 and θ_{102} are extremely small, the remaining parameters can be estimted almost consistently. Our viewpoint is that the approximate model f with θ_1 , θ_2 , θ_{101} and θ_{103} as the only non-zero coefficients to be estimated might be the best approximation according to the information criterion when say 50 observations are available. Although the maximum likelihood estimators of θ_1^* , θ_2^* , θ_{101}^* and θ_{103}^* will not consistently estimate the true θ_1 , θ_2 , θ_{101} and θ_{103} , the model f can still be the best approximation for prediction purposes.

Furthermore, Fisher (1961) is concerned with the "cost of approximate specification in simultaneous equation estimation," implying that something is lost by using an approximate model because of the inconsistencies in the estimation of the true parameters. We wish to emphasize the "benefits" of an approximate specification because specification errors are not necessarily bad. In equation (2.6), the first term measures sampling errors in estimating θ^* by $\hat{\theta}^*$, and the second term measures specification errors in using $f(y|\theta^*)$ to approximate $g(y)$. Large specification errors from assigning zeros to coefficients may be compensated by smaller sampling errors and may produce a better model for prediction. One is less concerned, as Fisher was,about whether an extremely small specification error would obtain if the sample were infinite. Rather, one is more concerned with the total error, due to both specification and sampling, in using an estimated model for forecasting, realizing that the specification error will almost always be present. Even when one knows that a large model is more nearly correctly specified than a small model, the latter can still be selected by the information criterion. It is possible for the true world to be completely interdependent, but for a block-recursive model, estimated from a finite sample, to be a better approximation than an estimated simultaneous model. We will apply the selection criterion of Section 3 to decide which of two models to use, one being simultaneous and the other block-recursive, or one being disaggregated and the other aggregated.

First, consider the choice between a simultaneous model

$$(4.1) \qquad [Y_1 \quad Y_2]\begin{bmatrix}\Gamma_{11} & \Gamma_{12}\\ \Gamma_{21} & \Gamma_{22}\end{bmatrix} + X[B_1 \quad B_2] = [U_1 \quad U_2]$$

and a block-recursive model obtained by the restrictions $\Gamma_{21} = 0$ and $\Sigma_{12} = \frac{1}{n}EU_1'U_2 = 0$. The information criterion (2.22) - (3.21) can be applied to choose between them if they are both estimated by the method of (full-information) maximum likelihood. A statistical criterion is thus provided to decide whether a system of simultaneous econometric equations should be decomposed into two recursive blocks. Equivalently, it can be used to decide whether a general equilibrium or a partial equilibrium model should be selected. The latter model is represented by a block recursive system which treats y_1 as exogenous in the explanation of y_2 .

The second issue is whether one should aggregate across equations. For example, real consumption expenditures y_{1t} and y_{2t} for two commodity groups may satisfy

$$(4.2) \qquad \begin{aligned} y_{1t} &= \theta_1 y_{3t} + \theta_2{,}y_{1,t-1} + \theta_3 x_{1t} + u_{1t} \\ y_{2t} &= \theta_4 y_{3t} + \theta_5 y_{2,t-1} + \theta_6 x_{2t} + u_{2t} \end{aligned}$$

where y_{3t} may be disposable income and x_{1t} and x_{2t} relative prices. The sum of these equations is

$$(4.3) \qquad y_{1t} + y_{2t} = (\theta_1+\theta_4)y_{3t} + \theta_2{,}y_{1,t-1} + \theta_5 y_{2,t-1} + \theta_3 x_{1t} + \theta_6 x_{2t} + (u_{1t}+u_{2t}) \; .$$

Let $y_{4t} = y_{1t} + y_{2t}$ be aggregate consumption and let $x_{4t} = w_1 x_{1t} + w_2 x_{2t}$ be an aggregate price index with constant weights. An aggregate equation for y_{4t} can be written as

$$(4.4) \qquad y_{4t} = \theta_7 y_{3t} + \theta_2 y_{4,t-1} + (\theta_3/w_1)x_{4t} + u_{4t}$$

provided that

$$(4.5) \qquad \theta_5 = \theta_2 \quad \text{and} \quad \theta_6 = \theta_3(w_2/w_1) \ .$$

This example illustrates that aggregation across equations can be expressed as linear restrictions on the parameters of the disaggregate model. The choice between a disaggregate model and an aggregate one can be made by the information criterion. Three cases will be distinguished depending on the common subset of endogenous variables which both models are supposed to explain or predict.

