BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR–89/08

LIKELIHOOD RATIO PROCEDURES FOR COMPARING NON–NESTED, POSSIBLY INCORRECT REGRESSORS

by

David F. Findley
U.S. Bureau of the Census
Statistical Research Division
Room 3000, FOB #4
Washington, DC 20233 U.S.A.

C.–Z. Wei
Department of Mathematics
University of Maryland
College Park, MD 20742

Report completed: August 10, 1989

Report issued: August 10, 1989

Report corrected: May 28, 1991

Beyond Chi–Square: Likelihood Ratio Procedures for Comparing Non–Nested, Possibly Incorrect Regressors.

David F. Findley, Bureau of the Census

C. Z. Wei, University of Maryland

1. INTRODUCTION AND OVERVIEW.

Applied work involving statistical modeling frequently leads to situations where models must be compared which are not related to one another by parameter restrictions. In such a situation, log–likelihood ratios of pairs of estimated models do not have a chi–square limiting distribution, and statisticians making model selection decisions frequently resort to rather complicated and subjective comparisons of residuals or other model artifacts to accomplish the selection. In this paper, we give some theoretical background for the use of the usual log–likelihood ratios for non–nested comparisons. The practical importance of this capability is magnified by the fact that maximized likelihood values are usually available from the software used for estimation. Thus comparisons can often be made quickly. This encourages inventiveness and experimentation by the modeler.

In fact, the model selection procedures we examine are the minimum AIC procedure of Akaike (1973) and related procedures like the minimum BIC procedure of Akaike (1977) and Schwarz (1978) and the criteria of Hannan and Quinn (1979) and Rissanen (1986). The contributions of the paper stem from its rather comprehensive analysis of situations where the models are non–nested and not necessarily correct, and from the mathematical completeness of the results presented for Gaussian situations with fixed regressors or with vector autoregressions and their

subregressions. We also provide a revision of the principle of parsimony away from its oversimplified emphasis on counts of parameters.

Our analysis is restricted to linear regression models estimated via least squares, because more intelligible and complete results can be obtained for these models. The regressors can be stochastic. The associated parameter estimates maximize a Gaussian likelihood function. (The true likelihood function could be non–Gaussian.) Some comments about generalizations to other models are given in section 11.

After introducing some terminology in section 2, we illustrate, in section 3, the use of the minimum AIC procedure with a regressor selection problem which arose in the design of a ship autopilot and which involves both nested and nonnested comparisons. Section 4 provides our basic theoretical assumptions and the measure of the coefficient estimation variability, $CVAR^{(x)}$, associated with a regressor process x_t, which is central to much of the subsequent discussion. In section 5, formulas for $CVAR^{(x)}$ are given which show that this quantity is equal to the number of coefficients estimated when x_t is complete in the sense that it contains the correct regressor as a subvector. Subsection 5.3 shows that $CVAR^{(x)}$ approximates this number when x_t is "almost complete." Subsection 5.1 contain the initial analysis of an important example of two asymptotically equivalent but incomplete autoregressions with the property that the value of $CVAR^{(x)}$ is larger for the model with fewer estimated coefficients, contradicting the principle of parsimony. Here, asymptotically equivalent means that the difference between the estimated regression functions tends to zero in probability as the sample size increases.

In section 6 some easy results are presented describing situations in which a variety of log–likelihood–ratio based model selection criteria prefer one regressor over another with asymptotic probability 1. Some criteria, like BIC, are seen to consistently prefer a model with fewer estimated coefficients whenever the

log–likelihood ratio is bounded in probability, a preference which is sometimes undesirable as the Example 5.1 shows.

The next several sections analyze the asymptotic behavior of the log–likelihood ratio $\hat{L}_N^{(1,2)}$ in the only simply defined situation in which the sequence $\hat{L}_N^{(1,2)}$, $N \geq N_0$ is bounded in probability, the situation of asymptotically equivalent regressors. Section 7 investigates the limiting distribution of the log–likelihood ratio and its connection with $CVAR^{(x)}$ values and with the MAIC criterion and a modification thereof. In subsection 7.1 we comment on the use and limitations of the complete–regressor form of the limiting distribution for hypothesis testing with non–nested regressors.

Section 8 shows that the difference of Kullback–Leibler ("entropy" or "information") numbers of the estimated models overcomes some of the deficiencies of the log–likelihood ratio and motivates the definition of an "ideal" minimum AIC criterion. The generalization (8.11) of a result of Akaike and Shimizu connecting K–L numbers and log–likelihood ratios plays an important role here.

In section 9, we show that under fairly general circumstances, when two regressors are asymptotically equivalent, one can expect the difference of their $CVAR^{(x)}$ values to be the limit of the differences of a normalized measure of the mean square prediction errors arising when the estimated regression coefficients are used to predict an independent replicate of the observations. Thus, there is a predictive interpretation of the results of sections 5, 7 and 8. Section 10 completes this discussion by presenting a strategy for showing that finite sample means of the log–likelihood ratios, of AIC differences, of K–L number differences, and of differences of mean square prediction errors, converge as expected. A new lemma on the rate of decrease of the inverse moments of the Wishart distribution makes it possible to verify the assumptions of section 10 for regressors which are fixed or are subvectors

of not necessarily stationary autoregressions with Gaussian noise processes. For these situations, we thereby achieve the first complete demonstration of the bias correction property used by Akaike (1973) to motivate the definition of AIC.

Section 11 contains comments and literature references concerning generalizations of the results of this paper to models different from linear regression models. The Appendices I and II contain proofs omitted from the initial discussion.

2. LINEAR LEAST SQUARES AND MAIC.

Let y_t be a q–dimensional regressand and x_t an r–dimensional candidate regressor process for y_t satisfying

$$\sum_{t=1}^{N} x_t x_t' > 0 \ , \ N \geq N_0 \tag{2.1}$$

with probability one (w.p.1). The coordinate entries of x_t can be fixed or random.

Although $y_1,\ldots,y_N$ need not be Gaussian, the least squares coefficient and error variance estimates for the regression of y_t on x_t,

$$\hat{A}_N^{(x)} \equiv (\sum_{t=1}^{N} y_t x_t')(\sum_{t=1}^{N} x_t x_t')^{-1} \tag{2.2}$$

and

$$\hat{\Sigma}_N^{(x)} \equiv N^{-1} \sum_{t=1}^{N} (y_t - \hat{A}_N x_t)(y_t - \hat{A}_N x_t)', \tag{2.3}$$

are the maximizers of a Gaussian log–(quasi)likelihood function

$$L_N^{(x)}[\Sigma,A] \equiv -\frac{N}{2} \log 2\pi|\Sigma| - \frac{1}{2} \mathrm{tr}\Sigma^{-1} \sum_{t=1}^{N} (y_t - Ax_t)(y_t - Ax_t)',$$

whose maximum value is

$$\hat{L}_N^{(x)} \equiv L_N^{(x)}[\hat{\Sigma}_N^{(x)}, \hat{A}_N^{(x)}] = -\frac{N}{2}(\log 2\pi|\hat{\Sigma}_N^{(x)}| + q). \tag{2.4}$$

(We use tr to denote trace and $\equiv$ to indicate the definition of a symbol.) When two competing regressor processes $x_t^{(i)}$, i=1,2 are being considered, we will replace the superscript $^{(x)}$ by the superscript $^{(i)}$ in the preceding notation to indicate quantities associated with $x_t^{(i)}$. Our investigation focuses on the log–likelihood ratio,

$$\hat{L}_N^{(1,2)} \equiv \hat{L}_N^{(1)} - \hat{L}_N^{(2)} = -\frac{N}{2}\log(|\hat{\Sigma}_N^{(1)}|/|\hat{\Sigma}_N^{(2)}|), \tag{2.5}$$

and several modifications thereof for regressor comparison purposes, the best known of which is due to Akaike (1973, 1974),

$$AIC_N^{(1,2)} \equiv (-2)\hat{L}_N^{(1,2)} + 2q(r^{(1)} - r^{(2)}), \tag{2.6}$$

with $r^{(i)} \equiv \dim x_t^{(i)}$, i = 1,2.

We will write $x_t^{(1)} \subseteq x_t^{(2)}$ to indicate that $x_t^{(1)} = Bx_t^{(2)}$ for some matrix B. In this case, we will say that $x_t^{(1)}$ is <u>nested</u> in $x_t^{(2)}$. When this happens, if $x_t^{(1)}$ is correct in a strong sense and has certain stability properties (see Lai and Wei (1982) and sections 4 and 5 below), then $(-2)\hat{L}_N^{(1,2)}$ will have an asymptotic $\chi^2(q(r^{(2)} - r^{(1)}))$ distribution. However, in this paper, we are interested in the situation in which the regressors may be non–nested and only approximately correct.

The quantity (2.6) is the difference of the two AIC statistics,

$$AIC_N^{(i)} \equiv -2\hat{L}_N^{(i)} + 2qr^{(i)} \quad (i=1,2)\ . \tag{2.7}$$

Akaike's minimum AIC criterion (MAIC) asserts that the regressor associated with the smaller AIC value should be preferred: thus, $x_t^{(1)}$ is preferred if $AIC_N^{(1,2)} < 0$. The next section presents an application of MAIC in which this criterion exhibits consistent performance across a range of nested and nonnested comparisons.

3. SHIP AUTOPILOT MODELING WITH MAIC: AMERIKA MARU DATA

In Kitagawa and Ohtsu (1976) and Ohtsu et al. (1979) and the papers referenced there, the design and testing of a stochastic–regression–model–based ship autopilot is described. The success of this experiment influenced the design of a new ship (Shoji Maru III) incorporating such an autopilot (K. Ohtsu, personal communication, January 1987). The principal variable to be controlled is yaw (Y), the angular deviation of the ship's forward movement from the intended direction, measured at the bridge. Other less important but useful variables to control include roll (R) and pitch (P). The rudder angle (RU) is the main controller input variable, but measured values of the lateral acceleration (LACC) and vertical acceleration (VACC) of the forepeak may also provide useful information for the controller/autopilot.

Our analysis will seek to determine the situations in which VACC is a useful controller input variable for a specific ship: we consider the problem of choosing between the regressors $x_t^{(m)}$ and $\tilde{x}_t^{(M)}$, these being defined by

$$x_t^{(m)} \equiv (Y_{t-1}, R_{t-1}, P_{t-1}, RU_{t-1}, LACC_{t-1}, \ldots Y_{t-m}, R_{t-m}, P_{t-m}, RU_{t-m}, LACC_{t-m})'$$

and

$$\tilde{x}_t^{(M)} \equiv (x_t^{(M)'}, \text{VACC}_{t-1}, \ldots, \text{VACC}_{t-M})',$$

for $1 \leq m, M \leq 10$. If $M < m$, these regressors are non–nested. The modeling will be done with N=894 observations made at 1 second intervals on the container ship Amerika Maru under manual control. These data are discussed in the papers cited above. If $\hat{m}$ and $\hat{M}$ denote the lags associated with minimum AIC values for the regressors $x_t^{(m)}$, $1 \leq m \leq 10$ and $\tilde{x}_t^{(M)}$, $1 \leq M \leq 10$, respectively, then the use of VACC in the autopilot model seems worth considering seriously when

$$\text{DAIC}_{894} \equiv \text{AIC}_{894}^{(\hat{M})} - \text{AIC}_{894}^{(\hat{m})}$$

is negative. Results obtained from the program MULCON of Akaike et al. (1985) for seven choices of the regressand y_t are included in Table 3.1 below. The choices for y_t are: Y_t, R_t, P_t, $(Y_t, R_t)'$, $(Y_t, P_t)'$, $(R_t, P_t)'$ and $(Y_t, R_t, P_t)'$. In the table, LAG denotes $\hat{M}$ or $\hat{m}$, as appropriate, and

$$\Delta\text{dimA} \equiv q(6\hat{M} - 5\hat{m}),$$

with $q = \text{dimy}_t$. The results are consistent: the use of VACC is favored only when P is one of the controlled variables. This conclusion has engineering plausibility: VACC is closely related to P but not to the other controlled variables. Thus, MAIC has functioned quite satisfactorily. Note also that in the two cases, $y_t = Y_t$ and $y_t = P_t$, the comparison is between non–nested regressors, since $\hat{M} < \hat{m}$. Hypothesis testing based on an asymptotic distribution for $L_N^{(1,2)}$ leads to the same conclusions, but this approach has some significant limitations, see subsection 7.1.

Table 3.1. DAIC Values for Various Choices of y

	With VACC			Without VACC			Difference	
y	LAG	dimA	AIC_{894}	LAG	dimA	AIC_{894}	ΔdimA	$DAIC_{894}$
Y,R,P	8	108	18683.	6	90	18693.	18	−10.
Y,R	6	72	12244.	6	60	12236.	12	8.
Y,P	7	84	12468.	7	70	12483.	14	−15.
R,P	8	96	13033.	8	80	13036.	16	−3.
Y	4	24	6027.	6	30	6025.	−6	2.
R	5	30	6574.	5	25	6569.	5	5.
P	7	42	6393.	8	40	6491.	2	−14.

An additional analysis with MAIC to determine which components' error processes are uncorrelated is discussed in Findley (1988).

4. BASIC ASSUMPTIONS

The fundamental issues we wish to discuss can be described in the context of selecting between two competing regressor processes $x_t^{(i)}$, i=1,2. Our minimal assumption beyond (2.1) is that the estimated error variance matrices $\hat{\Sigma}_N^{(i)}$ converge in probability to positive definite limits, $\hat{\Sigma}_N^{(i)} \xrightarrow[N]{p} \Sigma^{(i)} > 0$, so that the log–likelihood ratio satisfies

$$N^{-1}\hat{L}_N^{(1,2)} \xrightarrow[N]{p} -(1/2)\log(|\Sigma^{(1)}|/|\Sigma^{(2)}|). \tag{4.1}$$

In what follows, x_t usually designates either of the regressors $x_t^{(i)}$, i=1,2. We will assume that a matrix $A^{(x)}$ exists such that $\hat{A}_N^{(x)} \xrightarrow[N]{p} A^{(x)}$ holds. Defining $e_t^{(x)} \equiv y_t - A^{(x)}x_t$, we will call the equation

$$y_t = A^{(x)}x_t + e_t^{(x)}$$

the model associated with the regressor process x_t. We note that

$$\hat{A}_N^{(x)} - A^{(x)} = \sum_{t=1}^{N} e_t^{(x)} x_t' \Big(\sum_{t=1}^{N} x_t x_t' \Big)^{-1} , \tag{4.2}$$

and that $\hat{\Sigma}_N^{(x)}$ differs from $\Sigma_N^{(x)} \equiv N^{-1}\Sigma_{t=1}^{N} e_t^{(x)} e_t^{(x)'}$ by the quantity

$$\Sigma_N^{(x)} - \hat{\Sigma}_N^{(x)} = N^{-1}(\hat{A}_N^{(x)} - A^{(x)}) \sum_{t=1}^{N} x_t x_t' (A_N^{(x)} - A^{(x)})' . \tag{4.3}$$

We will now introduce a measure of model uncertainty (or variability) due to parameter estimation which is invariant under "scale" transformations $y_t \rightarrow By_t$, $x_t \rightarrow Cx_t$ with nonsingular B and C . In the situation of interest in sections 5–7, where the regressors are asymptotically equivalent, the effects of estimating $\hat{\Sigma}_N^{(x)}$ are the same for both regressors and cancel in the log–likelihood ratio $\hat{L}_N^{(1,2)}$, see Proposition 6.3. Our measure will therefore focus on the coefficient estimates. We define

$$Q_N^{(x)} \equiv \mathrm{tr}(\Sigma^{(x)})^{-1}(\hat{A}_N^{(x)} - A^{(x)}) \Big(\sum_{t=1}^{N} x_t x_t' \Big) (\hat{A}_N^{(x)} - A^{(x)})' . \tag{4.4}$$

For purposes of interpretation, we note that this reduces to the total squared estimation error of the coefficients, $\mathrm{tr}(\hat{A}_N^{(\tilde{x})} - A^{(\tilde{x})})(\hat{A}_N^{(\tilde{x})} - A^{(\tilde{x})})'$, if the y_t and x_t are transformed in such a way that $\Sigma^{(e)} = I_q$ and $\Sigma_{t=1}^{N} \tilde{x}_t \tilde{x}_t' = I_r$. Using (4.3), we could also write $Q_N^{(x)} = N\mathrm{tr}(\Sigma^{(x)})^{-1}(\hat{\Sigma}_N^{(x)} - \Sigma_N^{(x)})$ and observe that, since $\hat{\Sigma}_N^{(x)} \xrightarrow[p]{N} \Sigma^{(x)}$, the variate $Q_N^{(x)}$ is asymptotically equivalent to the final term of the decomposition $\hat{L}_N^{(x)} = -\frac{N}{2}\{\log 2\pi |\hat{\Sigma}_N^{(x)}| + \mathrm{tr}(\hat{\Sigma}_N^{(x)})^{-1}\Sigma_N\} + \mathrm{tr}(\hat{\Sigma}_N^{(x)})^{-1}(\Sigma_N -$

$\hat{\Sigma}_N^{(x)}$) obtained from (2.4) via the substitution $q = tr(\hat{\Sigma}_N^{(x)})^{-1}\hat{\Sigma}_N^{(x)}$. We shall assume that

(A1) $Q_N^{(x)} \xrightarrow[\text{dist.}]{N} Q^{(x)}$, and $EQ^{(x)} < \infty$.

