

No. 316

A Predictive Density Criterion for Selecting
Non-Nested Linear Models and Comparison with
Other Criteria

by

Hiroki Tsurumi and Hajime Wago

July 1986

A Predictive Density Criterion for Selecting
Non-Nested Linear Models and Comparison with
Other Criteria

Hiroki Tsurumi
Rutgers University
and
Hajime Wago
Tsukuba University

July 1986

I. Introduction

The mean squared-errors of forecasts (MSEF) is a statistic used to evaluate post-sample prediction performance. The MSEF has been used as a descriptive measure, but its exact distribution can be derived either from a sample theoretical or from a Bayesian perspective if the MSEF is computed from a linear regression model. In this paper, Bayesian and sampling distributions of the MSEF are derived, and it is suggested that the MSEF may be used as a statistic for linear model selection. Using sampling experiments, we compare the MSEF criterion with other model selection criteria. The organization of the paper is as follows. In section 2, we give the Bayesian and sampling distributions of the MSEF. In section 3, after presenting Akaike's information criterion, AIC, [Akaike (1974)], Efron's confidence interval for the mean squared errors, the N- and J-tests, we make sampling experiments to compare the Bayesian MSEF criterion with these other criteria.

2. Bayesian and Sampling Distributions of the MSEF

Let the linear model be given by

$$y = X\beta + u, \quad (1)$$

where y is an $(n \times 1)$ vector of observations on the dependent variable, X is an $(n \times k)$ matrix of observations on the explanatory variables with rank k , u is an $(n \times 1)$ vector of error terms, and β is a $(k \times 1)$ vector of unknown regression coefficients. Assume that $u \sim N(0, \sigma^2 I_n)$ and that β is estimated by $\hat{\beta} = (X'X)^{-1}X'y$.

The mean-squared-error for the post-sample period, $n+1, \dots, n+m$ is computed using the post-sample actual observations on y and X . Let y_* and

X_* be, respectively, an $(m \times 1)$ vector and an $(m \times k)$ matrix of post-sample observations and assume that the rank of X_* is $\min(m, k)$. Then the MSEF is

$$MSEF = \frac{1}{m}(\hat{y}_* - y_*)'(\hat{y}_* - y_*), \quad (2)$$

where $\hat{y}_* = X_*\hat{\beta}$. Given equation (1) and $\hat{\beta} = \beta + (X'X)^{-1}X'u$, equation (2) can be written as

$$MSEF = \frac{1}{m} \epsilon_*' B' B \epsilon_* = \frac{1}{m} \sum_{i=1}^m \mu_i \epsilon_i^2, \quad (3)$$

where: $\epsilon_* = (u', u_*')'$, $B = (A, -I_m)$, $A = X_*(X'X)^{-1}X'$, and the μ_i 's are the nonzero characteristic roots of $B'B$. The ϵ_i are elements of $\epsilon = c'\epsilon_*$, where c is the matrix of characteristic vectors of $B'B$. In passing, let us note that the μ_i 's are given by $\mu_i = 1 + \lambda_i$, $i = 1, \dots, m$ for $m \leq k$, and $\mu_i = 1 + \lambda_i$, $i = 1, \dots, k$; $\mu_i = 1$, $i = k+1, \dots, m$, for $m > k$, where λ_i is the i th nonzero characteristic root of AA' .

Since $\epsilon_i \sim NID(0, \sigma^2)$, $m \cdot MSEF$ is a quadratic form in normal variables. The distribution of quadratic forms or ratios of quadratic forms has been investigated by many; some of the earlier works are by McCarthy (1939), von Neumann (1941), and Bhattacharyya (1943). Bhattacharyya (1954) and Hotelling (1948) employed Laguerre expansion, and Gurland (1953) and Johnson and Kotz (1970) refined further the convergent Laguerre expansions. In this paper we use the degenerate hyperbolic function, which is convenient for computational purposes. Theorem 1 below summarizes the derivation.

After deriving the distribution of the MSEF using the degenerate hyperbolic function, we shall use the Laguerre expansions just for comparison sake.