In the first case, one is interested in predicting the individual components y_{1t} and y_{2t} as well as all other endogenous variables in the disaggregate model. One should retain equations (4.2) for the true model, and apply the information criterion to decide whether the restrictions (4.5) will yield a better approximate model. This is done by estimating the model using the method of maximum likelihood with and without these restrictions. The information criterion for the large model equals the maximum value of its log-likelihood minus the number of parameters. For the restricted model, it equals the maximum value of the log-likelihood minus an adjustment factor equal to $\mathrm{tr}\{J(\theta^*,\theta^o)(\mathrm{Cov}\hat{\theta}^*)\}$. Explicit expressions for $J(\theta^*,\theta^o)$ and $\mathrm{Cov}\hat{\theta}^*$ were given in Section 3.

In the second case, one is interested in predicting the aggregate $y_{4t} = y_{1t} + y_{2t}$ and all other endogenous variables in the model. One should then retain equation (4.3) instead of (4.2) for the true model, treating $(\theta_1+\theta_4)$ as one parameter. The approximate model imposes the restrictions (4.5) on the parameters of this equation.

In the third case, one is interested in predicting the aggregate y_{4t} and a (possibly small) subset of other endogenous variables, including the inflation rate and the unemployment rate, for example. The true model and the approximate model are as defined in the last paragraph. This case differs from the first two cases since only a subset of endogenous variables are of concern. We will have to consider the reduced-form equations for the subset in question. Two solutions to this model selection problem can be given.

For the first solution, the reduced-form equations for each model are estimated by the method of least squares, or maximum likelihood without allowing for the overidentifying restrictions from the structure. Here the two models explaining the common subset of endogenous variables are treated simply as two linear systems of regression equations. If the true model is written to include $y_{4,t-1}$, $y_{2,t-1}$, $x_{4,t}$ and $x_{2,t}$ as its predetermined variables, the approximate model excludes $y_{2,t-1}$ and $x_{2,t}$. To estimate the expected information for the approximate model, one can subtract the adjustment constant given by (3.22) from the maximum likelihood of the reduced-form explaining the subset of endogenous variables of interest.

For the second solution, the estimates of the reduced-form parameters are derived from the full-information maximum likelihood estimates of the parameters of the corresponding structures. The expected information for the approximate model can be estimated by evaluating the two terms given in (2.12). The first term $\log L(Y;\hat{\theta}^*)$ is the log-likelihood of the reduced-form for the selected endogenous variables evaluated at $\hat{\theta}^*$, which here denotes the above derived estimates of the reduced-form parameters. The second term equals the trace of the product of $J(\theta^*,\theta^o)$ and $\mathrm{Cov}(\hat{\theta}^*)$. $-J(\theta^*,\theta^o)$ is the expectation of the matrix of the second partials of the above log-likelihood with respect to the elements of θ^* . Explicit formulas for its elements are given by (3.15) and (3.19). The remaining task is the estimation of $\mathrm{Cov}(\hat{\theta}^*)$. The covariance matrix of the estimates $\hat{\alpha}^*$ and $\hat{r}^*_{ij}$ of the structural parameters, from which the reduced-form parameters $\hat{\theta}^*$ are derived, can be obtained by using the formulas given in Section 3. Given this covariance matrix, the covariance matrix $\mathrm{Cov}(\hat{\theta}^*)$ of the estimates of the corresponding reduced-form parameters can be estimated by the formula given in Dhrymes (1973, p. 122). This solution is applicable to the choice between any two linear simultaneous-equation models for the purpose of explaining a common subset of endogenous variables, provided that one can write down a general model as the true model and express both models by suitable linear restrictions on the parameters of the true model. One of the two models might serve as the true model if they are nested, as in our discussion of aggregation.