Our measure of (asymptotic) model uncertainty due to coefficient estimation is defined to be

$$CVAR^{(x)} \equiv EQ^{(x)} . \tag{4.5}$$

Explicit formulas for $CVAR^{(x)}$ will be given in the next section. Its connection with the MAIC procedure will be revealed in sections 5, 8 and 11. In section 9 an alternative measure is described which provides a connection between model uncertainty and prediction error.

Our usual method of verifying (A1) will involve establishing that there is a vector variate $t_n^{(x)}$ satisfying

$$Q_N^{(x)} = t_N^{(x)\prime} t_N^{(x)}, \tag{4.6}$$

which has a limiting distribution with finite mean and variance. To define this variate, we need some notation. Given a positive definite matrix Σ, we will use $\Sigma^{1/2}$ to denote any matrix S with the property that $\Sigma = SS'$, providing it is formed continuously, meaning that $\Sigma_N \xrightarrow{N} \Sigma$ implies $\Sigma_N^{1/2} \xrightarrow{N} \Sigma^{1/2}$. The Cholesky factorization is an example. For a matrix explicitly of the form $C\Sigma C'$, the square root of choice will be $C\Sigma^{1/2}$. We will denote the inverse of $\Sigma^{1/2}$ by $\Sigma^{-1/2}$, this being different from $(\Sigma^{-1})^{1/2} = \{(\Sigma^{1/2})'\}^{-1}$ in general. We define

$$t_N^{(x)} \equiv \text{vec}\left[(\Sigma^{(x)})^{-1/2} \sum_{t=1}^{N} e_t^{(x)} x_t' \left\{ \left(\sum_{t=1}^{N} x_t x_t' \right)^{-1} \right\}^{1/2} \right], \qquad (4.7)$$

where vec[·] denotes the column vector obtained by stacking the columns of the matrix [·]. This satisfies (4.6). It is easy to check that the variates defined by (4.7) are invariant under nonsingular linear transformations of x_t or y_t.

It follows from (A1) that Q_N is bounded in probability ($0_p(1)$). Thus the term on the right in (4.3) converges to 0 in probability, with the result that $\Sigma_N^{(x)} \xrightarrow[p]{N} \Sigma^{(x)}$.

We will occasionally need to assume

(A2) $N^{1/2}(\hat{\Sigma}_N^{(x)} - \Sigma^{(x)})$ is bounded in probability.

This condition is satisfied when $N^{1/2}(\hat{\Sigma}_N^{(x)} - \Sigma^{(x)})$ has a limiting distribution. Two further simplifying assumptions sometimes called upon are

(A3) $Ee_t^{(x)} x_t' = 0$,

(A4) $Ee_t^{(x)} e_t^{(x)} = \Sigma^{(x)}$.

Note that, for <u>non-stochastic</u> regressors, (A3) is equivalent to $Ey_t = A^{(x)} x_t$, meaning that x_t has been chosen well enough to capture the mean behavior of y_t. Shibata (1981) presents results for fixed regressors when (A3) fails, for the case in which $y_t - Ey_t$ is i.i.d. and Gaussian. We discuss his results briefly in section 11.

To obtain formulas for $CVAR^{(x)}$, we also require

$$\text{(A5)} \qquad \left\{E\left(\sum_{t=1}^{N} x_t x_t'\right)\right\}^{-1/2}\left\{\sum_{t=1}^{N} x_t x_t'\right\}^{1/2} \xrightarrow[N]{p} I_r \ .$$

Here I_r denotes the identify matrix of order r≡dim x_t. (A5) is satisfied, for example, if x_t is nonstochastic, or if x_t is stationary and $N^{-1}\Sigma_{t=1}^{N} x_t x_t' \xrightarrow[N]{p} \Gamma^{(x)} \equiv Ex_t x_t'$. It implies that the difference between $t_N^{(x)}$

and

$$Z_N^{(x)} \equiv \text{vec}\left[(\Sigma^{(x)})^{-1/2}\left(\sum_{t=1}^{N} e_t^{(x)} x_t'\right)\left\{\left(E\sum_{t=1}^{N} x_t x_t'\right)^{-1}\right\}^{1/2}\right] \qquad (4.8)$$

tends to zero in probability, a situation we denote by

$$t_N^{(x)} \sim_p Z_N^{(x)}. \qquad (4.9)$$

Hence, under (A5), also, $Z_N^{(x)\prime} Z_N^{(x)} \xrightarrow[N]{\text{dist.}} Q^{(x)}$, the limiting distribution in (A1).

5. FORMULAS FOR $CVAR^{(x)}$ WHEN THE LIMITING DISTRIBUTION IS GAUSSIAN.

We will present formulas in subsection 5.1 for the situation in which the limiting distribution of $t_N^{(x)}$ in (4.7) is Gaussian with mean zero, and the regressor process x_t is not complete, meaning that neither x_t nor any subvector is a correct regressor in the sense of subsection 5.2. These results require the joint stationarity

of x_t and y_t (or e_t), but stationarity is not needed for the familiar formula, $CVAR^{(x)} = qr$ (=dim $A^{(x)}$), obtained in subsection 5.2 for complete regressors. We will assume throughout this section that (A3) – (A5) hold and will refer to theorems in the literature verifying (A1).

5.1. Stationary Case.

Making joint stationarity assumptions for x_t, y_t, we define $\Gamma^{(x)} \equiv Ex_t x_t'$ and note that, since $Ee_t^{(x)} x_t' = 0$, the random vector

$$Z_N^{(x)} = N^{-1/2} \mathrm{vec}\left[\sum_{t=1}^{N} \left\{(\Sigma^{(x)})^{-1/2} e_t^{(x)}\right\}\left\{(\Gamma^{(x)})^{-1/2} x_t\right\}'\right] \tag{5.1}$$

is $N^{1/2}$ times the sample mean of the mean zero stationary vector process

$$M_t^{(x)} = \mathrm{vec}\left[\left\{(\Sigma^{(x)})^{-1/2} e_t^{(x)}\right\}\left\{(\Gamma^{(x)})^{-1/2} x_t\right\}'\right]. \tag{5.2}$$

Therefore, a variety of Central Limit Theorem results apply to (5.1), see Theorem 5.2 of Brillinger (1969), Hannan (1970, pp. 220–228), Corollary (3.9) of McLeish (1975), Dahlhaus (1985) and Eberlein (1986). Under diverse assumptions, these results yield

$$Z_N^{(x)} \xrightarrow[\text{dist.}]{N} \mathcal{N}(0,V) \quad , \tag{5.3}$$

with

$$CVAR^{(x)} = \mathrm{tr}V = \lim_{N\to\infty} EZ_N^{(x)\prime} Z_N^{(x)}$$

$$= \sum_{k=-\infty}^{\infty} E\left\{\left[e_t^{(x)\prime}\Sigma^{(x)-1}e_{t+k}^{(x)}\right]\left[x_{t+k}'\Gamma^{(x)-1}x_t\right]\right\}, \qquad (5.4),$$

or an equivalent expression involving integrals of cumulant spectral densities, see Brillinger (1969) and (5.6) below.

The fourth cumulants associated with the fourth moment quantities in (5.4) vanish when e_t and x_t are jointly Gaussian and also in the not necessarily Gaussian univariate autogression situation, where y_t is scalar with mean zero and $x_t = [y_{t-m_1} \cdots y_{t-m_r}]'$ for positive integers $m_1 < m_2 < \cdots < m_r$, see Remark 3.2 of Hosoya and Taniguchi (1982, p. 138). In these cases, Isserlis' formula (Brillinger, 1975, p. 21) and (A3) can be used to show that (5.4) reduces to

$$CVAR_G^{(x)} = \sum_{k=-\infty}^{\infty} tr\{\Sigma^{(x)-1}\Gamma^{(e)}(k)\}tr\{\Gamma^{(x)-1}\Gamma^{(x)}(k)\}$$

$$+ \sum_{k=-\infty}^{\infty} tr\{\tilde{\Gamma}^{ex}(k)\tilde{\Gamma}^{ex}(-k)'\}, \qquad (5.5)$$

where $\Gamma^{(x)}(k) \equiv Ex_t x_{t+k}'$, $\Gamma^{(e)}(k) \equiv Ee_t^{(x)}e_{t+k}^{(x)\prime}$, and

$$\tilde{\Gamma}^{ex}(k) \equiv E[\{(\Sigma^{(x)})^{-1/2}e_t^{(x)}\}\{(\Gamma^{(x)})^{-1/2}x_{t+k}\}'] .$$

A convenient spectral density form of (5.5) follows via Parseval's formula:

$$CVAR_G^{(x)} = 2\pi\int_{-\pi}^{\pi} tr\left\{\Sigma^{(e)-1}f^{ee}(\lambda)\right\}tr\left\{\Gamma^{(x)-1}f^{xx}(\lambda)\right\}d\lambda$$

$$+ 2\pi\int_{-\pi}^{\pi} \mathrm{tr}\left\{\Sigma^{(x)-1} f^{ex}(\lambda)\Gamma^{(x)-1} f^{xe}(-\lambda)\right\} d\lambda, \tag{5.6}$$

where $f^{ee}(\lambda)$ and $f^{xx}(\lambda)$ are the spectral density matrices of e_t and x_t, respectively, $f^{ex}(\lambda) = (2\pi)^{-1}\Sigma_{k=-\infty}^{\infty} \tilde{\Gamma}^{ex}(k)e^{-ik\lambda}$, and $f^{xe}(\lambda) = (2\pi)^{-1}\Sigma_{k=-\infty}^{\infty} \tilde{\Gamma}^{ex}(-k)'e^{-ik\lambda}$.

For scalar processes y_t, set $\rho_k \equiv Ey_t y_{t+k}/Ey_t^2$. If

$$\rho_m = \rho_{m_1} = \rho_{m_2} = \rho_{m_1 - m_2} = 0, \tag{5.7}$$

for distinct lags m, m_1, m_2, two autoregressions, with $x_t^{(1)} \equiv y_{t-m}$ and $x_t^{(2)} \equiv [y_{t-m_1}\ y_{t-m_2}]'$, will turn out to be of special interest. The condition (5.7) implies that $A^{(1)} = 0$ and $A^{(2)} = 0$, so that $e_t^{(1)} = e_t^{(2)} = y_t$. It follows from (5.5) that if $V_n \equiv \sum_{k=-\infty}^{\infty}\{\rho_k^2 + \rho_{k+n}\rho_{k-n}\}$, then

$$\mathrm{CVAR}^{(1)} = V_m, \tag{5.8}$$

and

$$\mathrm{CVAR}^{(2)} = V_{m_1} + V_{m_2}. \tag{5.9}$$

Example 5.1. Suppose y_t is a stationary autoregressive process of order 6 with variance 1 whose first six partial autocorrelations are 0.0, 0.0, 0.0, .80, −.41, −.64. The Levinson–Durbin algorithm (Box and Jenkins (1976, p. 83) can be used to

calculate the autoregressive coefficients and the autocorrelations ρ_k. The vanishing of the first three partial autocorrelations is equivalent to

$$\rho_1 = \rho_2 = \rho_3 = 0 \quad . \tag{5.10}$$

With $x_t^{(1)} = y_{t-2}$ and $x_t^{(2)} = [y_{t-1}\ y_{t-3}]'$, the formulas (5.8) and (5.9) yield

$$\begin{aligned} &\text{CVAR}^{(1)} = 26.3 \\ \text{and} \qquad &\text{CVAR}^{(2)} = 2.9 + 2.4 = 5.3 \quad . \end{aligned} \tag{5.11}$$

Thus, although the regressors $x_t^{(1)}$ and $x_t^{(2)}$ are asymptotically equivalent in the sense that $e_t^{(1)} = e_t^{(2)}$, the more parsimonious regressor $x_t^{(1)}$ has greater coefficient estimation variability as measured by CVAR. In fact, for 577 out of 1000 models obtained under (5.7) by choosing the partial autocorrelations at lags 4–6 uniformly and independently, it happened that the regression on one of y_{t-1}, y_{t-2}, y_{t-3} had a larger value of CVAR than the regression on the remaining pair of lagged y–variates. Some implications of this will be discussed in sections 7–9. This phenomenon does not occur when both regressors are complete in the sense we will now describe.

5.2. Complete Regressors.

Let I_t denote an information set (σ–algebra) containing the information generated by the "past history up to t" of all regressors under consideration, $I_t \supseteq \sigma(x_t^{(1)},\ldots,x_1^{(1)};\ x_t^{(2)},\ldots,x_1^{(2)};\ y_{t-1},\ldots,y_0;\ldots)$. It would be natural to say that a regressor x_t which is determined by I_t (that is, is I_t–measurable) for each $t = 1, 2,\ldots$, is correct if for some matrix $A^{(x)}$, all of whose columns are non–zero, we have $E(y_t|I_t) = A^{(x)}x_t$, or, equivalently, with $e_t^{(x)} = y_t - A^{(x)}x_t$, if

$$E(e_t^{(x)}|I_t) = 0 \tag{5.12}$$

holds. However, the case when some columns of $A^{(x)}$ are 0 needs to be considered and additional conditions, such as $\sup_t E\{e_t^{(x)\prime} e_t^{(x)}\}^{1+\epsilon} < \infty$ for some $\epsilon > 0$, need to be imposed to obtain the expected limiting distribution for $t_N^{(x)}$. We shall say that x_t is a complete regressor (process) for y_t if, in addition to (5.12) and

$$E(e_t^{(x)} e_t^{(x)\prime}|I_t) = \Sigma^{(e)}, \tag{5.13}$$

two other conditions hold,

$$\left\{E \sum_{u=1}^{N} x_u x_u'\right\}^{-1/2} x_t \xrightarrow[p]{N} 0 \quad , \tag{5.14}$$

and (A5). It follows then, from a multivariate generalization of Theorem 3 of Lai and Wei (1982), that

$$t_N^{(x)},\ Z_N^{(x)} \xrightarrow[\text{dist.}]{N} \mathcal{N}(0,\ I_{qr}) \ . \tag{5.15}$$

Hence, for complete regressors,

$$CVAR^{(x)} = qr \ . \tag{5.16}$$

Any two complete regressors are asymptotically equivalent, since $A^{(1)}x_t^{(1)} = E(y_t|I_t) = A^{(2)}x_t^{(2)}$. Thus, if we regard $CVAR^{(x)}$ as a cost function, then (5.16) embodies the principle of parsimony ("the fewer coefficients estimated the

better") for complete regressors. The example of subsection 5.1 shows that when asymptotically equivalent but incomplete regressors are considered, the principle of parsimony is no longer valid: the more parsimonious regressor can have greater cost.

For the stationary case, (5.16) follows from (5.4), (5.12) and (5.13): using the formula $E(\cdot) = E\{E(\cdot|I_t)\}$, one sees immediately from (5.12) that the terms in (5.4) with $k \neq 0$ are 0, and, from (5.13) one then obtains

$$CVAR^{(x)} = tr\left\{\Sigma^{(x)-1}Ee_t^{(x)}e_t^{(x)\prime}\right\}\cdot tr\left\{\Gamma^{(x)-1}Ex_tx_t'\right\} = qr \ .$$

In the next subsection, we shall describe some continuity properties of $CVAR^{(x)}$. These imply that (5.16) holds approximately if x_t is "almost complete." We will also give a simple example to show that, although it can be weakened as in Lai and Wei (1982), a condition like (5.13) cannot be completely dispensed with.

5.3. Continuity of $CVAR^{(x)}$ near Complete Regressors for Stationary Regressions.

Let us consider $CVAR_G^{(x)}$ first. If x_t is complete, then, by (5.12) and (5.13), $f^{xe}(\lambda) = f_-^{xe}(\lambda) \equiv (2\pi)^{-1}\Sigma_{k=-\infty}^{-1} \tilde{\Gamma}^{ex}(-k)'e^{-ik\lambda}$ and $(\Sigma^{(x)})^{-1}f^{ee}(\lambda) = (2\pi)^{-1}I_q$. Clearly, $CVAR_G^{(x)}$ will be close to qr if $f^{xe}(\lambda) - f_-^{xe}(\lambda)$ and $f^{ee}(\lambda) - (2\pi)^{-1}\Sigma^{(x)}$ are close to zero in any of a variety of senses. For example, if the entries of the spectral density matrices in (5.6) are square integrable over 2π, one can obtain such a result from the fact that the left–hand side of

$$\int_0^{2\pi}|g(\lambda)h(\lambda)|d\lambda \leq \left\{\int_0^{2\pi}|g(\lambda)|^2d\lambda\right\}^{1/2}\left\{\int_0^{2\pi}|h(\lambda)|^2d\lambda\right\}^{1/2}$$

will be small if the integral of $|g(\lambda)|^2$ is small enough.

We will now indicate how, if fourth moments exist, the Cauchy–Schwarz inequality for expectations can be applied to obtain an analogous result for $CVAR^{(x)}$ via (5.4). For a random h–vector $w = [w_1 \cdots w_h]'$, we will use $||w||_2$ to denote $\max_{1 \le j \le h} \{Ew_j^2\}^{1/2}$.

With $M_t^{(x)}$ as in (5.2) and $c_2 \equiv ||M_t^{(x)}||_2$, we will first show, following an approach suggested by Madga Peligrad, that, for all N,

$$|EZ_N^{(x)\prime} Z_N^{(x)} - EM_t^{(x)\prime} M_t^{(x)}| \le 2c_1c_2 \tag{5.17}$$

where c_1 is a measure of the t–dependencies among the entries m_t of $M_t^{(x)}$, $t = 1,2,...$, which is described below. Observe that if (5.12) holds, then for each entry m_t, the quantity

$$\Delta_N = N^{-1}E\{(\Sigma_{t=1}^N m_t)^2 - \Sigma_{t=1}^N m_t^2\}$$

is zero, as is also the expected value of $S_n(p) \equiv \Sigma_{t=n+1}^{n+p} m_t$ conditional on I_{n+1}.