Theorem 1: Let $x = m \cdot \text{MSEF} / \sigma^2$. Then the distribution of x is given by

$$f(x) = \frac{e^{-x/2\mu_m} x^{\frac{1}{2}m-1}}{2^{\frac{1}{2}m} \pi^{\frac{1}{2}} \prod_{i=1}^m \mu_i^{\frac{1}{2}}} \sum_{p=0}^{\infty} c(m,p) x^p \quad (4)$$

where $c(m,p)$ is the recursive coefficient given by

$$c(m,p) = \frac{\Gamma(p + \frac{m-1}{2})}{\Gamma(p + \frac{m}{2})} \sum_{j=0}^p \frac{c(m-1,j) a_m^{p-j}}{(p-j)!}, \quad \text{for } m \geq 2, \quad (5)$$

and $c(1,0)=1$, $c(1,j)=0$ for $j \geq 1$; $a_m = \frac{1}{2} \left(\frac{1}{\mu_m} - \frac{1}{\mu_{m-1}} \right)$, $\mu_1 \geq \mu_2 \geq \dots \geq \mu_m$, for $m \geq 2$, $a_1 = \frac{1}{2\mu_1}$, and $a_i^0 = 1$ for all $i=1, \dots, m$.

Proof: We shall prove the theorem by induction. For $m=1$, $x = \mu_1 z_1^2 / \sigma^2$, and $z_1^2 / \sigma^2 \sim \chi_1^2$. Hence, we have

$$f(x) = \frac{1}{\sqrt{2\pi\mu_1}} x^{-\frac{1}{2}} e^{-x/2\mu_1} \quad (6)$$

which is equal to equation (4) for $m=1$. Assume that equation (5) holds for $m=q$. Then $m=q+1$, we have $x = z_q^2 + z_{q+1}^2$, where z_q^2 has the probability density function (pdf) of equation (5) with $m = q$, and

z_{q+1} has the pdf $\frac{1}{\sqrt{2\pi\mu_{q+1}}} x^{-1/2} e^{-x/2\mu_{q+1}}$. Since z_q and z_{q+1} are independent, the distribution of x is given by

$$f(x) = \int_0^x \left[\frac{e^{-t/\mu_q} t^{q/2-1}}{2^{q/2} \sqrt{\pi} \prod_{i=1}^q \mu_i^{1/2}} \sum_{p=0}^{\infty} c(q,p) t^p \right] \frac{1}{\sqrt{2\pi\mu_{q+1}}} (x-t)^{-1/2} e^{-(x-t)/2\mu_{q+1}} dt \quad (7)$$

$$= \frac{e^{-x/2\mu_{q+1}}}{2^{(q+1)/2} \pi \prod_{i=1}^{q+1} \mu_i^{1/2}} \sum_{p=0}^{\infty} c(q,p) \int_0^x t^{p+q/2-1} (x-t)^{-1/2} e^{-a_{q+1}t} dt$$

Since $\int_0^x t^{p+q/2-1} (x-t)^{-1/2} e^{-a_{q+1}t} dt = B(\frac{1}{2}, p + \frac{q}{2}) x^{p + \frac{q+1}{2} - 1}$

${}_1F_1(p + \frac{q}{2}, p + \frac{q+1}{2}, a_{q+1}x)$, where ${}_1F_1(\alpha, \gamma, z) = \frac{\Gamma(\gamma)}{\Gamma(\alpha)} \sum_{j=0}^{\infty} \frac{\Gamma(\alpha+j)}{j! \Gamma(\gamma+j)} z^j$

is the degenerate hyperbolic function and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

[Gradshteyn and Ryzhik (1980), p.318], we have

$$f(x) = \frac{e^{-x/2\mu_{q+1}} x^{\frac{1}{2}(q+1)-1}}{2^{\frac{1}{2}(q+1)} \sqrt{\pi} \prod_{i=1}^{q+1} \mu_i^{1/2}} \sum_{p=0}^{\infty} c(q,p) \sum_{j=0}^{\infty} \frac{\Gamma(p + \frac{1}{2}q + j)}{j! \Gamma(p + \frac{q+1}{2} + j)} a_{q+1}^j x^{p+j} \quad (8)$$