5. Concluding Remarks

It is not difficult, at least in principle, to extend our result to the selection of nonlinear simultaneous equations and of equations estimated by methods other than full-information maximum likelihood. No matter whether the model is linear or not, provided that the estimate $\hat{\theta}^*$ of θ^* is consistent and satisfies the restriction $H'\hat{\theta}^* = -b$ and $(\theta^*-\hat{\theta}^*)'[\partial \log L(Y;\hat{\theta}^*)/\partial\theta]$ is approximately zero, our information criterion (2.12) remains valid as it can be seen by reviewing the five steps used in its derivation. To estimate (2.12), one can easily evaluate $\log L(Y;\hat{\theta}^*)$ and approximate $J(\theta^*,\theta^o)$ by taking analytical or numerical derivatives for $-\partial^2 \log L(Y;\hat{\theta}^*)/\partial\theta\partial\theta'$. The more difficult problem is to estimate the covariance matrix of $\hat{\theta}^*$ when the approximate model is incorrect, i.e., when $\theta^o \neq \theta^*$. If computational expenses are not an issue, one can always apply Monte Carlo to find the covariance matrix of $\hat{\theta}^*$ under the assumption that the true parameter vector equals its estimate $\hat{\theta}^o$ which is obtained by the same method as $\hat{\theta}^*$ is. It remains a problem to find a computationally less expensive way to estimate the covariance matrix of $\hat{\theta}^*$.

An alternative approach to model selection is to rank a model by the

Jeffreys-Bayes posterior probability for it to be true after the data Y are observed. If $L(Y;\theta)$ is the likelihood function specified by the model M, the posterior probability $P(M|Y)$ for the model to be true equals the prior probability $P(M)$ for the model times

$$\int L(Y;\theta)p(\theta|M)d\theta = E_{\theta}L(Y;\theta)$$

where $p(\theta|M)$ is the prior density of the parameter of the model M. If $P(M)$ are equal for all models, the posterior probability criterion selects the model having the highest $\log E_{\theta}L(Y;\theta)$ where the expectation E_{θ} is evaluated by the prior density of θ. By contrast, the information criterion selects the model having the highest $E_{\hat{\theta}}E_{\tilde{Y}}\log L(\tilde{Y};\hat{\theta})$ where $\tilde{Y}$ denotes future observations and the expectation $E_{\hat{\theta}}$ is evaluated by the sampling distribution of $\hat{\theta}$ based on the data Y. The former criterion uses the data Y to judge a model specified by $L(\cdot;\theta)$ <u>and</u> by the prior density $p(\theta|M)$. The latter criterion uses future observations $\tilde{Y}$ to judge a model specified by $L(\cdot;\hat{\theta})$ where $\hat{\theta}$ has been estimated by the sample data Y. Insofar as the econometric models to be selected refer to models which have been estimated by the sample data for future prediction, and not models which had been specified before the sample period together with some prior density function $p(\theta|M)$ of its parameter vector, the information criterion appears to be more relevant.

[*] I would like to thank an anonymous reviewer for his helpful comments and acknowledge support from the National Science Foundation under grant No. SOC77-07677.

References

[1] Akaike, H.: "Information Theory and an Extension of the Maximum Likelihood Principle," in Proc. 2nd Int. Symp. Information Theory, ed. by B. N. Petrov and F. Csåki. Budapest: Academiai Kiadó,1973, 267-281.

[2] __________: "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, AC-19 (1974), 716-723.

[3] Dhrymes, P. J.: "Restricted and Unrestricted Reduced Forms: Asymptotic Distribution and Relative Efficiency," Econometrica, 41 (1973), 119-134.

[4] Fisher, F. M.: "On the Cost of Approximate Specification in Simultaneous Equation Estimation," Econometrica, 29, (1961), 139-170.

[5] Koopmans, T. C.: "When Is an Equation System Complete for Statistical Purposes?" in Chapter XVII, T. C. Koopmans, ed., Statistical Inference in Dynamic Economic Models. New York: John Wiley and Sons, 1950.

[6] Kullback, S.: Information Theory and Statistics. New York: John Wiley and Sons, 1950.

[7] Kullback, S. and R. A. Leibler: "On Information and Sufficiency," Annals of Mathematical Statistics, 22 (1951), 79-86.

[8] Liu, T. C.: "A Simple Forecasting Model for the U. S. Economy," International Monetary Fund, Staff Papers, (1955), 434-466.

[9] __________: "Underidentification, Structural Estimation, and Forecasting," Econometrica, 28 (1960), 855-865.

[10] Silvey, S. D.: "The Lagrangian Multiplier Test," Annals of Math. Statistics, 30 (1959), 389-407.

[11] Wold, H. in association with L. Juréen: Demand Analysis, New York: John Wiley and Sons, 1953.