One verifies as in Eberlein (1986) that

$$\sup_{n,p} ||E(S_n(p)|I_{n+1})||_2 \le c_1 \tag{5.18}$$

where c_1 is the maximum over the components of $M_t^{(x)}$ of the sum of the mixingale coefficients as defined in McLeish (1975). Since $\Delta_N = (2/N)\Sigma_{t=1}^N E\{m_t S_t(N-t)\}$, and since $|E\{m_t E(S_t(N-t)|I_{t+1})\}| \le c_1c_2$ by Cauchy–Schwarz, (5.17) follows.

The left hand side of (5.17) will be small if c_1 is small enough. We will complete our examination of $CVAR^{(x)} - qr$ by showing that $\delta \equiv EM_t^{(x)\prime} M_t^{(x)} - qr$ is neglible if $||E(e_t^{(x)\prime} \Sigma^{(x)-1} e_t^{(x)}|I_t) - q||_2$ is small enough. Noting that

$Ee_t^{(x)\prime}\Sigma^{(x)-1}e_t^{(x)} = q$ and $Ex_t'\Gamma^{(x)-1}x_t = r$, this assertion follows from the identity $\delta = E[x_t'\Gamma^{(x)-1}x_t\{E(e_t^{(x)\prime}\Sigma^{(x)-1}e_t^{(x)}|I_t) - q\}]$ via the Cauchy–Schwarz inequality.

This last argument is clearly related to (5.13). We close this section with an elementary stationary example for which (A1)–(A5), (5.3), (5.12) and (5.14) hold, but not (5.13), and for which (5.16) does not hold, because the asymptotic variance matrix of $t_N^{(x)}$ and $Z_N^{(x)}$ is different from the identity matrix indicated in (5.15). The basic construction is due to Andrew Siegel (personal communication, March 1987). Let F be the distribution on the eight number pairs $\pm(\sqrt{3/2},\sqrt{1/2})$, $\pm(\sqrt{3/2}, -\sqrt{1/2})$, $\pm(\sqrt{1/2}, \sqrt{3/2})$, $\pm(\sqrt{1/2}, -\sqrt{3/2})$, which assigns probability 1/8 to each pair. Let (x_t, e_t), $t = 0,1,...$ be an i.i.d. sequence with distribution F. If $y_t = ax_t + e_t$ for some a, we have a regression with q=r=1 and $e_t^{(x)} = e_t$. Also, $Ee_t^2 = Ex_t^2 = 1$, and $Ex_t = Ee_t = Ee_tx_t = 0$. If $I_t \equiv \sigma(x_1,...,x_t,y_0,...,y_{t-1})$, then $E(e_t|I_t) = Ee_t = 0$, but $E(e_t^2|x_t^2 = 3/2) = 1/2$, whereas $E(e_t^2|x_t^2 = 1/2) = 3/2$, so (5.13) fails. Finally, (5.3) holds, with $V = Ee_t^2x_t^2 = 3/4$, so that $CVAR^{(x)} = 3/4$. Thus $CVAR^{(x)} \neq qr(=1)$ in contrast to (5.16).

6. SOME REGRESSOR SELECTION CRITERIA AND THEIR CONSISTENCY PROPERTIES.

To obtain a broader perspective on MAIC and the role of CVAR, we now consider additional adjusted log–likelihood ratios,

$$D_N^{(1,2)}[c_N^{(1,2)}] = -2\hat{L}_N^{(1,2)} + c_N^{(1,2)}, \qquad (6.1)$$

and their allied criteria, according to which the regressor process $x_t^{(1)}$ is preferred if $D_N^{(1,2)}[c_N^{(1,2)}] < 0$. For fixed regressors, all such criteria are admissible in the

decision–theoretic sense, see Takada (1982). Some examples of $c_N^{(1,2)}$ and the names of their associated criteria are given in (6.2):

$$c_N^{(1,2)} = \begin{cases} 2q(r^{(1)} - r^{(2)}); & AIC_N^{(1,2)} \quad (Akaike(1973)) \\ 2(CVAR^{(1)} - CVAR^{(2)}); & Ideal\text{–}AIC_N^{(1,2)} \quad (Section\ 8) \\ q(r^{(1)} - r^{(2)}) \log N; & BIC_N^{(1,2)} \quad (Akaike(1977), Schwarz(1978), Rissanen(1986)) \\ 2q(r^{(1)} - r^{(2)}) \log\log N; & Hannan\ and\ Quinn(1979) \end{cases} \tag{6.2}$$

The minimum Ideal–AIC criterion is not implementable because the quantities $CVAR^{(i)}$, $i = 1,2$ are unknown. The following proposition, an immediate consequence of (4.1), shows that all the criteria of (6.2) consistently prefer $x_t^{(1)}$ if it provides a better fit asymptotically, in the sense that $|\Sigma^{(1)}| < |\Sigma^{(2)}|$.

Proposition 6.1. If $|\Sigma^{(1)}| < |\Sigma^{(2)}|$ and if $c_N^{(1,2)}/N \xrightarrow{N} 0$, then $P(D_N^{(1,2)}[c_N^{(1,2)}] < 0) \xrightarrow{N} 1$.

For example, if $x_t^{(1)}$ is a complete regressor process, as defined in section 5, and $x_t^{(2)}$ is not complete, then $A^{(1)}x_t^{(1)} - A^{(2)}x_t^{(2)} \neq 0$, but $Ee_t^{(1)}(A^{(1)}x_t^{(1)} - A^{(2)}x_t^{(2)})' = 0$, by (5.12). Hence (A4) and the decomposition $e_t^{(2)} = e_t^{(1)} + \{A^{(1)}x_t^{(1)} - A^{(2)}x_t^{(2)}\}$ yield $\Sigma^{(1)} \leq \Sigma^{(2)}$ and $\Sigma^{(1)} \neq \Sigma^{(2)}$, from which it follows that $|\Sigma^{(1)}| < |\Sigma^{(2)}|$. Proposition 6.1 shows, therefore, in particular, that the criteria defined by (6.1) and (6.2) consistently prefer a complete regressor over an incomplete regressor. The Cox tests discussed briefly in subsection 7.1 are intended to provide a more traditional model selection approach for this situation.

Another immediate result applies to the BIC and Hannan and Quinn criteria and also to the criteria of Rissanen (1986). We will call a criterion of the form (6.1) strongly parsimonious if $c_N^{(1,2)} \xrightarrow{N} -\infty$ whenever $r^{(1)} < r^{(2)}$.

Proposition 6.2. If $r^{(1)} < r^{(2)}$, and if $D_N^{(1,2)}[c_N^{(1,2)}]$ is strongly parsimonious, then $P(D_N^{(1,2)}[c_N^{(1,2)}] < 0) \xrightarrow{N} 1$ whenever the log-likelihood ratio $L_N^{(1,2)}$ is bounded in probability.

The next result shows that Proposition 6.2 applies to asymptotically equivalent regressors. Its proof is given in Appendix I.

Proposition 6.3. Under assumption (A2), if the regressor processes $x_t^{(1)}$ and $x_t^{(2)}$ are asymptotically equivalent in the sense that their error processes coincide, $e_t^{(1)} = e_t^{(2)}$ (w. p. 1), then

$$(-2)L_N^{(1,2)} \sim_p Q_N^{(2)} - Q_N^{(1)}. \tag{6.3}$$

Therefore, if (A1) holds for both regressors, as well as (A2), then $\hat{L}_N^{(1,2)}$ is bounded in probability.

These last two propositions show that when asymptotically equivalent regressors are being compared (and (A1) and (A2) hold), then the minimum BIC and Hannan–Quinn criteria, among others, consistently prefer the regressor with smaller dimension (fewer estimated coefficients). Example 5.1 reveals that this preference can be undesirable. The deep results of Shibata(1980,1981) also show that the strongly parsimonious criteria can perform poorly relative to MAIC when the regressors are

not complete. Our Example 5.1 is simpler and more accessible than Shibata's results. It too has implications for prediction, see section 9.

The discussion after Proposition 7.3 below shows that, in typical situations involving asymptotically equivalent regressors, each regressor has a non–zero probability of selection by MAIC, so this criterion does not have a consistency property.

A theoretical prototype for $AIC_N^{(1,2)}$ and Ideal–$AIC_N^{(1,2)}$ which has a more focused consistency property than the strongly parsimonions criteria is investigated in section 8.

Except in Corollary 7.3, we will not establish any further theoretical properties of the strongly parsimonious criteria. These criteria are as easily calculated as AIC's and, for certain applications, might be preferable on the basis of experiments and subject–matter considerations, see Franke et al.(1985). If the model selection need not be done automatically, most users will want to examine several criteria.

7. LIMITING DISTRIBUTIONS OF $\hat{L}_N^{(1,2)}$ AND $AIC_N^{(1,2)}$ FOR ASYMPTOTICALLY EQUIVALENT REGRESSORS.

We would like to conclude from (6.3) that the asymptotic mean of $(-2)\hat{L}_N^{(1,2)}$ is $CVAR^{(2)} - CVAR^{(1)}$. In section 10, we will obtain this quantity as the limit of the finite–sample means $E\{-2\hat{L}_N^{(1,2)}\}$. Here we ignore the finite–sample means and invoke less restrictive assumptions than those of section 10. Our goal is to obtain the existance of a random variable $Q^{(1,2)}$ such that (7.1) holds,

$$(-2)\hat{L}_N^{(1,2)} \xrightarrow[\text{dist.}]{N} Q^{(1,2)}, \text{ with } EQ^{(1,2)} = CVAR^{(2)} - CVAR^{(1)}. \quad (7.1)$$

This property will follow from the additive property of means if we can show that $Q_N^{(1)}$ and $Q_N^{(2)}$ (or $t_N^{(1)}$ and $t_N^{(2)}$) have a joint asymptotic distribution, which we will do for a rather broad class of regressors in Proposition 7.1.

The mean formula in (7.1) reveals that, with asymptotically equivalent regressors, $\hat{L}_N^{(1,2)}$ will have an asymptotic tendency to be positive if $x_t^{(1)}$ is the less desirable regressor (that is, if $CVAR^{(1)} > CVAR^{(2)}$). This is the opposite of what happens if $|\Sigma^{(1)}| \neq |\Sigma^{(2)}|$, when, according to Proposition 6.1, a positive tendency of $\hat{L}_N^{(1,2)}$ means that $|\Sigma^{(1)}| < |\Sigma^{(2)}|$, so that $x_t^{(1)}$ is the better regressor. It is this dichotomous behavior for which a log–likelihood–ratio–based regressor selection procedure must compensate.

We pointed out in section 4 that the t_N–variates are invariant under non–singular transformations of the regressor processes. We will make frequent use of this property now. Throughout this section, the coinciding regression error processes $e_t^{(1)}$ and $e_t^{(2)}$ will be denoted by e_t.

The simplest path to (7.1) uses the familiar device of isolating the effect of the factor $(\Sigma_{t=1}^N x_t^{(i)} x_t^{(i)})^{-1}$ in $t_N^{(i)}$ by suitably normalizing $x_t^{(i)}$. We assume that there exist nonsingular matrices $C_N^{(i)}$, $N \geq N_0$ such that the transformed variates $x_{t,N}^{(i)} = C_N^{(i)-1} x_t^{(i)}$ satisfy

$$\sum_{t=1}^{N} x_{t,N}^{(i)} x_{t,N}^{(i)\prime} \longrightarrow_p V^{(i)} > 0, \tag{7.2}$$

with $V^{(i)}$ a <u>non–stochatistic</u>, positive definite matrix, for $i = 1, 2$, and such that an appropriate joint limiting distribution exists for the fundamental sums,

$$\text{vec}\left[\sum_{t=1}^{N} e_t [x_{t,N}^{(1)\prime} \; x_{t,N}^{(2)\prime}]\right] \longrightarrow_{\text{dist.}} U, \text{ with } EU'U < \infty. \tag{7.3}$$

(In the most familiar cases, $C_N^{(i)} = N^{1/2}I_r(i)$). Using the property vec[BC′] = (C⊗I)vecB, it follows from (7.2) and (7.3) that the joint distribution of $t_N^{(1)}$ and $t_N^{(2)}$ is obtained by left multiplying U by the block diagonal matrix

$$\begin{bmatrix} (V^{(1)})^{-1/2}\otimes I_r(1) & 0 \\ 0 & (V^{(2)})^{-1/2}\otimes I_r(2) \end{bmatrix}.$$

Now, from Proposition (6.3), we obtain

Proposition 7.1. Under assumption (A2) of section 4, if (7.2) and (7.3) hold, so does (7.1).

The approach taken above to (7.1) obscures the role of components common to $x_t^{(1)}$ and $x_t^{(2)}$: if $A^{(1)}x_t^{(1)} = A^{(2)}x_t^{(2)}$ is non–zero, the distribution of U in (7.3) will be singular (have a singular variance matrix, etc.). It also does not make clear the form of the asymptotic distribution of $-2\hat{L}_N^{(1,2)}$; this can have a rather simple form that does not depend on the nature of common components, as we shall demonstrate in Proposition 7.3. To motivate this result, we consider the simplifications which occur when $x_t^{(1)}$ and $x_t^{(2)}$ are jointly stationary. The following propostion is proved in Appendix I.

Proposition 7.2. Suppose that $x_t^{(1)}$ and $x_t^{(2)}$ are non–nested, jointly covariance stationary regressor processes for y_t, with mean 0 and nonsingular variance matrices, and with the property that the variance matrix of the joint process $[x_t^{(1)\prime}\ x_t^{(2)\prime}]$ is singular. Suppose also that the sample variance matrices $N^{-1}\Sigma_{t=1}^{N} x_t^{(i)}x_t^{(j)\prime}$ converge

in probability to $Ex_0^{(i)}x_0^{(j)\prime}$, i,j = 1, 2. Then there exist non-singular matrices $B^{(1)}$ and $B^{(2)}$ such that

$$B^{(i)}x_t^{(i)} = \begin{bmatrix} x_t^c \\ z_t^{(i)} \end{bmatrix}, \quad i = 1,2 \tag{7.4}$$

and such that the combined process $x_t \equiv [x_t^{c\prime} z_t^{(1)\prime} z_t^{(2)\prime}]'$ satisfies

$$N^{-1} \sum_{t=1}^{N} x_t x_t' \longrightarrow_p \begin{bmatrix} W^c & 0 \\ 0 & W^z \end{bmatrix},$$

with $W^c (\equiv Ex_t^c x_t^{c\prime})$ and W^z both positive definite. Consequently, the regression model equations can be put in the form

$$y_t = Ax_t^c + \bar{A}^{(i)}z_t^{(i)} + e_t^{(i)}, \quad i = 1, 2.$$

The regressor processes $x_t^{(1)}$ and $x_t^{(2)}$ are asymptotically equivalent if and only if $\bar{A}^{(1)}z_t^{(1)} = \bar{A}^{(2)}z_t^{(2)} = 0$ (w.p.1). In this case, the combined process x_t is also asymptotically equivalent to $x_t^{(1)}$ and $x_t^{(2)}$.

Without assuming the regressors are jointly stationary, we will, for the remainder of this section, suppose that matrices $B^{(1)}$ and $B^{(2)}$ exist such that (7.4) holds and such that there is a sequence of invertible, block-diagonal "weighting" matrices,

$$D_N = \begin{bmatrix} D_N^c & 0 & 0 \\ 0 & D_N^{(1)} & 0 \\ 0 & 0 & D_N^{(2)} \end{bmatrix}, \ (N \geq N_0)$$

with the property that

$$D_N^{-1} \sum_{t=1}^{N} x_t x_t' (D_N')^{-1} \longrightarrow_p \begin{bmatrix} W^c & 0 \\ 0 & W^z \end{bmatrix} \tag{7.5}$$

holds for the process $x_t \equiv [x_t^{c\prime} z_t^{(1)\prime} z_t^{(2)\prime}]'$, where W^z is a positive definite nonstochastic matrix and W^c is positive definite w.p.1.

In the stationary case, $D_N = N^{1/2}$ times the identity matrix of appropriate order. For other types of regressors, including sinusoids and polynomials, see Theorems 10.2.6–7 and pages 581–584 of Anderson (1971) and the discussion of Grenander's conditions in Hannan (1970). For complete, unstable autoregressors of the form $(y_{t-1},\ldots,y_{t-p})'$ (no lags missing, $\dim y_t = 1$), see Chan and Wei (1988).

Theorem 10.2.11 of Anderson (1971) and Theorem VII.10 of Hannan (1970) describe somewhat different conditions under which $t_N^{(1)}$ and $t_N^{(2)}$ have (nonsingular) Gaussian limiting distributions. In Chan and Wei (1988), the common component x_t^c is nonstationary (and W^c is random), but their results show that

$$t_n^c = \text{vec}[\sum_{t=1}^{N} e_t x_t^c \{(\sum_{t=1}^{N} x_t^c x_t^{c\prime})^{-1}\}^{1/2}] \sim 0_p(1) \tag{7.6}$$

and, with $z_t = [z_t^{(1)\prime} z_t^{(2)\prime}]$, that

$$t_N^z \equiv \text{vec}\left[\Sigma^{(x)-1/2} \sum_{t=1}^{N} e_t z_t \{(\sum_{t=1}^{N} z_t z_t')^{-1}\}^{1/2}\right] \longrightarrow_{\text{dist.}} \mathcal{N}(0, \tilde{W}). \quad (7.7)$$

For simplicity, we shall assume that $\tilde{W} > 0$, so $\text{var}(e_t z_t)$ must be non–singular.