$$= \frac{e^{-x/2\mu_{q+1}} x^{\frac{1}{2}(q+1)-1}}{2^{\frac{1}{2}(q+1)} \sqrt{\pi} \prod_{i=1}^{q+1} \mu_i^{\frac{1}{2}}} \sum_{p=0}^{\infty} \frac{\Gamma(p + \frac{q}{2})}{\Gamma(p + \frac{q+1}{2})} \sum_{j=0}^p \frac{c(q,j) a_{q+1}^{p-j}}{(p-j)!} x^p$$

$$= \frac{e^{-x/2\mu_{q+1}} x^{\frac{1}{2}(q+1)-1}}{2^{\frac{1}{2}(q+1)} \sqrt{\pi} \prod_{i=1}^{q+1} \mu_i^{\frac{1}{2}}} \sum_{p=0}^{\infty} c(q+1,p) x^p$$

where $c(q+1,p) = \frac{\Gamma(p + \frac{q}{2})}{\Gamma(p + \frac{q+1}{2})} \sum_{j=0}^p \frac{c(q,j) a_{q+1}^{p-j}}{(p-j)!}$.

Remark 1: If $m > k$, then equation (4) becomes

$$f(x) = \frac{e^{-x/2} x^{m/2-1}}{2^{m/2} \sqrt{\pi} \prod_{i=1}^k \mu_i^{\frac{1}{2}}} \sum_{p=0}^{\infty} \frac{\Gamma(\frac{k}{2} + p)}{\Gamma(\frac{m}{2} + p)} \sum_{j=0}^p \frac{c(k,j) a_{k+1}^{p-j}}{(p-j)!} x^p \quad (9)$$

where $a_{k+1} = \frac{1}{2}(1 - \frac{1}{\mu_k})$.

Remark 2: Equation (4) is the pdf of $x=m \cdot \text{MSEF}/\sigma^2$. The pdf of $z=\text{MSEF}$ may be obtained by transforming $z = (\frac{\sigma^2}{m})x$, and this becomes

$$f(z|m, \sigma^2) = c_1 m^{\frac{1}{2}m} z^{\frac{1}{2}m-1} \sigma^{-m} \exp(-\frac{m}{2\mu_m} z \sigma^{-2}) \sum_{p=0}^{\infty} c(m,p) m^p z^p \sigma^{-2p} \quad (10)$$

where c_1 is the constant given by $c_1 = 1/(2^{\frac{1}{2}m} \sqrt{\pi} \prod_{i=1}^m \mu_i^{\frac{1}{2}})$.

Equations (4) and (10) have upper and lower bounds that are chi-square distributions, and this is stated in the following lemma.

Lemma 1: Let the pdf of z be denoted by $f(z|m, \sigma^2)$ as in equation (10).

Then

$$\begin{aligned} \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{m}{2})} c_1^m m^{\frac{1}{2}m} z^{\frac{1}{2}m-1} \sigma^{-m} \exp(-\frac{mz\sigma^{-2}}{2\mu_m}) &\leq f(z|m, \sigma^2) & (11) \\ &\leq \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{m}{2})} c_1^m m^{\frac{1}{2}m} z^{\frac{1}{2}m-1} \sigma^{-m} \exp(-\frac{mz\sigma^{-2}}{2\mu_1}). \end{aligned}$$

Proof: The first inequality is obvious. To prove the second inequality, we first use the following inequality which can be easily proven by induction and by the properties of the binomial distribution:

$$c(m, p) \leq \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{m}{2})} \frac{b_m^p}{p!} \quad (12)$$

where $b_1^0 = 1$, $b_1^i = 0$ for $i \geq 1$, and $b_i = \frac{1}{2}(\frac{1}{\mu_i} - \frac{1}{\mu_1})$, for $i=2, \dots, m$.

Substituting (12) in (10) we have

$$\begin{aligned} f(z|m, \sigma^2) &\leq c_1 \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{m}{2})} m^{\frac{1}{2}m} z^{\frac{1}{2}m-1} \sigma^{-m} \exp(-\frac{mz\sigma^{-2}}{2\mu_m}) \sum_{p=0}^{\infty} \frac{(mz b_m \sigma^{-2})^p}{p!} \\ &= c_1 \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{m}{2})} c_1^m m^{\frac{1}{2}m} z^{\frac{1}{2}m-1} \sigma^{-m} \exp(-\frac{mz\sigma^{-2}}{2\mu_m} + mz b_m \sigma^{-2}) \end{aligned}$$

$$= c_1 \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{m}{2})} m^{\frac{1}{2}m} z^{\frac{1}{2}m-1} \sigma^{-m} \exp\left(-\frac{mz\sigma^{-2}}{2\mu_1}\right).$$

In proving the lemma above, we have shown that $\sum_{p=0}^{\infty} c(m,p) m^p z^p \sigma^{-2p}$ converges. Equation (11) shows that the pdf of the MSEF is bounded by two chi-square distributions that are adjusted by maximum and minimum characteristic roots, μ_1 and μ_m , respectively, and equalities hold if $\mu_m = \mu_1$.

A predictive density of the MSEF will be given by

$$p(z|\text{data}) = \int_0^{\infty} f(z|\sigma^2, m) p(\sigma^2|\text{data}) d\sigma^2 \quad (13)$$

where $p(\sigma^2|\text{data})$ is the posterior pdf of σ^2 , which may be given by

$$p(\sigma^2|\text{data}) = \sigma^{-(v+1)} \exp\left(-\frac{vs^2}{2\sigma^2}\right) \quad (14)$$

where $v=n-k$, and $vs^2 = y'(I-X(X'X)^{-1}X')y$. Carrying out the integration in (13) we obtain

$$p(z|s^2, v, m) = \frac{z^{\frac{1}{2}m-1}}{(vs^2 + mz/\mu_m)^{(m+v)/2}} \sum_{p=0}^{\infty} \Gamma\left(\frac{m+v}{2} + p\right) 2^p c(m,p) \left[\frac{mz}{vs^2 + mz/\mu_m} \right]^p \quad (15)$$

Using equation (11) we can show that the predictive pdf of z is bounded by two F distributions:

$$c_1 \frac{\Gamma(\frac{1}{2}) \Gamma(\frac{m+v}{2})}{\Gamma(\frac{m}{2})} 2^{(m+v)/2-1} \frac{m^{\frac{1}{2}m} z^{\frac{1}{2}m-1}}{(vs^2 + mz/\mu_m)^{(m+v)/2}}$$

$$\leq P(z|s^2, v, m) \leq c_1 \frac{\Gamma(\frac{1}{2}) \Gamma(\frac{m+v}{2})}{\Gamma(\frac{m}{2})} 2^{(m+v)/2-1} \frac{m^{\frac{1}{2}m} z^{\frac{1}{2}m-1}}{(vs^2 + mz/\mu_1)^{(m+v)/2}$$

(16)

The equalities hold if $\mu_m = \mu_1$. When $m=1$, the predictive density of the squared root of the MSEF is identical to the predictive density for one period ahead forecast (Zellner (1971, pp.72-73)), which in turn is equal to the predictive density in the sampling theory framework.

In Theorem 1 we used the degenerate hyperbolic function. The distribution of quadratic forms is often given by the Laguerre polynomials. If we rearrange the Laguerre expansions given in Johnson and Kotz (1970, pp.159-160) to fit more conveniently in our case, the distribution of $x=m \cdot \text{MSEF}/\sigma^2$ is given by

$$f(x) = \frac{1}{\beta} P_m(x/\beta, 1) \Gamma(\frac{m}{2}) \sum_{p=0}^{\infty} \frac{(-2)^{-p}}{p! \Gamma(p+\frac{1}{2}m)}$$

$$\cdot \left(\sum_{j=p}^{\infty} \frac{j!}{(j-p)!} c_j \beta^{-j} \right) (x/\beta)^p \quad (17)$$

where $p_m(x/\beta, 1)$ is the chi-square distribution with m degrees of freedom; $c_0=1$, $c_r = (2r)^{-1} \sum_{j=0}^{r-1} s_{r-j} c_j$, ($r \geq 1$), $S_r = \sum_{j=1}^m (\beta - \mu_j)^r$, and β is a number such that $\beta > \frac{1}{2} \max_i \mu_i$.