These conditions suffice to establish the results we are after. In fact, setting $\tilde{t}_N^{(i)} \equiv \Sigma_{t=1}^N e_t z_t^{(i)'} \{(\Sigma_{t=1}^N z_t^{(i)} z_t^{(i)'})^{-1}\}^{1/2}$, it follows from (7.5), and the fact that the t_N – statistics are bounded in probability, that $t_N^{(i)'} t_N^{(i)} \sim_p t_N^{c'} t_N^c + \tilde{t}_N^{(i)'} \tilde{t}_N^{(i)}$. Therefore,

$$Q_N^{(2)} - Q_N^{(1)} \sim_p \tilde{t}_N^{(2)'} \tilde{t}_N^{(2)} - \tilde{t}_N^{(1)'} \tilde{t}_N^{(1)} . \quad (7.8)$$

Now set $d^{(i)} \equiv q(r^{(i)} - r^c)$, where $r^c \equiv \dim x_t^c$. Since $\tilde{W}$ is positive definite, the matrix

$$\tilde{W}^{1/2} \begin{bmatrix} I_d(2) & 0 \\ 0 & -I_d(1) \end{bmatrix} [\tilde{W}^{1/2}]'$$

will have $d^{(2)}$ positive eigenvalues $\lambda_1^2,\ldots,\lambda_d^2(2)$ and $d^{(1)}$ negative eigenvalues $-\mu_1^2,\ldots,-\mu_d^2(1)$, see Noble(1969, p. 419). A standard argument applied to the right hand side of (7.8) leads to

Proposition 7.3 If (7.4) – (7.7) hold, then

$$(-2)\hat{L}_N^{(1,2)} \xrightarrow{N}_{\text{dist.}} \sum_{j=1}^{d^{(2)}} \lambda_j^2 \chi_j^2(1) - \sum_{k=1}^{d^{(1)}} \mu_k^2 \chi^2_{\{d^{(2)}+\, k\}}(1), \quad (7.9)$$

where $\chi^2_j(1)$, $j=1,\ldots,d^{(1)} + d^{(2)}$ are independent chi-square variates with one d.f. In particular, if both regressor processes are complete (in the sense of subsection 5.2), then

$$(-2)\hat{L}_N^{(1,2)} \xrightarrow[\text{dist.}]{N} \chi^2(d^{(2)}) - \chi^2(d^{(1)}), \tag{7.10}$$

a difference of independent chi-square variates with d.f.'s $d^{(2)}$ and $d^{(1)}$ respectively.

Davies(1980) describes an algorithm suitable for calculating values of the distributions in (7.9) and (7.10). The variance of the distribution in (7.10) is $2\{d^{(2)} + d^{(1)}\}$. This is greater than the variance $2|d^{(2)} - d^{(1)}|$ of $\chi^2(|d^{(2)} - d^{(1)}|)$, which is the limiting distribution of $(-2)\hat{L}_N^{(1,2)}$ for the comparison of complete, nested regressors when the parameter excess of the larger regressor is $|d^{(2)} - d^{(1)}|$. In this sense, non-nested comparisons are more problematic than nested comparisons. The instability of $(-2)\hat{L}_N^{(1,2)}$ is further increased when the non-nested regressors are weakly equivalent but not strongly equivalent, see the discussions below (8.7) and (8.12).

Remark When the D_N in (7.5) are multiples of the identity matrix, say $D_N = N^{1/2}I$, as in the case of bounded regressors which do not decrease too rapidly, then the block diagonal form in (7.5) can be achieved starting from the weaker requirement

$$N^{-1}\sum_{t=1}^{N} x_t x_t' \longrightarrow_p \begin{bmatrix} W^c & W^{c,1'} & W^{c,2'} \\ W^{c,1} & * & * \\ W^{c,2} & * & * \end{bmatrix} \quad (W^c > 0 \text{ w.p.1}),$$

through replacement of $z_t^{(i)}$ by $z_t^{(i)} - W^{c,i}(W^c)^{-1}x_t^c$.

Example 5.1 continued. For this example, the distribution in (7.9) is

$$5.1\chi_1^2(1) + 0.03\chi_2^2(1) - 26.0\chi_3^2(1). \tag{7.11}$$

If the variate on the right in (7.9) is denoted by $\delta(\underline{\lambda}, \underline{\mu})$, then the limiting distribution of $AIC_N^{(1,2)}$ is that of $\delta(\underline{\lambda}, \underline{\mu}) + 2q(r^{(1)} - r^{(2)})$. If, say, $r^{(1)} < r^{(2)}$, it follows that $\lim_{N\to\infty} P(AIC_N^{(1,2)} < 0)$ is always non–zero. The same is true also of $\lim_{N\to\infty} P(AIC_N^{(1,2)} > 0)$. That is, each regressor has a positive asymptotic probability of being chosen by MAIC. Some tables related to (7.10) and further discussion of Example 5.1 are given in subsection 8.3.

If $x_t^{(1)} \subseteq x_t^{(2)}$ we can arrange (7.2) so that $x_t^c = x_t^{(1)}$. Then $z_t^{(1)}$ and all terms related to it should be removed from the formulas and discussion above. In this case, (7.10) reduces to the familiar assertion of a limiting $\chi^2(q(r^{(2)} - r^{(1)}))$ distribution for $(-2)\hat{L}_N^{(1,2)}$.

These results yield corresponding results for $AIC_N^{(1,2)}$ and Ideal–$AIC_N^{(1,2)}$ by the addition of an appropriate constant. We note the following corollaries.

Corollary 7.2. If the assumptions of Proposition 7.3 hold, then the mean of the asymptotic distribution of Ideal–$AIC_N^{(1,2)}$ is $CVAR^{(1)} - CVAR^{(2)}$.

Corollary 7.3. If the Assumptions of Proposition 7.2 are satisfied, and if the regressors $x_t^{(1)}$ and $x_t^{(2)}$ are complete and have the same dimension, $r^{(1)} = r^{(2)}$, then all of the statistics $D_N^{(1,2)}[c_N^{(1,2)}]$ defined in (6.2) coincide with $(-2)L_N^{(1,2)}$ and have a symmetric limiting distribution. Thus, in this situation these criteria give equal preference to both regressors, asymptotically.

7.1. Use of (7.10) for Hypothesis Testing.

To complete this section, we remark that if the dimension r^c of the common component x_t^c in (7.4) is known or can be estimated reliably, and if it is assumed that at least one of the regressors $x_t^{(1)}$ and $x_t^{(2)}$ is complete (and that other appropriate assumptions described above hold), then (7.10) can be used to test the null hypothesis that both regressors are complete, and therefore the regressor x_t^c should be preferred, against the alternative that one regressor, presumably the one with smaller maximized likelihood value (larger $|\hat{\Sigma}_N^{(i)}|$), is not complete, see the discussion after Proposition 6.1. This is a generalization of the familiar test of the nested model against the nesting model. Consider the autopilot design problem of section 3, for example. With $x_t^{(1)} \equiv \tilde{x}_t^{(\hat{M})}$ and $x_t^{(2)} \equiv x_t^{(\hat{m})}$, values of $(-2)\hat{L}_N^{(1,2)}$ can be calculated from Table 3.1 as $DAIC_{894} - 2\Delta dimA$. Thus, the asymptotic p–value associated with the observed value of $(-2)\hat{L}_N^{(1,2)}$ under (7.10) for the regression with $y_t = P_t$, where $r^c = 7{\cdot}5 = 35$, $d^{(1)} = 7$ and $d^{(2)} = 5$, is

$$p_P = P(\chi^2(5) - \chi^2(7) \leq -18.) = 0.003 \quad . \tag{7.12}$$

The asymptotic p–value for the regression with $y_t = Y_t$, where $r^c = 4{\cdot}5 = 20$, $d^{(1)} = 4$ and $d^{(2)} = 10$, is

$$p_Y = P(\chi^2(10) - \chi^2(4) \geq 14.) = 0.07. \tag{7.13}$$

The value of p_P would cause the null hypothesis to be rejected in favor of a preference for the regressor including VACC values at the significance levels usually used. The value of p_Y would lead to acceptance of the null hypothesis at some popular significance levels. The decision to reject the null hypothesis coincides in

these examples with the decision reached via MAIC. However, in the case of (7.13), acceptance of the null hypothesis, favoring the "intersection" regressor x_t^c containing just the variables which both regressors share, would also exclude VACC. Thus (7.10) does not provide a way of testing for the inclusion of a specific variable unless this variable only occurs in the regressor process having greater dimension.

There are other modified–log–likelihood–ratio procedures for doing regressor selection via hypothesis–testing with non–nested models deriving from Cox (1961, 1962). These have been extensively developed in the theoretical econometrics literature, but not widely used, it appears. Their adjustments to the log–likelihood ratio are more difficult to calculate than those we have discussed because they require an approximation to the expected value of the log–likelihood ratio under the null hypothesis. The test statistic requires a consistent estimate of the corresponding asymptotic variance as well. It seems to happen rather frequently in applications of Cox tests that each model of the pair under consideration is rejected in favor of the other. For a survey of the econometrics literature concerned with these procedures see Judge et al.(1984, pp. 883–888) and White (1989), and their references. In some limited simulation experiments by Tsurumi and Wago (1987), MAIC did a more satisfactory job of regressor selection than the Cox–test procedure they investigated.

Hypothesis testing would seem to be an appropriate tool for vindicating one theory over another when the theories specify competing regression models. It is a less natural procedure for trying to decide which of two possibly incomplete regressors has greater predictive power.

8. AN IDEAL CRITERION: MAXIMIZING THE KULLBACK–LEIBLER NUMBER

We will now investigate some properties relevant for regressor selection which are possessed by the expected log–likelihood function,

$$\mathcal{E}_N^{(x)}[\Sigma, A] \equiv E\{L_N^{(x)}[\Sigma, A]\},$$

a quantity we will call the Kullback–Leibler (K–L) number, or cross–entropy, associated with A and Σ and the regressor x_t. We will assume that (A3) and (A4) of section 2 hold for the regressor processes under consideration. Then, from the decomposition,

$$e_t^{(A)} \equiv y_t - Ax_t = e_t^{(x)} + (A^{(x)} - A)x_t$$

we obtain

$$Ee_t^{(A)}e_t^{(A)\prime} = \Sigma^{(x)} + (A - A^{(x)})Ex_t x_t'(A - A^{(x)})' ,$$

which leads via the definition of $L_N[\Sigma, A]$ to the basic formula for $\mathcal{E}_N^{(x)}[\Sigma, A]$,

$$\mathcal{E}_N^{(x)}[\Sigma, A] = -\frac{N}{2}(\log 2\pi|\Sigma| + tr\Sigma^{-1}\Sigma^{(x)})$$

$$-\frac{1}{2} tr\Sigma^{-1}(A - A^{(x)})\Big\{E\sum_{t=1}^{N} x_t x_t'\Big\}(A - A^{(x)})' . \qquad (8.1)$$

If $N \geq N_0$ as in (2.1), then $E\Sigma_{t=1}^{N} x_t x_t'$ is positive definite and the quadratic expression in $A - A^{(x)}$ in (8.1) takes on its maximum value, 0, only when $A = A^{(x)}$. An elementary analysis of the eigenvalues (see the proof of Lemma 3.1 of Hosoya and Taniguchi (1982)) shows that the other term in (8.1) is uniquely maximized at $\Sigma = \Sigma^{(x)}$. Thus, $\mathcal{E}_N^{(x)}[\Sigma, A]$ is uniquely maximized at $\Sigma = \Sigma^{(x)}$, $A = A^{(x)}$:

$$\Sigma^{(x)},\ A^{(x)} : \mathcal{E}_N^{(x)}[\Sigma, A] \rightarrow \max! \quad (N \geq N_0). \tag{8.2}$$

From (8.1) we obtain that, for all N=1,2,...,

$$N^{-1}\mathcal{E}_N^{(x)}[\Sigma^{(x)}, A^{(x)}] = -\tfrac{1}{2}\{\log 2\pi|\Sigma^{(x)}| + q\}\ . \tag{8.3}$$

Given regressor processes $x_t^{(i)}$, i=1,2, we define $\hat{\mathcal{E}}_N^{(i)} \equiv \mathcal{E}_N^{(i)}[\hat{\Sigma}_N^{(i)}, \hat{A}_N^{(i)}]$, $\mathcal{E}_{N,\infty}^{(i)} \equiv \mathcal{E}_N^{(i)}[\Sigma^{(i)}, A^{(i)}]$, and $\hat{\mathcal{E}}_N^{(1,2)} \equiv \hat{\mathcal{E}}_N^{(1)} - \hat{\mathcal{E}}_N^{(2)}$. Note, from (8.3) that

$$\mathcal{E}_{N,\infty}^{(1)} > \mathcal{E}_{N,\infty}^{(2)} \text{ if and only if } |\Sigma^{(1)}| < |\Sigma^{(2)}|\ . \tag{8.4}$$

From (8.1) we also obtain

$$\hat{\mathcal{E}}_N^{(i)} = -\tfrac{N}{2}\{\log 2\pi|\hat{\Sigma}_N^{(i)}| + \mathrm{tr}(\hat{\Sigma}_N^{(i)})^{-1}\Sigma^{(i)}\}$$

$$-\tfrac{1}{2}\mathrm{tr}[(\hat{\Sigma}_N^{(i)})^{-1}(\hat{A}_N^{(i)} - A^{(i)})\{E\sum_{t=1}^{N} x_t^{(i)}x_t^{(i)\prime}\}(\hat{A}_N^{(i)} - A^{(i)})']. \tag{8.5}$$

The regression analogue of the entropy maximization principle of Akaike (1977) asserts that the regressor $x_t^{(i)}$ with larger $\hat{\mathcal{E}}_N^{(i)}$ is to be preferred:

(EMP) Prefer $x_t^{(1)}$ if $\hat{\mathcal{E}}_N^{(1,2)} > 0$.

In this section, we shall show that this principle favors the desired regressor both when $|\Sigma^{(1)}| \neq |\Sigma^{(2)}|$ and when $x_t^{(1)}$ and $x_t^{(2)}$ are asymptotically equivalent. In other words, $\hat{\mathcal{E}}_N^{(1,2)}$ does not exhibit the dichotomous behavior of $\hat{L}_N^{(1,2)}$ discussed in section 5. We will give separate analyses for the situations

$$|\Sigma^{(1)}| < |\Sigma^{(2)}| \tag{8.6}$$

and

$$|\Sigma^{(1)}| = |\Sigma^{(2)}| \ . \tag{8.7}$$

The situation $|\Sigma^{(2)}| < |\Sigma^{(1)}|$ is covered by interchanging indices, $(1,2) \longleftrightarrow (2,1)$, in the discussion of (8.6).

In the nested case, $x_t^{(1)} = Bx_t^{(2)}$, it follows from the uniqueness of the maximizer of $\mathcal{E}_N^{(2)}[\Sigma,A]$ that (8.7) is equivalent to $A^{(2)} = A^{(1)}B$, which is equivalent to $e_t^{(1)} = e_t^{(2)}$. In the non–nested case, however, the condition (8.7), defining what we shall call weak equivalence of regressors, is weaker than $e_t^{(1)} = e_t^{(2)}$. For instance, let y_t be a mean zero stationary process for which non–zero autocorrelations at two different lags coincide, $\rho_k(1) = \rho_k(2) = \rho$ $(k^{(1)} \neq k^{(2)};\ 0 < |\rho| < 1)$; an example would be a stationary third–order autoregressive process with $\rho_1 = \rho_3 = 0.2$

and $\rho_2 = 0.5$. Then, for $x_t^{(i)} \equiv y_{t-k}(i)$, the error processes $e_t^{(i)} = y_t - \rho x_t^{(i)}$ (i = 1,2) are distinct, but $\Sigma^{(1)} = \Sigma^{(2)} = (Ey_o^2)(1 - \rho^2)$.

8.1 A Consistency Result for EMP when $|\Sigma^{(1)}| < |\Sigma^{(2)}|$.

If the trace term on the right in (8.5) is bounded in probability, or even just $o_p(N)$, for i=1,2, then it follows from $\hat{\Sigma}_N^{(i)} \xrightarrow[p]{N} \Sigma^{(i)}$ and (4.1) that, under (8.6),

$$\lim_{N\to\infty} P(\hat{\mathcal{E}}_N^{(1,2)} > 0) = 1 \ , \tag{8.8}$$

showing that $x_t^{(i)}$ is preferred by EMP with asymptotic probability one. For example, if (A1) and (A5) hold, then

$$R_N^{(i)} \equiv \mathrm{tr}(\hat{\Sigma}_N^{(i)})^{-1}(\hat{A}_N^{(i)} - A^{(i)})\{E \sum_{t=1}^{N} x_t^{(i)} x_t^{(i)\prime}\}(\hat{A}_N^{(i)} - A^{(i)})' \tag{8.9}$$

is easily seen to satisfy $R_N^{(i)} \sim_p Q_N^{(i)}$, so $R_N^{(i)}$ is bounded in probability. Hence the same is true of the trace term in (8.5) and (8.8) applies.

8.2 Results for the Cases $|\Sigma^{(1)}| = |\Sigma^{(2)}|$ and $e_t^{(1)} = e_t^{(2)}$.

When (8.7) holds, then $\mathcal{E}_{N,\infty}^{(1)} = \mathcal{E}_{N,\infty}^{(2)}$ and we have

$$\hat{\mathcal{E}}_N^{(1,2)} = \left\{\mathcal{E}_{N,\infty}^{(2)} - \hat{\mathcal{E}}_N^{(2)}\right\} - \left\{\mathcal{E}_{N,\infty}^{(1)} - \hat{\mathcal{E}}_N^{(1)}\right\} \ , \tag{8.10}$$

a decomposition in which the bracketed terms are nonnegative, by (8.2). In Appendix I, we will demonstrate that, under (A1) – (A5), the likelihoods and K–L numbers deviate from their maximum values in the same way,

$$\hat{L}_N^{(x)} - L_{N,\infty}^{(x)} \sim_p \mathcal{E}_{N,\infty}^{(x)} - \hat{\mathcal{E}}_N^{(x)} , \tag{8.11}$$

a phenomenon first examined by Akaike and Shimizu in the case of overparameterized autoregressions, see Shimizu (1978). When (8.11) holds, then (8.10) yields

$$\hat{\mathcal{E}}_N^{(1,2)} \sim_p \left\{\hat{L}_N^{(2)} - L_{N,\infty}^{(2)}\right\} - \left\{\hat{L}_N^{(1)} - L_{N,\infty}^{(1)}\right\} . \tag{8.12}$$

It follows from (8.12) and the analysis of the expressions (I.5)–(I.7) in Appendix I that, when Assumptions (A1)–(A5) are satisfied, then $\hat{\mathcal{E}}_N^{(1,2)}$ is bounded in probability if (and only if) (8.7) holds, a cleaner result than is possible for $\hat{L}_N^{(1,2)}$. In fact,

$$\hat{L}_N^{(1,2)} = \{\hat{L}_N^{(1)} - L_{N,\infty}^{(1)}\} - \{\hat{L}_N^{(2)} - L_{N,\infty}^{(2)}\} + \{L_{N,\infty}^{(1)} - L_{N,\infty}^{(2)}\}, \tag{8.13}$$

and, under (8.7),

$$L_{N,\infty}^{(1)} - L_{N,\infty}^{(2)} = (1/2)\sum_{t=1}^{N} \{e_t^{(2)\prime}\Sigma^{(2)-1}e_t^{(2)} - e_t^{(1)\prime}\Sigma^{(1)-1}e_t^{(1)}\},$$

which has mean zero but magnitude $0_p(N^{1/2})$ when $N^{-1/2}$ times the right hand sum has a nondegenerate limiting distribution, as in the example discussed after (8.7).

Since $L_{N,\infty}^{(1)} = L_{N,\infty}^{(2)}$ when $e_t^{(1)} = e_t^{(2)}$, we obtain the following fundamental result, from which regressor comparison properties of EMP follow.

Proposition 8.1. If the regressor processes $x_t^{(1)}$ and $x_t^{(2)}$ for y_t are asymptotically equivalent and satisfy assumptions (A1)–(A5) of section 4, then $\hat{E}_N^{(1,2)}$ and $\hat{L}_N^{(1,2)}$ behave oppositely for large N, in the precise sense of (8.14):

$$\hat{\mathcal{E}}_N^{(1,2)} \sim_p -\hat{L}_N^{(1,2)} \quad . \tag{8.14}$$

Note that if $x_t^{(1)} \subseteq x_t^{(2)}$ and $-2\hat{L}_N^{(1,2)} \to_{\text{dist.}} \Sigma_{j=1}^{r^{(2)}-r^{(1)}} \lambda_j^2 \chi_j^2(1)$ as in (7.9), then (8.14) implies

$$\lim_{N\to\infty} P(\hat{\mathcal{E}}_N^{(1,2)} > 0) = 1 \quad .$$

Thus, among nested, asymptotically equivalent regressors satisfying (7.9) and (8.14), EMP consistently prefers the regressor with smallest dimension. This consistency property is more limited in scope than that of the strongly parsimonious criteria, see Proposition 6.2. In light of Example 5.1, this more limited scope may be desirable. What are the large sample properties of $\hat{\mathcal{E}}_N^{(1,2)}$ for this example? The result (8.12) and some results of section 7 combine to provide the basis for an answer to this question.

Proposition 8.2. If (A1)–(A4) and the other assumptions of Proposition 7.1 or Proposition 7.3 are satisfied by the asymptotically equivalent regressors $x_t^{(1)}$ and $x_t^{(2)}$, then $2\hat{\mathcal{E}}_N^{(1,2)}$ has an asymptotic distribution whose mean is $CVAR^{(2)} - CVAR^{(1)}$.

Thus, asymptotically, (EMP) has a tendency to favor the regressor with smaller CVAR.

Example 5.1 continued. For this example, $CVAR^{(2)} < CVAR^{(1)}$. The asymptotic probability that $x_t^{(2)}$ is selected by EMP can be obtained from (7.11):

$$\lim_{N\to\infty} P(\hat{\mathcal{E}}_N^{(1,2)} < 0) = 0.73.$$

Similarly, the asymptotic probabilities that $AIC_N^{(1,2)}$ and Ideal–$AIC_N^{(1,2)}$ lead to the selection of $x_t^{(2)}$ are

$$\lim_{N\to\infty} P(AIC_N^{(1,2)} > 0) = 0.17$$

and

$$\lim_{N\to\infty} P(\text{Ideal–}AIC_N^{(1,2)} > 0) = 0.81,$$

respectively. See (6.2) for the definition of Ideal–$AIC_N^{(1,2)}$.

It can be shown that if $x_t^{(1)}$ and $x_t^{(2)}$ are only weakly equivalent and $\Sigma^{(1)} \neq \Sigma^{(2)}$, then the asymptotic distribution of $2\hat{\mathcal{E}}_N^{(1,2)}$ (which exists rather generally) has a mean which can involve additional terms related to the variances of the estimates of $\Sigma^{(1)}$ and $\Sigma^{(2)}$, see (10.16) below. Arguments like those of subsection 5.3 can be used to show that the additional terms will be negligible if the regressors are nearly complete, see Findley (1985).

8.3 Ideal–$AIC_N^{(1,2)}$ and $AIC_N^{(1,2)}$.

The attractive properties of EMP described in subsection 8.2 provide the motivation for our definition in section 6 of

$$\text{Ideal–AIC}_N^{(1,2)} = (-2)\hat{L}_N^{(1,2)} + 2\{\text{CVAR}^{(1)} - \text{CVAR}^{(2)}\}.$$

This was defined in such a way that it has the same asymptotic mean as $\hat{\mathcal{E}}_N^{(1,2)}$, see Proposition 8.2 and Corollary 7.2. Similarly, $\text{AIC}_N^{(1,2)}$ is motivated by the special case in which the regressors are complete in the sense of subsection 5.2. In these contexts, our results show that $\text{Ideal–AIC}_N^{(1,2)}$ and $\text{AIC}_N^{(1,2)}$ are asymptotically unbiased estimators of $\hat{\mathcal{E}}_N^{(1,2)}$, the property emphasized by Akaike (1973, 1977), who discusses $\text{AIC}_N^{(i)}$ as a bias–corrected estimate of $\hat{\mathcal{E}}_N^{(i)}$. It is clear from (8.14) that these estimators are not consistent.

One of us (D.F.F.) will report elsewhere on simulation experiments concerning the estimation of $\text{CVAR}^{(x)}$ for scalar autoregressions, in order to directly estimate $\text{Ideal–AIC}_N^{(1,2)}$. Lacking such estimates, it is properties of $\text{AIC}_N^{(1,2)}$ which are of practical interest.

To get some sense of the asymptotic behavior of MAIC for non–nested models, we will now look at the case of complete regressors, where (7.10) holds. We assume that $d \equiv q(r^{(2)} - r^{(1)}) > 0$. Let $m \equiv q(r^{(1)} - r^c)$, where r^c is the dimension of the shared regressor x_t^c in (7.4). Then the variate on the right in (7.10) becomes

$$\delta(m,d) \equiv \chi^2(m+d) - \chi^2(m) \ . \tag{8.15}$$

In Table 8.1 below, three sets of $\delta(m,d)$–probabilities are given for a range of values of m and d. These are asymptotic probabilities of selection of the more parsimonious regressor $x_t^{(1)}$, which has the smaller CVAR value, see (5.16):

$$\lim_{N\to\infty} P(\text{AIC}_N^{(1,2)} < 0) = P(\delta(m,d) < 2d);$$

$$\lim_{N\to\infty} P(\hat{\varepsilon}_N^{(1,2)} > 0) = P(\delta(m,d) > 0);$$

and

$$\lim_{N\to\infty} P(AIC_N^{(1,2)} < 0,\ \hat{\varepsilon}_N^{(1,2)} > 0) = P(0 < \delta(m,d) < 2d)\ ,$$

the last being the asymptotic probability that MAIC and EMP agree on the choice of $x_t^{(1)}$. For this situation, the low probabilities which arise when m is larger than d are a consequence of (8.14).

Table 8.1 Asymptotic Probabilities of Parsimonious Choice Between Complete, Non–Nested Regressors, by AIC, EMP and Both Simultaneously. $d \equiv \dim A^{(2)} - \dim A^{(1)}$; m is the number of estimated coefficients for variables in $x_t^{(1)}$ which are not linear combinations of those in $x_t^{(2)}$; $\delta(m,d)$ is defined in (8.15).

$P(\delta(m,d) < 2d)$

m/d	1	2	6	12	18	∞
0	.84	.87	.94	.98	.99	1.00
1	.74	.81	.92	.98	.99	1.00
2	.68	.77	.90	.97	.99	1.00
6	.59	.67	.85	.95	.98	1.00
12	.56	.62	.79	.92	.97	1.00
18	.55	.59	.75	.89	.95	1.00
∞	.50	.50	.50	.50	.50	

$P(\delta(m,d) > 0)$

m/d	1	2	6	12	18	∞
0	1.00	1.00	1.00	1.00	1.00	1.00
1	.71	.82	.97	1.00	1.00	1.00
2	.65	.75	.94	.99	1.00	1.00
6	.58	.65	.86	.97	.99	1.00
12	.56	.61	.79	.93	.98	1.00
18	.55	.59	.75	.90	.96	1.00
∞	.50	.50	.50	.50	.50	

$P(0 < \delta(m,d) < 2d)$

m/d	1	2	6	12	18	∞
0	.84	.87	.94	.98	.99	1.00
1	.45	.63	.89	.98	.99	1.00
2	.33	.52	.84	.96	.99	1.00
6	.17	.32	.71	.92	.97	1.00
12	.12	.23	.58	.85	.95	1.00
18	.10	.10	.50	.79	.91	1.00
∞	0.00	0.00	0.00	0.00	0.00	

9. A SECOND COST FUNCTION: NORMALIZED MEAN SQUARE PREDICTION ERROR WITH INDEPENDENT REPLICATES.

One would expect that, between asymptotically equivalent regressors, one important consequence of greater coefficient estimation variability would be diminished predictive performance. In this section, we establish a connection, between $CVAR^{(2)} - CVAR^{(1)}$ and the corresponding difference of a measure of mean square prediction error in two situations: predicting independent replicates of the data used to estimate the regression coefficients; and predicting the observation set used for estimation.

Let $\hat{A}_N^{(x)}$ denote the least squares coefficient estimate of $A^{(x)}$ in the model

$$y_t = A^{(x)}x_t + e_t^{(x)} \quad ,$$

from data y_t, x_t, t=1,...,N. We assume that (A1)–(A5) hold. As before, $Ee_t^{(x)}e_t^{(x)'}$ is denoted by $\Sigma^{(x)}$. Let $\bar{y}_t$, $\bar{x}_t$, t=1,...,N denote an independent replicate of the data which were used to determine $\hat{A}_N^{(x)}$ and let $\bar{E}$ denote the expectation operator for this replicate. Consider the normalized mean square prediction error measure defined by

$$\mathrm{MSPE}_N^{(x)} \equiv E\bar{E}[\sum_{t=1}^{N}(\bar{y}_t - \hat{A}_N^{(x)}\bar{x}_t)'\Sigma^{(x)-1}(\bar{y}_t - \hat{A}_N^{(x)}\bar{x}_t)]. \qquad (9.1)$$

From the decomposition $\bar{y}_t - \hat{A}_N^{(x)}\bar{x}_t = \bar{e}_t^{(x)} + (A^{(x)} - \hat{A}_N^{(x)})\bar{x}_t$ and (A3) – (A4), we obtain

$$\mathrm{MSPE}_N^{(x)} = Nq + \mathrm{tr}E[(\Sigma^{(x)})^{-1}(\hat{A}_N^{(x)} - A^{(x)})\Big[\bar{E}\sum_{t=1}^{N}\bar{x}_t\bar{x}_t\Big](\hat{A}_N^{(x)} - A^{(x)})'].$$

Hence, for $R_N^{(i)}$ defined in (8.9), we have

$$\mathrm{MSPE}_N^{(2)} - \mathrm{MSPE}_N^{(1)} = E\{R_N^{(2)} - R_N^{(1)}\} .$$

Since $R_N^{(i)} \sim_p Q_N^{(i)}$ under our assumptions, we would expect to have

$$\lim_{N\to\infty}\{\mathrm{MSPE}_N^{(2)} - \mathrm{MSPE}_N^{(1)}\} = \lim_{N\to\infty} E\{Q_N^{(2)} - Q_N^{(1)}\}$$

$$= \mathrm{CVAR}^{(2)} - \mathrm{CVAR}^{(1)} . \qquad (9.2)$$

It follows from taking expectations in (4.3) that $E\{Q_N^{(2)} - Q_N^{(1)}\}$ corresponds to the difference of the normalized mean square prediction error obtained if, instead of the independent replicate in (9.1), the data used to estimate the $\hat{A}_N^{(i)}$ are predicted. The equalities in (9.2) establish a connection between estimation variability and prediction error. In section 10, we shall describe how (9.2) can be verified for some important classes of models.

The results of Kunitomo and Yamamoto (1985) show that the analogue of (9.2) for the same–realization forecast error quantities $N^{1/2}(y_{N+1} - \hat{A}^{(x)}x_N)(\Sigma^{(x)-1})^{1/2}$

contains additional terms. The examples in their Table 3 can be rescaled by $(\Sigma^{(x)})^{-1}$as in (9.1) to show that $MSPE_N^{(x)}$ and $CVAR^{(x)}$ can be smaller for an incorrect regressor than for the correct regressor. Thus these quantities by themselves (when they can be adequately estimated) do not provide completely satisfactory regressor selection criteria. In theory, they can be used to discriminate between weakly equivalent regressors, as defined in the preceding section, a different situation from that of Table 3 of the above reference.

10. CONVERGENCE OF FINITE–SAMPLE MEANS TO THE ASYMPTOTIC MEANS.

To increase our confidence in the relevence to the moderate sample size situation of the asymptotic results given in sections 5, 7 and 8, we would like to know that convergence in distribution or probability leads to convergence of the means, for example,

$$\lim_{N\to\infty} E\{-2\hat{L}_N^{(1,2)}\} = CVAR^{(2)} - CVAR^{(1)} \qquad (10.1)$$

This is the same issue that arose with (9.2). This chapter shows how such results can be obtained, including complete verifications for two important Gaussian cases: non–stochastic regressors; and subregressions of full–rank autoregressive processes. The Gaussian version of the Example 5.1 will be encompassed by our discussion.

Let Q denote a matrix with stochastic entries and let $|\cdot|$ denote a convenient matrix norm, see Noble (1969). For any $\beta \geq 1$, define the β–mean (or L^β–) norm of Q by $|Q|_\beta = \{E|Q|^\beta\}^{1/\beta}$. This will be finite if and only if all the entries Q_{ij} of Q satisfy $E|Q_{ij}|^\beta < \infty$. Our basic strategy can be summarized in two lemmas.

Lemma 10.1. (Billingsley (1985, p. 348)) If $Q_N \xrightarrow{N}_{dist.} Q$, and also, for some N_0 and some $\epsilon>0$, the condition, $\sup_{N\geq N_0}\|Q_N\|_{1+\epsilon} < \infty$ is satisfied, then $EQ_N \xrightarrow{N} EQ$.

Using the matrix norm inequality $\|Q_1Q_2 \cdots Q_m\| \leq \|Q_1\|\|Q_2\| \cdots \|Q_m$ and Hölder's inequality, it is easy to verify

Lemma 10.2. Given $\epsilon>0$ and $\beta_j \geq 1$, j=1,...,m such that $\beta_1^{-1} + \cdots + \beta_m^{-1} = 1$,

then $\|Q_1Q_2 \cdots Q_m\|_{1+\epsilon} \leq \|Q_1\|_{(1+\epsilon)\beta_1} \cdots \|Q_m\|_{(1+\epsilon)\beta_m}$.

We are investigating equivalent regressors, with $e_t^{(1)} = e_t^{(2)} = e_t$. All of the quantities we wish to examine, $\hat{L}_N^{(1,2)}$, $Q_N^{(i)}$, etc. are unchanged by the transformation $e_t \rightarrow (\Sigma^{(i)})^{-1/2}e_t$, so we will assume for the remainder of our discussion that

$$\Sigma^{(i)} = I_q \quad . \tag{10.2}$$

Then the eigenvalues $\hat{\lambda}_{j,N}^{(i)}$ of $\hat{\Sigma}_N^{(i)}$ converge to 1 in probability. We will only consider the mean of $\hat{L}_N^{(1,2)}$; the arguments for the other quantities are similar. Using a first degree Taylor expansion of $\log\lambda$ about $\lambda = 1$, we show in Appendix I that

$$|(-2)\hat{L}_N^{(1,2)}| \leq |N\mathrm{tr}(\hat{\Sigma}_N^{(2)} - \hat{\Sigma}_N^{(1)})|$$

$$+ \sum_{i=1}^{2} \mathrm{tr}[\{(\hat{\Sigma}_N^{(i)})^{-2} + I_q\}\{N^{1/2}(\hat{\Sigma}_N^{(i)} - I_q)\}^2]. \tag{10.3}$$

Set $\Sigma_N \equiv N^{-1}\Sigma_{t=1}^{N} e_t e_t'$. Now,

$$|\mathrm{Ntr}(\hat{\Sigma}_N^{(2)} - \hat{\Sigma}_N^{(1)})| \leq \sum_{i=1}^{2} \mathrm{Ntr}(\Sigma_N - \hat{\Sigma}_N^{(i)}) , \tag{10.4}$$

and

$$\hat{\Sigma}_N^{(i)} - I_q = (\hat{\Sigma}_N^{(i)} - \Sigma_N) + (\Sigma_N - I_q) . \tag{10.5}$$

Using (4.3), we can rewrite $\Sigma_N - \hat{\Sigma}_N^{(i)}$ as a product of analyzable factors,

$$\Sigma_N - \hat{\Sigma}_N^{(i)} = N^{-1}\{\sum_{t=1}^{N} e_t x_t^{(i)} C_N^{'-1}\}\{C_N'(\sum_{t=1}^{N} x_t^{(i)} x_t^{(i)'})^{-1} C_N\}\{\sum_{t=1}^{N} e_t x_t^{(i)'} C_N^{'-1}\}' . \tag{10.6}$$

By substituting (10.4) – (10.6) into the right–hand side of (10.3), one obtains an upper bound for $|(-2)\hat{L}_N^{(1,2)}|$ which is a sum of products involving up to eight factors. Since $|(-2)L_N^{(1,2)}|_{1+1/8}$ will be less than the sum of the (1+1/8)–norms of each of the products, we can establish

$$\sup_{N \geq N_0} |(-2)\hat{L}_N^{(1,2)}|_{1+1/8} < \infty$$

by verifying the moment conditions (AM1) – (AM4) below and applying Lemma 10.2. Then, if $C_N^{(1)}$ and $C_N^{(2)}$ are matrices such that (7.2) and (7.3) hold, or if (7.4) – (7.7) hold, and

$$C_N^{(i)} \equiv \begin{bmatrix} D_N^c & 0 \\ 0 & D_N^{(i)} \end{bmatrix} B^{(i)},$$

we can use Propositions 7.1 or 7.3 and Lemma 10.1 to obtain (10.1).

The remaining subsections are devoted to describing the situations in which we have been able to verify the following ninth moment conditions:

(AM1) $$\sup_{N \geq N_0} \left| \sum_{t=1}^{N} e_t x_t' C_N'^{-1} \right|_9 < \infty \quad ,$$

(AM2) $$\sup_{N \geq N_0} |N^{1/2}(\Sigma_N - I_q)|_9 < \infty \quad ,$$

(AM3) $$\sup_{N \geq N_0} |(\hat{\Sigma}_N^{(x)})^{-1}|_9 < \infty \quad ,$$

(AM4) $$\sup_{N \geq N_0} |C_N'(\sum_{t=1}^{N} x_t x_t')^{-1} C_N|_9 < \infty \quad ,$$

for some sufficiently large N_0.

Substantially greater generality can be achieved for (AM1) and (AM2). We start with these.

10.1. Regressions Satisfying (AM1) and (AM2).

The easiest results deal with the case in which e_t is <u>independent</u> of x_t and e_s, $s < t$, and, in addition, the moment conditions

$$\sup_t |e_t|_k < \infty \qquad (10.7)$$

and

$$\sup_{N \geq N_0} |C_N^{-1} \sum_{t=1}^{N} x_t x_t' C_N'^{-1}|_{k/2} < \infty \qquad (10.8)$$

hold for some $k \geq 9$. Indeed, if $\tilde{e}_t$ and $\tilde{x}_t$ $(= \tilde{x}_t(N))$ denote entries of e_t and $C_N^{-1}x_t$, respectively, then it follows from Burkholder's inequality, see Hall and Heyde (1980, p. 23), that there is a constant K_0 such that

$$|\sum_{t=1}^{N} \tilde{e}_t \tilde{x}_t|_k \leq K_0 |(\sum_{t=1}^{N} \tilde{e}_t^2 \tilde{x}_t^2)|_{k/2}^{1/2}$$

$$\leq K_0 (\sum_{t=1}^{N} |\tilde{e}_t^2 \tilde{x}_t^2|_{k/2})^{1/2} \quad \text{(triangle inequality)}$$

$$\leq K_0 (\sup_t |\tilde{e}_t|_k)(\sum_{t=1}^{N} |\tilde{x}_t|_{k/2}^2)^{1/2} ,$$

since $E|\tilde{e}_t \tilde{x}_t|^k = E|\tilde{e}_t|^k E|\tilde{x}_t|^k$. Thus the boundedness of $|\Sigma_{t=1}^{N} e_t x_t' C_N'^{-1}|_k$, $N = 1,2,...$, which implies (AM1), follows from (10.7) and (10.8). The condition (10.8) is satisfied, for example, if x_t is non–stochastic and if $C_N x_t x_t' C_N'^{-1}$ is convergent, as we assumed in section 7, or if x_t has stationary k–th order moments and $C_N = N^{1/2} I_r$.

A variety of results are available which lead to (AM1) and (AM2) <u>without independence assumptions</u>, using instead either a linear representation assumption, see Lemma 3.3 of Bhansali (1981), or mixing assumptions, see Yokoyama (1980), Theorem 5.1 of Brillinger (1969) and sections 3 and 4 of Chapter 1 of Zhurbenko (1986). Yokoyama's, Bhansali's and Brillinger's results require that e_t be stationary

with mean 0 or, for (AM1), that e_t and x_t be jointly stationary with mean 0 and $Ee_t x_t' = 0$ (our (A4)), and they cover the stationary Gaussian subregressions considered in the next subsection. Zhurbenko's results do not require stationarity and could be used when x_t is a bounded, non–stochastic regressor sequence. Brillinger's and Zhurbenko's results establish the boundedness of cumulants. Since a moment of order k can be obtained from sums of products of cumulants of orders k and less, see McCullagh(1987), the boundedness of k–th moments follows. The interested reader may consult these references for further details.

For (AM3) and (AM4), we utilize Gaussian assumptions.

10.2 Regressions Satisfying (AM3) and (AM4).

If ϵ_t is a sequence of independent $\mathcal{N}(0,\Sigma)$ random h–vectors with $\Sigma > 0$ (positive definite), then $W_N = \sum_{t=1}^{N} \epsilon_t \epsilon_t'$ has the Wishart distribution $W_h(\Sigma,N)$. If $\lambda_{min}(W_N)$ denotes the minimum eigenvalue of W_N, then $\lambda_{min}^{-1}(W_N)$ is the maximum eigenvalue of W_N^{-1}, which is a convenient matrix norm for W_N^{-1}. The following lemma concerning the Wishart distribution is fundamental to our investigation of (AM3) and (AM4). It appears to be new. We will use $\sim$ for "is distributed as."

Lemma 10.3 If $W_N \sim W_h(\Sigma, N)$, N=1,2 ..., then for every $k \geq 1$,

$$\sup_{N \geq h+2k} N^k E\{\lambda_{min}^{-k}(W_N)\} < \infty . \qquad (10.9)$$

The proof of this lemma, and of the Propositions 10.1 and 10.2 below, are given in Appendix II.

Remark. Complete, Non–Stochastic Regressors with Gaussian Errors. We use the notation N.I.D.$(0,\Sigma)$ to indicate an i.i.d. $\mathcal{N}(0,\Sigma)$ process. Observe that if the regressors x_t are non–stochastic and the errors e_t are N.I.D.$(0,\Sigma^{(x)})$, then $N\hat{\Sigma}_N^{(x)} \sim W_r(\Sigma^{(x)},N-qr)$, and (AM3) follows from this lemma. The condition (AM4) holds if the sequence $C_N^{-1}\Sigma_{t=1}^{N} x_t x_t' C_N'^{-1}$ has a nonsingular limit, as in the examples in Hannan(1970) and Anderson(1971) referred to after (7.5). Then (10.8) also holds, as well as (10.7), and (10.1) follows.

For the case of stochastic regressors, we will obtain our most general verification of (AM3) in the Corollary of the following result.

Proposition 10.1. Suppose that $\tilde{x}_t$ is a $\tilde{r}$–dimensional, not necessarily stationary, autoregressive process of order p, which satisfies

$$\tilde{x}_t = A_1\tilde{x}_{t-1} + \cdots + A_p\tilde{x}_{t-p} + a_t \,, \tag{10.10}$$

where $a_t \sim$ N.I.D.$(0,\Sigma)$ $(\Sigma > 0)$ is independent of $\tilde{x}_s$, $-p+1 \le s < t$. Let the initializing values $\tilde{x}_{-p+1},\ldots,\tilde{x}_0$ have a joint density function, $f(\tilde{x}_{-p+1},\ldots,\tilde{x}_0)$. Then, if $\hat{A}_1,\ldots,\hat{A}_p$ denote the least squares estimates of $A_1,\ldots,A_p$ from the data $\tilde{x}_{-p+1},\ldots,\tilde{x}_N$, the error variance matrix estimate defined by

$$\hat{\Sigma}_N = N^{-1}\sum_{t=1}^{N}(\tilde{x}_t - \hat{A}_1\tilde{x}_{t-1} - \cdots - \hat{A}_p\tilde{x}_{t-p})(\tilde{x}_t - \hat{A}_1\tilde{x}_{t-1} - \cdots - \hat{A}_p\tilde{x}_{t-p})' \tag{10.11}$$

has the property that for every k=1,2,..., the k–th moments of $\hat{\Sigma}_N^{-1}$ are ultimately bounded: that is, there is an N(k) such that

$$\sup_{N \geq N(k)} E\{\lambda_{min}^{-k}(\hat{\Sigma}_N)\} < \infty \ . \tag{10.12}$$

Now we introduce the concept of a subregression and show how (10.12) can be applied to subregressions of (10.10). Suppose that $\tilde{x}_t$ in (10.10) has the form $\tilde{x}_t = [y_t' \ v_t']'$. Then the residual variance matrix $\bar{\Sigma}_N$ of the regression of y_t on $\bar{x}_t = [\tilde{x}_{t-1}', \ldots, \tilde{x}_{t-p}']'$ is a submatrix of $\hat{\Sigma}_N$, and therefore $\lambda_{min}(\hat{\Sigma}_N) \leq \lambda_{min}(\bar{\Sigma}_N)$. If the regressor x_t of interest for y_t is a subvector of $\bar{x}_t$, we will further have $\bar{\Sigma}_N \leq \hat{\Sigma}_N^{(x)}$, which leads to $\lambda_{min}(\bar{\Sigma}_N) \leq \lambda_{min}(\hat{\Sigma}_N^{(x)})$ and, therefore, finally to $\lambda_{min}^{-k}(\hat{\Sigma}_N^{(x)}) \leq \lambda_{min}^{-k}(\hat{\Sigma}_N)$ for $k=1,2,\ldots$. We will summarize the regressor situation just described by saying that the regression of y_t on x_t is a <u>subregression</u> of (10.10), or alternatively, is a <u>subautoregression</u>. For example, the regressors $x_t^{(1)}$ and $x_t^{(2)}$ of Example 5.1 are subregressions of the correct AR(6) autoregression for y_t. In general, $\dim\tilde{x}_t$ and p could be unknown and quite large relative to $\dim y_t$ and $\dim x_t$. Then the regression seeks to approximate the dynamics of a small subsystem y_t of the complex process $\tilde{x}_t$.

The following result is apparent.

<u>Corollary 10.1</u>. <u>If the regression of</u> y_t <u>on</u> x_t <u>is a subregression of</u> (10.10), <u>then</u> (AM3) <u>is satisfied</u>.

Our main result establishing (AM4) generalizes a result of Fuller and Hasza (1981) which concerned the more restricted situation of (stationary) univariate autoregressions. We now suppose that $\bar{x}_t$ is an $\bar{r}$–vector process satisfying

$$\bar{x}_t = A\bar{x}_{t-1} + \bar{e}_t, \tag{10.13}$$

with $\bar{e}_t \sim$ N.I.D.$(0,\bar{\Sigma})$, where, although $\bar{\Sigma}$ may be singular,

$$\sum_{j=0}^{\bar{r}} A^j \bar{\Sigma} (A^j)' \text{ is nonsingular.} \tag{10.14}$$

Full–rank autoregressions of order greater than one, such as (10.10), can be rewritten in the form (10.13) in such a way that (10.14) is satisfied. In Appendix II, we will prove the following result.

Proposition 10.2. If $\bar{x}_t$ is a process satisfying the conditions above, then for every k=1,2,..., there is an N(k) such that

$$\sup_{N \geq N(k)} E\left\{\lambda_{min}^{-k}(N^{-1} \sum_{t=1}^{N} \bar{x}_t \bar{x}_t')\right\} < \infty \ . \tag{10.15}$$

From an argument used to establish Corollary 10.1, we obtain our result for (AM4).

Corollary 10.2. If x_t is an r–dimensional subvector of a process satisfying (10.13) – (10.14), then (AM4) holds with $C_N = N^{1/2} I_r$.

In summary, since the matrices C_N must be the same in (AM1) and (AM4), the only stationary stochastic regressors for which we have completely verified (AM1) – (AM4) and (10.1) are the Gaussian subautoregressions. For non–stochastic regressors, see the Remark above.

Results analogous to (10.6) hold for K–L number differences and for $2\{\mathcal{E}_{N,\infty}^{(x)} - \hat{\mathcal{E}}_N^{(x)}\}$ and $2\{\hat{L}_N^{(x)} - L_{N,\infty}^{(x)}\}$ under the same assumptions The latter variates have the same asymptotic mean, $CVAR^{(x)} + \Sigma VAR^{(x)}$, where $\Sigma VAR^{(x)}$ is the trace of

the variance of the asymptotic distribution of $(N/2)^{1/2}(\mathrm{vec}\tilde{\Sigma}_N^{(x)} - I_q)$, with $\tilde{\Sigma}_N^{(x)} = \Sigma^{(x)-1/2}\Sigma_N^{(x)}(\Sigma^{(x)-1/2})'$. This follows from applying the Taylor expansion (I.2) to the bracketed term on the right in (I.6) of Appendix I. As a consequence, for weakly equivalent regressors ($|\Sigma^{(1)}| = |\Sigma^{(2)}|$), one obtains from (8.13) and from $E\{L_{N,\infty}^{(2)} - L_{N,\infty}^{(2)}\} = 0$ that

$$\lim_{N\to\infty} E\{-2\hat{L}_N^{(1,2)}\} = \lim_{N\to\infty} E\{2\hat{\mathcal{E}}_N^{(1,2)}\} =$$

$$\{CVAR^{(2)} + \Sigma VAR^{(2)}\} - \{CVAR^{(1)} + \Sigma VAR^{(1)}\}. \qquad (10.16)$$

Finally, we note that the conditions (AM3) and (AM4) are easily verified for some special non–Gaussian situations, such as the example at the end of section 5, where the processes e_t and x_t are bounded away from zero.

11. GENERALIZATIONS

In this section, we will briefly describe some elements of a natural conceptual framework for generalizations of the main results of the previous sections to model comparison problems different from regressor selection. For additional details, see Findley(1985) for time series models and Findley(1989) for models for independent observations. For an interesting application, see Ogata (1988). Suppose $L_N[\theta]$ denotes a log–likelihood function for N observations with parameter vector θ, having the property that $N^{-1}L_N[\theta]$ and also its first and second partial θ–derivatives converge in probability as $N\to\infty$ uniformly on compact subsets of the convex parameter space. The limit function $\mathcal{E}[\theta] \equiv \lim_{N\to\infty} E\{N^{-1}L_N[\theta]\}$ is a type of Kullback–Leibler number for the model defined by θ. $\mathcal{E}[\theta]$ is assumed to have a

unique maximum in the interior of the parameter space at a point θ_∞ where the matrix of second partial derivatives, $\mathcal{E}''[\theta]$, is non–singular. Then, under rather general circumstances, see Pollard(1985) or White(1989), maximum likelihood estimates $\hat{\theta}_N$ satisfying $\partial L_N[\hat{\theta}_N]/\partial\theta = 0$ will converge to θ_∞ in such a way that $N^{1/2}(\hat{\theta}_N - \theta_\infty)$ has a Gaussian limiting distribution. In this situation, the Taylor expansions

$$2\{L_N[\theta_\infty] - L_N[\hat{\theta}_N]\} = (\hat{\theta}_N - \theta_\infty)' L_N''[\bar{\theta}_N](\hat{\theta}_N - \theta_\infty)$$

and

$$2\{\mathcal{E}_N[\hat{\theta}_N] - \mathcal{E}_N[\theta_\infty]\} = (\hat{\theta}_N - \theta_\infty)' \mathcal{E}_N''[\tilde{\theta}_N](\hat{\theta}_N - \theta_\infty) ,$$

with $\bar{\theta}_N$ and $\tilde{\theta}_N$ on the line segment between $\hat{\theta}_N$ and θ_∞, motivate a generalization of the Q_N–statistic (4.4), namely

$$Q_N \equiv (\hat{\theta}_N - \theta_\infty)' L_N''[\theta_\infty](\hat{\theta}_N - \theta_\infty) ,$$

and lead to a generalization of the Akaike–Shimizu relation (8.11),

$$L_N[\hat{\theta}_N] - L_N[\theta_\infty] \sim_p N\{\mathcal{E}[\theta_\infty] - \mathcal{E}[\hat{\theta}_N]\} .$$

Two competing families of log–likelihoods $L_N^{(i)}[\theta^{(i)}]$ with these properties, with m.l.e.'s $\hat{\theta}_N^{(i)} \to_p \theta_\infty^{(i)}$, i=1,2, are said to be asymptotically equivalent if $L_N^{(1)}[\theta_\infty^{(1)}] = L_N^{(2)}[\theta_\infty^{(2)}]$ (w. p. 1) for $N \geq N_0$. Distributional results like those of section 7 can be obtained if twice differentiable, nonsingular parameter transformations $g^{(i)}$

exist such that $g^{(i)}(\theta^{(i)}) = [\eta^{c\prime}\eta^{(i)\prime}]'$, i=1,2, and if the model defined by the log–likelihood $L_N[\eta]$ with $\eta = [\eta^{c\prime}\eta^{(1)\prime}\eta^{(2)\prime}]'$ is asymptotically equivalent to those defined by the $L_N^{(i)}[\theta^{(i)}]$, i=1,2 . To establish results like

$$\lim_{N\to\infty} 2E\{L_N^{(i)}[\hat{\theta}_N^{(i)}] - L_N^{(i)}[\theta_\infty^{(i)}]\} = CVAR^{(i)} \qquad (11.1)$$

(as before, $CVAR^{(i)} = EQ^{(i)}$, where $Q_N^{(i)} \xrightarrow[dist.]{N} Q^{(i)}$) following the strategy of section 10, it is necessary to have explicit formulas for the m.l.e.'s $\hat{\theta}_N^{(i)}$. Also, conditional expectations must sometimes be used. In Findley (1989), the formula (11.1) is established for some models related to the multinomial distribution (histograms, contingency tables). For this analysis, the expectation operator E, when applied to $L_N[\theta]$, was taken to be the conditional expectation conditioned on cells with non–zero probability having at least one observation, in order to have $\mathcal{E}_N[\hat{\theta}_N] > -\infty$. We mention this to illustrate that there are a variety of ways, depending on the models under consideration, of filling in the theoretical structure outlined in this section. For the case of density models estimated from i.i.d. data, a formula for $CVAR^{(i)}$ can be obtained from Takeuchi (1976), see also Härdle (1987) and Findley (1989).

Shibata (1981) considers the case of fixed regressors, with $\dim y_t = 1$ and with $y_t - Ey_t$ being i.i.d. and Gaussian. He takes a very interesting and different approach from ours. He assumes that the correct regressor x_t has infinite dimension, but the not necessarily nested regressors $x_t^{(i)}$ under consideration are finite dimensional subvectors of x_t whose dimension increases with N (the range of i can increase also). For a modified version of the mean square prediction error criterion of section 9, he shows that MAIC is optimally efficient and that the strongly parsimonious criteria we discussed in section 6 are not. Härdle (1987) has considered

the extension to the case of i.i.d. non–Gaussian $y_t - Ey_t$, allowing the error density to be misspecified. The extension of Shibata's and Härdle's results to the case of stochastic regressors appears to be very difficult, see Shibata (1980), where increasing–order autoregressions are considered.

Finally, it should be mentioned that various cross–validation procedures for model selection are asymptotically equivalent to MAIC or simple variants thereof, see Stoica et al. (1986) and the references given there.

12. CONCLUDING REMARKS

Our goal in this paper has been to provide a coherent theory supporting the use of the ordinary log–likelihood ratio for making non–nested regressor comparisons. We were motivated to do this by the importance for applications of the non–nested comparison problem and by a desire to understand the substantial industrial successes of Akaike's MAIC procedure (which, for linear regressions, is asymptotically equivalent to the minimum FPEC criterion of Akaike(1971)). Some of these successes are described in Akaike and Nakagawa(1988), Nakamura et al.(1986), Otomo et al.(1972) and Ohtsu et al.(1979). (There are many industrial applications which are not publicly documented for company confidentiality reasons: Mr. K. Toki of System Sougou Kaihatsu in Tokyo kindly told one of the authors in 1987, in response to a query, that his company has implemented more than sixty statistical model–based controllers using the regressor selection procedures described here and in these references.) Akaike developed AIC as an asymptotically unbiased approximation to the Kullback–Leibler number, see This Week's Citation Classic (1981) and Akaike (1985). The results of sections 8 and 10 reveal attractive properties of K–L numbers for model comparison and clarify the nature of the connection with MAIC .

We hope that the results presented here will stimulate further research on non–nested model comparisons and the role of the likelihood ratio therein. It would be attractive to have generalizations of our results, or a reasonably comprehensive alternative theory, for the situation in which the number of variables in each regressor, $\dim x_t^{(i)}$, is permitted to increase as the sample size increases: there are circumstances in which the number of estimated variables must increase if the sequence of log–likelihood ratios, $\hat{L}_N^{(1,2)}$, $N \geq N_0$, is to be bounded in probability. (This is the relevent situation because, in practice, statisticians are only concerned about the interpretation of small–to–moderate values of $\hat{L}_N^{(1,2)}$). Such results should shed light on finite–sample properties.

*The authors wish to gratefully acknowledge the excellent computing support they received from E. Arahata and M. Pugh for the calculations presented in this paper. Some of the results presented here were obtained by the first–named author while he was a Visiting Professor at the Institute of Statistical Mathematics in Tokyo. He wishes to express his gratitude for the support and hospitality he received during this visit, especially from professor G. Kitagawa, who provided the data analyzed in section 3, along with insightful comments and advice.

REFERENCES

Akaike, H. (1971). "Autoregressive Model Fitting for Control," Annals of the Institute of Statistical Mathematics, 23, 163–180.

Akaike, H. (1973). "Information Theory and an Extension of the Likelihood Principle," in 2nd International Symposium on Information Theory, Eds. B. N. Petrov and F. Czaki pp. 267–287, Budapest: Akademia Kiado.

Akaike, H. (1977). "On Entropy Maximisation Principle," in Applications of Statistics (ed. P.R. Krishnaiah), Amsterdam: North Holland, 27–41.

Akaike, H., T. Ozaki, M. Ishiguro, Y. Ogata, G. Kitagawa, Y.-H. Tamura, E. Arahata, K. Katsura, and Y. Tamura (1985), TIMSAC 84, Part 2, Computer Science Monographs No. 23, Tokyo: Institute of Statistical Mathematics.

Akaike, H. (1985). "Prediction and Entropy," in A Celebration of Statistics (eds. A. C. Atkinson and S. E. Fienberg), New York: Springer Verlag, 1–24.

Akaike, H. and T. Nakagawa (1988). Statistical Analysis and Control of Dynamic Systems, Dordrecht: Kluver.

Anderson, T. W. (1971). The Statistical Analysis of Time Series, New York: Wiley.

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis, 2nd. Ed., New York: Wiley.

Bhansali, R. J. (1981). "Effects of Not Knowing the Order of an Autoregressive Process I." Journal of the American Statistical Association 76, 588–597.

Billingsley, P. (1985). Probability and Measure, 2nd Ed., New York: Wiley.

Brillinger, D. R. (1969). "Asymptotic Properties of Spectral Estimates of Second Order," Biometrika 56, 375–390.

Chan, N. H. and C. Z. Wei (1988). "Limiting Distributions of Least Squares Estimators of Unstable Autoregressive Processes," Annals of Statistics 16, 367–401.

Cox, D. R. (1961). "Tests of Separate Families of Hypotheses." Proceedings of the 4th Berkeley Symposium, 1, 105–123.

Cox, D. R. (1962). "Further Results on Tests of Separate Families of Hypotheses." Journal of the Royal Statistical Society, Series B 24, 406–424.

Dahlhaus, R. (1985). "A Functional Central Limit Theorem for Tapered Empirical Spectral Functions," Stochastic Processes and Their Applications 19, 135–149.

Davies, R. B. (1980). "AS155. The distribution of a linear combination of chi-squared random variables," Applied Statistics, 323–333.

Eberlein, E. (1986). "On Strong Invariance Principles under Dependence Assumptions," Annals of Probability, 14, 260–270.

Findley, D. F. (1985). "On the Unbiasedness Property of AIC for Exact or Approximating Linear Stochastic Time Series Models," Journal of Time Series Analysis, 6, 229–252.

Findley, D. F. (1988). "An Analysis of AIC for Linear Stochastic Regression and Control," 1988 American Control Conference, Piscataway: IEEE, 1281–1288.

Findley, D. F. (1989). "Beyond Chi–Square: Likelihood Ratio Procedures for Comparing Models for Independent Observations, Including Conditional Models," in preparation.

Franke, J., Th Gasser and H. Steinberg, (1985). "Fitting Autoregressive Processes to EEG Time Series: An Empirical Comparison of Estimates of Order," IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP–33, 143–150.

Fuller, W. A. and D. P. Hasza (1981). "Properties of Predictors for Autoregressive Time Series," Journal of the American Statistical Association 75, 155–161.

Hall, P. and C. C. Heyde (1980). Martingale Limit Theory and Its Application, New York: Academic Press.

Hannan, E. J. (1970). Multiple Time Series, New York: Wiley.

Hannan, E. J. and B. Quinn (1979). "The Determination of the Order of an Autoregression," Journal of the Royal Statistical Society, Series B 41, 190–195.

Härdle, W. (1987). "An Effective Selection of Regression Variables When the Error Distribution is Incorrectly Specified," Annals of the Institute of Statistical Mathematics, 39, 533–548.

Hosoya, Y. and M. Taniguchi (1982). "A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes," Annals of Statistics 10, 132–153.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl and T.–C. Lee (1984). Theory and Practice of Econometrics, 2nd Ed., New York: Wiley.

Kitagawa, G. (1987), "Rejoinder," Journal of the American Statistical Association 82, 1060–1063.

Kitagawa, G. and K. Ohtsu (1976). "The Statistical Control of Ship's Course Keeping Motion," Proceedings of the Institute of Statistical Mathematics, 23,105–128. (in Japanese)

Kunitomo, N. and T. Yamamoto (1985). "Properties of Predictors in Misspecified Autoregressive Time Series Models," Journal of the American Statistical Association 80, 941–950.

Lai, T. L. and C. Z. Wei (1982). "Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems," Annals of Statistics, 10, 154–166.

Lai, T. L. and C. Z. Wei (1985). "Asymptotic Properties of Multivariate Weighted Sums with Applications to Stochastic Regression in Linear Dynamic Systems," in Multivariate Analysis VI. ed. P. R. Krishnaiah, Amsterdam: North Holland, pp. 375 – 393.

McCullagh, P. (1987). Tensor Methods in Statistics, London: Chapman and Hall.

McLeish, D. L. (1975). "Invariance Principles for Dependent Variables," Zeitschrift für Wahrscheinlichkeitstheorie, 32, 165–178.

Nakamura, H., M. Uchida, Y. Toyota and M. Kushihashi (1986). "Optimal Control of Thermal Power Plants," ASME Winter Annual Meetings Proceedings, 86–4A/DSC–14.

Noble, B. (1969). Applied Linear Algebra, Englewood Cliffs: Prentice–Hall.

Ogata, Y. (1988). "Statistical Models for Earthquake Occurrances and Residual Analysis for Point Processes," Journal of the American Statistical Association 83, 9–27.

Ohtsu, K., M. Horigome and G. Kitagawa (1979). "A New Ship's Autopilot Design Through a Stochastic Model," Automatica, 15, 255–268.

Otomo, T., T. Nakagawa and H. Akaike (1972). "Statistical Approach to Computer Control of Cement Rotary Kilns," Automatica, 8, 35–48.

Pollard, D. (1985). "New Ways to Prove Central Limit Theorems," Econometric Theory 1, 295–313.

Rissanen, J. (1986). "Stochastic Complexity and Modeling," Annals of Statistics 14, 1080–1100.

Schwarz, G. (1978). "Estimating the Dimension of a Model," Annals of Statistics 6, 461–464.

Shibata, R. (1980). "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," Annals of Statistics 8, 147–164.

Shibata, R. (1981). "An Optimal Selection of Regression Variables," Biometrika 68, 45–54 and Biometrika 69, 494 (correction).

Shimizu, R. (1978). "Entropy Maximization Principle and Selection of the Order of an Autoregressive Gaussian Process," Annals of the Institute of Statistical Mathematics 30, 263–270.

Stoica, P. and P. Eykhoff, P. Janssen. T. Söderström (1986). "Model Structure by Cross–Validation," International Journal of Control, 43, 1841–1878.

Takada, Y. (1982). "Admissibility of Some Variable Selection Rules in the Linear Regression Model," Journal of the Japan Statistical Society 12, 45–49.

Takeuchi,K. (1976). "The Information Statistic of a Distribution and Criteria for Model Fitting," Mathematical Sciences (Sūri–Kagaku), 14, 14–18. (In Japanese)

This Week's Citation Classic (1981). Current Contents 51, 22.

Tsurumi, H. and Wago, H. (1987). "Mean Square Errors of Forecast for Selecting Non-Nested Linear Models and Comparison with Other Criteria," Journal of Econometrics, to appear.

White, H. (1989). Estimation, Inference and Specification Analysis. New York: Cambridge University Press.

Yokoyama, R. (1980). "Moment Bounds for Stationary Mixing Sequences," Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 52, 45–47.

Zhurbenko, I. G. (1986). The Spectral Analysis of Time Series, Amsterdam: North Holland.

APPENDIX I: PROOFS FOR SECTIONS 6 – 10.1.

We begin with the arguments leading to Propositions 6.3 and the results (10.3) and (8.11), which concern $\hat{L}_N^{(1,2)}$, $\hat{L}_N^{(x)} - L_{N,\infty}^{(x)}$ and $\mathcal{E}_{N,\infty}^{(x)} - \hat{\mathcal{E}}_N^{(x)}$. These quantities are unchanged by the transformation $e_t^{(x)} \to (\Sigma^{(x)})^{-1/2} e_t^{(x)}$, so we shall assume that $\Sigma^{(x)} = I_q$ $(= \Sigma^{(1)}, \Sigma^{(2)}$ etc.). Then the assumption (A2) implies that the eigenvalues $\hat{\lambda}_{N,1}^{(x)},\ldots,\hat{\lambda}_{N,q}^{(x)}$ of $\hat{\Sigma}_N^{(x)}$ converge in probability to 1 at the rate $N^{-1/2}$,

$$N^{1/2}(\hat{\lambda}_{N,j}^{(x)} - 1) \sim O_p(1), \ (1 \le j \le q). \tag{I.1}$$

In general, if Σ is a positive definite matrix of order q with eigenvalues $\lambda_1,\ldots,\lambda_q$, then it follows from Taylor's formula that there exists an $\alpha = \alpha(\lambda_1,\ldots,\lambda_q)$ between 0 and 1 such that, with $\tilde{\lambda}_j = 1 + \alpha(\lambda_j - 1)$, $1 \le j \le q$, we have

$$\sum_{j=1}^{q} \log\lambda_j = \sum_{j=1}^{q} (\lambda_j - 1) - \frac{1}{2} \sum_{j=1}^{q} \tilde{\lambda}_j^{-2} (\lambda_j - 1)^2 \quad . \tag{I.2}$$

From (I.1) and (I.2), we obtain

$$N\log|\hat{\Sigma}_N^{(x)}| \sim_p N\mathrm{tr}(\hat{\Sigma}_N^{(x)} - I_q) - (N/2)\mathrm{tr}(\hat{\Sigma}_N^{(x)} - I_q)^2 . \qquad (I.3)$$

Proof of Proposition 6.3: Set $e_t \equiv e_t^{(1)} = e_t^{(2)}$ and $\Sigma_N \equiv N^{-1}\Sigma_{t=1}^N e_t e_t'$. Observe via the identity (4.3) that $N\mathrm{tr}(\Sigma_N - \hat{\Sigma}_N^{(i)}) = Q_N^{(i)}$, so that

$$N\mathrm{tr}(\hat{\Sigma}_N^{(2)} - \hat{\Sigma}_N^{(1)}) = Q_N^{(1)} - Q_N^{(2)} . \qquad (I.4)$$

It follows from assumption (A1) that $N\mathrm{tr}(\hat{\Sigma}_N^{(2)} - \hat{\Sigma}_N^{(1)})$ is bounded in probability. Consequently,

$$N\mathrm{tr}(\hat{\Sigma}_N^{(1)} - I_q)^2 - N\mathrm{tr}(\hat{\Sigma}_N^{(2)} - I_q)^2 =$$

$$N\mathrm{tr}[(\hat{\Sigma}_N^{(1)} - \hat{\Sigma}_N^{(2)})\{(\hat{\Sigma}_N^{(1)} - I_q) + (\hat{\Sigma}_N^{(2)} - I_q)\}]$$

converges to zero in probability, and Proposition 6.3 follows immediately from (I.3) and (I.4).

Proof of (10.3). If $0 \le \alpha \le 1$ and $\lambda \ge 0$, then $(1 + \alpha(\lambda-1))^{-2} \le \lambda^{-2} + 1$. Hence, the remainder term in the expansion (I.2) of $\log|\hat{\Sigma}_N^{(i)}|$ has the upper bound

$$\sum_{j=1}^{q} \left\{\left[\hat{\lambda}_{N,j}^{(i)}\right]^{-2} + 1\right\}\left\{\hat{\lambda}_{N,j}^{(i)} - 1\right\}^2$$

$$= \mathrm{tr}\left\{\hat{\Sigma}_N^{(i)-2} + I_q\right\}\left\{\hat{\Sigma}_N^{(i)} - I_q\right\}^2. \qquad (I.5)$$

The inequality (10.3) follows easily from (I.5) and the expansion (I.2) of $\log|\hat{\Sigma}_N^{(2)}|$ and $\log|\hat{\Sigma}_N^{(1)}|$.

Proof of (8.11). Using (4.3) and $\Sigma^{(x)} = I_q$, one sees that

$$(-2)\left\{\hat{L}_N^{(x)} - L_{N,\infty}^{(x)}\right\} = N\left\{\log|\hat{\Sigma}_N^{(x)}| - \mathrm{tr}\left[\hat{\Sigma}_N^{(x)} - I_q\right]\right\} - Q_N^{(x)}, \quad \text{(I.6)}$$

and, from (8.1), that

$$(-2)\left\{\mathcal{E}_{N,\infty}^{(x)} - \hat{\mathcal{E}}_N^{(x)}\right\} = N\left\{\log|\hat{\Sigma}_N^{(x)-1}| - \mathrm{tr}[\hat{\Sigma}_N^{(x)-1} - I_q]\right\} - R_N^{(x)}, \quad \text{(I.7)}$$

with $Q_N^{(x)}$ defined by (4.4) and $R_N^{(x)}$ defined as in (8.9). Since $Q_N^{(x)} \sim_p R_N^{(x)}$, see section 8, it remains to show the asymptotic coincidence of the terms in curly brackets in (I.6) and (I.7). From the expansions (I.2) for $\hat{\Sigma}_N^{(x)}$ and $\hat{\Sigma}_N^{(x)-1}$, we deduce that the bracketed terms differ by

$$N\sum_{j=1}^{q}\left[\left\{1 - (\hat{\lambda}_{N,j}^{(x)})^{-1}\right\}^2(\bar{\lambda}_{N,j}^{(x)})^2 - \left\{1 - \hat{\lambda}_{N,j}^{(x)}\right\}^2(\tilde{\lambda}_{N,j}^{(x)})^{-2}\right]$$

$$= \sum_{j=1}^{q} N(1 - \hat{\lambda}_{N,j}^{(x)})^2\left\{(\bar{\lambda}_{N,j}^{(x)}/\hat{\lambda}_{N,j}^{(x)})^2 - (\tilde{\lambda}_{N,j}^{(x)})^{-2}\right\}, \quad \text{(I.8)}$$

where $\bar{\lambda}_{N,j}^{(x)}$ and $\tilde{\lambda}_{N,j}^{(x)}$ are between $\hat{\lambda}_{N,j}^{(x)}$ and 1, for j=1,..,q. Since the factors in curly brackets on the right hand side of (I.8) tend to zero in probability while their multipliers $N(1-\hat{\lambda}_{N,j}^{(x)})^2$ are bounded, by (I.1), both sides of (I.8) tend to 0, and (8.11) follows.

Now we turn to the proof required for section 7.

Proof of Proposition 7.2: Let $V_t^{(i)}$ denote the vector space of linear combinations of the $r^{(i)}$ entries of $x_t^{(i)}$, i = 1,2. If $V_t \equiv V_t^{(1)} \cap V_t^{(2)}$ has dimension r^c, let $x_{t,j}^c$, j = 1,..., r^c denote a basis for V_t and let $z_{t,j}^{(i)}$, j = 1,...,$r^{(i)} - r^c$ denote a basis for the orthogonal complement of V_t in $V_t^{(i)}$, orthogonal in the sense of the inner product defined by covariance. By stationarity, the coefficients of the linear combinations of the entries of $x_t^{(i)}$ used to produce these bases can be chosen independently of t. If we do this, and define $x_t^c = [x_{t,1}^c,...,x_{t,r}^c c]'$ and $z_t^{(i)} = [z_{t,1}^{(i)},...,z_{t,r}^{(i)}(i)_{-r}c]'$, then the assertions of Proposition 7.2 follow easily.

APPENDIX II: PROOFS FOR SUBSECTION 10.2

Proof of Lemma 10.3: The proof involves a sequence of reductions to simpler cases. Reduction to the case N=mn, n=1,2,..., for any fixed m. Given m, we choose n so that $(m-1)n \leq N \leq mn$. With $W_N = \Sigma_{t=1}^N \epsilon_t \epsilon_t'$ as in the subsection, we have $\lambda_{min}(W_{(m-1)n}) \leq \lambda_{min}(W_N) \leq \lambda_{min}(W_{mn})$, so

$$(m-1)n\lambda_{min}^{-1}(W_{mn}) \leq N\lambda_{min}^{-1}(W_N) \leq mn\lambda_{min}^{-1}(W_{(m-1)n})$$

holds. This reveals that (10.9) will follow provided we can show that the sequence $n^k E\lambda_{min}^{-k}(W_{mn})$, n = 1, 2,..., is bounded whenever $m \geq h + 2k$.

Reduction to the case n = 1. First, we observe that if $M_1,...,M_n$, are positive definite matrices of order h, then from the arithmetic–geometric mean inequality, for any h–vector x,

$$n^{-1}x'(\sum_{j=1}^{n} M_j)x \geq \prod_{j=1}^{n} (x'M_jx)^{1/n},$$

which implies that

$$n\lambda_{min}^{-1}(\sum_{j=1}^{n} M_j) \leq \prod_{j=1}^{n} \lambda_{min}^{-1/n}(M_j). \qquad \text{(II.1)}$$

Now set $M_j=\Sigma_{t=m(j-1)+1}^{mj}\epsilon_t\epsilon_t'$, so that $M_j \sim W_h(\Sigma,m)$ and $W_{mn}=\Sigma_{j=1}^{n}M_j$. Then from (II.1),

$$n^k E\lambda_{min}^{-k}(W_{mn}) \leq E\prod_{j=1}^{n} \lambda_{min}^{-k/n}(M_j)$$

$$\leq \prod_{j=1}^{n} \{E\lambda_{min}^{-k}(M_j)\}^{1/n} \qquad \text{(Hölder's inequality)}$$

$$= E\lambda_{min}^{-k}(M_1) \qquad \text{(identical distribution).}$$

Reduction to the case $\Sigma=I_h$. Let $\Sigma^{1/2}$ denote a symmetric square root of Σ. Then $\Sigma^{-1/2}W_m\Sigma^{-1/2} \sim W_h(I_h,m)$, and we will show that

$$\lambda_{min}^{-1}(W_m) \leq \lambda_{min}^{-1}(\Sigma^{-1/2}W_m\Sigma^{-1/2})\lambda_{min}^{-1}(\Sigma) \; . \qquad \text{(II.2)}$$

In fact, for any h–vector x,

$$x'W_mx = (\Sigma^{1/2}x)'\Sigma^{-1/2}W_m\Sigma^{-1/2}(\Sigma^{1/2}x)$$

$$\geq \lambda_{min}(\Sigma^{-1/2}W_m\Sigma^{-1/2'})x'\Sigma x \ ,$$

so that $\lambda_{min}(W_m) \geq \lambda_{min}(\Sigma^{-1/2}W_m\Sigma^{-1/2})\lambda_{min}(\Sigma)$, from which (II.2) follows. Our proof of (10.9) will therefore be complete if we verify (II.3):

$$\underline{\text{If}}\ W_m \sim W_h(I_h,m)\ \underline{\text{and}}\ m \geq h+2k,\ \underline{\text{then}}\ E\lambda_{min}^{-k}(W_m) < \infty \ . \quad \text{(II.3)}$$

In fact, the density function of the joint distribution of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_h$ (see formula (11) of Anderson (1984, p. 534)) is bounded above by

$$C_{h,m} \prod_{i=1}^{h} \lambda_i^{(m+h-2i-1)}\exp(-\lambda_i/2)$$

for some constant $C_{h,m}$. Hence, for some constant $\overline{C}_{h,m}$,

$$E\lambda_{min}^{-k}(W_m) \leq \overline{C}_{h,m}\int_0^{\infty}\lambda_h^{(m-h-1-2k)/2}\exp(-\lambda_h/2)d\lambda_h \ .$$

The exponent of λ_h in this integral is greater than -1 when $m \geq h+2k$, so the integral is finite, and (II.3) follows. This shows that (10.9) holds.

Proof of Proposition 10.1: Set $\tilde{S}_N \equiv N\hat{\Sigma}_N$. Then $\lambda_{min}(\tilde{S}_N) = N\lambda_{min}(\hat{\Sigma}_N)$. Our proof of (10.12) is based mainly on an elaboration of an argument sketched in Appendix I of Fuller and Hasza (1981). Let the sample residual variance matrix of the least squares regression of $\tilde{x}_{m(p+1)}$ on $\tilde{x}_{m(p+1)\pm1},\ldots,\tilde{x}_{m(p+1)\pm p}$, $m=1,\ldots,n$, be denoted by $n^{-1}\tilde{S}_{n(p+1)}$. Then

$$\bar{S}_{n(p+1)} \leq \sum_{m=1}^{n} (\tilde{x}_{m(p+1)} - \hat{A}_1 \tilde{x}_{m(p+1)-1} - \cdots - \hat{A}_p \tilde{x}_{m(p+1)-p}) \cdot$$

$$(\tilde{x}_{m(p+1)} - \hat{A}_1 \tilde{x}_{m(p+1)-1} - \cdots - \hat{A}_p \tilde{x}_{m(p+1)-p})' \leq \tilde{S}_{n(p+1)} \cdot$$

Thus, to establish (10.12), it is sufficient to prove that for any $k \geq 1$, there is an $n(k)$ such that

$$\sup_{n \geq n(k)} n^k E \lambda_{min}^{-k}(\bar{S}_{n(p+1)}) < \infty. \tag{II.4}$$

We shall establish the existence of coefficient matrices C_j, $j = \pm 1, \ldots \pm p$, such that, conditional on $\tilde{x}_{m(p+1)\pm 1}, \ldots, \tilde{x}_{m(p+1)\pm p}$, $m = 1, \ldots, n$, the random variables $\tilde{e}_m$ defined by

$$\tilde{e}_m \equiv \tilde{x}_{m(p+1)} - \sum_{|j|=1}^{p} C_j \tilde{x}_{m(p+1)-j} \tag{II.5}$$

are N.I.D.$(0, \tilde{\Sigma})$ for some $\tilde{\Sigma} > 0$ which does not depend on the values of the conditioning variables. This result, which follows from Lemma II.1 below, implies that for $n > 2p\tilde{r}$, the conditional distribution of $\bar{S}_{n(p+1)}$ given $u_m = (\tilde{x}'_{m(p+1)-p}, \ldots, \tilde{x}'_{m(p+1)-1})'$, $m = 1, \ldots, n+1$, is $W_{\tilde{r}}(\tilde{\Sigma}, n - 2p\tilde{r})$, by Theorem 8.22 of Anderson (1984). Therefore, from (10.9), there are constants C and $n(k)$ such that the conditional $(-k)$-th moments of $\lambda_{min}(\bar{S}_{n(p+1)})$ satisfy

$$\sup_{n \geq n(k)} n^k E\{\lambda_{min}^{-k}(\bar{S}_{n(p+1)}) | u_1, \ldots, u_{n+1}\} \leq C. \tag{II.6}$$

Since

$$E\{\lambda_{min}^{-k}(S_{n(p+1)})\} = E_{u_1,\ldots,u_{n+1}}(E\{\lambda_{min}^{-k}S_{n(p+1)})|u_1,\ldots,u_{n+1}\}),$$

property (II.4) follows from (II.6).

We return to the discussion of (II.5) to complete the proof. Let g(a) denote the density of $a \sim \mathcal{N}(0,\Sigma)$. The result needed is the following lemma.

Lemma II.1. The conditional distribution of $z_m = \tilde{x}_{m(p+1)}$, $1 \le m \le n$, given $u_1,\ldots,u_{n+1}$, has the form

$$f(z_1,\ldots,z_n|u_1,\ldots,u_{n+1}) = \prod_{m=1}^{n} \{h(z_m,u_m,u_{m+1})/\int_{\mathbb{R}^r} h(z_m,u_m,u_{m+1})dz_m\}, \qquad (II.7)$$

where $h(z_m,u_m,u_{m+1})$ is the p-fold product function defined by $h(z_m,u_m,u_{m+1}) \equiv$

$$g(z_m - A_1\tilde{x}_{m(p+1)-1} \cdots - A_p\tilde{x}_{m(p+1)-p}) \cdots g(\tilde{x}_{(m+1)(p+1)-1} - A_1\tilde{x}_{(m+1)(p+1)-2} \cdots - A_p z_m). \qquad (II.8)$$

Proof: Let $f(\tilde{x}_{1-p},\ldots,\tilde{x}_0)$ denote the joint density of $\tilde{x}_{1-p},\ldots\tilde{x}_0$. Then the joint density of $\tilde{x}_{1-p},\ldots,\tilde{x}_{(n+1)(p+1)-1}$ is

$$f(\tilde{x}_{1-p},\ldots,\tilde{x}_0)g(\tilde{x}_1 - \sum_{j=1}^{p} A_j\tilde{x}_{1-j}) \cdots g(\tilde{x}_{(n+1)(p+1)-1} - \sum_{j=1}^{p} A_j\tilde{x}_{(n+1)(p+1)-1-j}).$$

So, with the function K_0 defined by

$$K_0 \equiv \int_{R^{p\tilde{r}}} f(\tilde{x}_{1-p},\ldots,\tilde{x}_0)g(\tilde{x}_1-\sum_{j=1}^{p} A_j\tilde{x}_{1-j})\cdots g(\tilde{x}_p-\sum_{j=1}^{p} A_j\tilde{x}_{p-j})d\tilde{x}_{1-p}\cdots d\tilde{x}_0,$$

the joint density of $z_1,\ldots,z_n,u_1,\ldots,u_{n+1}$ is given by

$$f(z_1,\ldots,z_n,u_1,\ldots u_{n+1}) \equiv K_0 \cdot \prod_{t=p+1}^{(n+1)(p+1)-1} g(\tilde{x}_t - \sum_{j=1}^{p} A_j\tilde{x}_{t-j}). \qquad \text{(II.9)}$$

Integrating over $z_1,\ldots z_n$, we obtain the joint density of $u_1,\ldots,u_{n+1}$,

$$f(u_1,\ldots,u_{n+1}) = K_0 \cdot \prod_{m=1}^{n} \{\int_{R^{\tilde{r}}} h(z_m,u_m,u_{m+1})dz_m\},$$

and the assertation of (II.7) follows from dividing (II.9) by this expression.

By adding the exponents of the $N(0,\Sigma)$ density g–functions in (II.8), one sees from (II.7) that, conditioned on $u_1,\ldots u_{n+1}$, the random variables z_m are independently normally distributed with means of the form $\sum_{|j|=1}^{p} C_j\tilde{x}_{m(p+1)-j}$ and nonsingular variance matrix $\tilde{\Sigma} = \{\Sigma^{-1} + \sum_{j=1}^{p} A_j\Sigma^{-1}A_j'\}^{-1}$. Hence, the $\tilde{e}_m$, $1\leq m\leq n$ of (II.5) are N.I.D.$(0,\tilde{\Sigma})$, which is the result needed to complete the verification of (10.12).

Proof of Proposition 10.2: We shall obtain (10.15) from the special case (10.9). Let $\det(A-\lambda I_p)=\lambda^{\bar{r}}+a_1\lambda^{\bar{r}-1}+\cdots+a_{\bar{r}}$ and $\tilde{z}_t=\bar{x}_t+\Sigma_{j=1}^{\bar{r}}a_j\bar{x}_{t-j}$. Then $\tilde{z}_t=\tilde{e}_t+(A+a_1I_{\bar{r}})\tilde{e}_{t-1}+\cdots+(A^{\bar{r}-1}+a_1A^{\bar{r}-2}+\cdots+a_{\bar{r}-1}I_{\bar{r}})\tilde{e}_{t-\bar{r}+1}$, see Lai and Wei (1985), equation (3.17). Clearly $\tilde{z}_t \sim N(0,\Sigma)$, with

$$\Sigma = \bar{\Sigma}+(A+a_1 I_{\bar{r}})\bar{\Sigma}(A+a_1 I_{\bar{r}})' +\cdots+(A^{\bar{r}-1}+\cdots+a_{\bar{r}-1}I_{\bar{r}})\bar{\Sigma}(A^{\bar{r}-1} +\cdots+a_{\bar{r}-1}I_{\bar{r}})'.$$

It follows from (10.14) that Σ is nonsingular, see Lai and Wei (1985, p. 381). Let $a_0=1$. Then, by (3.20) of the same reference,

$$\lambda_{\min}\Big(\sum_{t=\bar{r}+1}^{N} \tilde{z}_t\tilde{z}_t'\Big) \le \bar{r}\Big(\sum_{j=0}^{\bar{r}} a_j^2\Big)\lambda_{\min}\Big(\sum_{t=1}^{N} \bar{x}_t\bar{x}_t'\Big) . \qquad \text{(II.10)}$$

Consider the time series e_t obtained by observing every $\bar{r}$–th value of $\tilde{z}_t$: $e_t \equiv \tilde{z}_{\bar{r}t}$, t=1, 2,... . Then e_t is N.I.D.$(0,\Sigma)$ and

$$\lambda_{\min}\Big(\sum_{t=2}^{N} e_t e_t'\Big) \le \lambda_{\min}\Big(\sum_{t=\bar{r}+1}^{N\bar{r}} \tilde{z}_t\tilde{z}_t'\Big) . \qquad \text{(II.11)}$$

Now (10.15) follows immediately from (II.10), (II.11), and (10.9) via a reduction argument of the sort used at the start of the proof of Lemma 10.3 in Appendix I.