If we use equation (17) the predictive density becomes

$$f(z|s^2, \nu, m, \beta) = \frac{z^{m-1}}{(\nu s^2 + mz/\beta)^{(m+\nu)/2}} \sum_{p=0}^{\infty} \frac{(-1)^p \Gamma(\frac{m+\nu}{2} + p)}{p! \Gamma(p + \frac{m}{2})} \cdot \left(\sum_{j=p}^{\infty} \frac{j!}{(j-p)!} c_j^{-j} \right) m^p \beta^{-p} \left(\frac{z}{\nu s^2 + mz/\beta} \right)^p \quad (18)$$

Equation (18) is conditioned on one additional parameter, β , whereas equation (15) is free of such an additional parameter.

Remark 3: The Bayesian predictive density of the MSEF in equation (15) may be derived in another way. We may start with the joint density for y_* , β , and σ^2 , and integrate β first and transform y_* into the MSEF given σ^2 , and finally we integrate σ^2 out:

$$p(y_*|y, X, X_*) = \iint p(y_*|\beta, \sigma^2, X_*) p(\beta, \sigma^2|y, X) d\sigma^2 d\beta$$

$$\propto \int \sigma^{-(\nu+m+1)} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) \cdot \exp\left[-\frac{1}{2\sigma^2}(y_* - X_*\hat{\beta})'H(y_* - X_*\hat{\beta})\right] d\sigma, \quad (19)$$

where $H = [I + X_*(X'X)^{-1}X_*']^{-1}$. The right-hand side of equation (19) shows that given σ , $y_* - X_*\hat{\beta}$ is distributed as $N(0, \sigma^2 H^{-1})$. Let $w = y_* - X_*\hat{\beta}$ and $H = R'R$, where R is a nonsingular matrix. Then $m \cdot z = w'w - \eta'(RR')^{-1}\eta$, with $\eta = Rw \sim N(0, \sigma^2 I_m)$. Since the nonzero characteristic roots of $H^{-1} = (R'R)^{-1}$ are the same as those of $(RR')^{-1}$, we see that $m \cdot z = \sum_{i=1}^m \mu_i \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2)$, and μ_i is the i th characteristic root of H^{-1} . Hence, given σ , $m \cdot z$ has the distribution of a quadratic form in normal variables that is given in equation (4). Thus

$$p(z | s^2, m) \propto \int \sigma^{-(\nu+m+1)} p(z | \sigma^2, m) \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) d\sigma, \quad (20)$$

and integrating σ out, we obtain equation (15).

Equation (15) is a Bayesian predictive density of the MSEF given s^2 . A sampling distribution of the MSEF is given in Theorem 2 below:

Theorem 2: Let z be the MSEF. Then $u = z / (\mu_m s^2)$ has the probability density function given by

$$f(u | \mu_1, \dots, \mu_m, \nu) = \text{const.} \cdot \frac{u^{m/2-1}}{\left(1 + \frac{m}{\nu} u\right)^{(m+\nu)/2}} \cdot \sum_{p=0}^{\infty} \Gamma\left(\frac{m+\nu}{2}\right) 2^p c(m,p) \mu_m^p \left(\frac{m}{\nu}\right)^p \left(\frac{u}{1 + mu/\nu}\right)^p, \quad (21)$$

where

$$\text{const} = \frac{(m/\nu)^{m/2} \mu_m^{m/2}}{\pi^{1/2} \prod_{i=1}^k \mu_i^{1/2} \Gamma\left(\frac{\nu}{2}\right)}$$

and $\nu s^2 = y'[I - X(X'X)^{-1}X']y$, $\nu = n - k$.

Proof: Equation (4) in Theorem 1 is the probability density function of $x=m \cdot \text{MSEF}/\sigma^2$. The probability density function of $z=\text{MSEF}$ may be obtained by transforming $z=(\sigma^2/m)x$, and this is given by equation (10). On the other hand, the pdf of s^2 is given by

$$f(s^2 | \sigma^2, \nu) = \frac{\nu \sigma^{-2}}{2^{1/2} \Gamma(\frac{\nu}{2})} \left(\frac{\nu s^2}{\sigma^2} \right)^{\nu/2-1} \exp\{-\nu s^2/(2\sigma^2)\}$$

and it is easy to show that z and s^2 are independent. Hence, the joint pdf of z and s^2 is

$$f(z, s^2 | m, \sigma^2) = c_3 \nu^{\nu/2} z^{m/2-1} (s^2)^{\nu/2} \sigma^{-(m+\nu)} \cdot \exp\left\{-\frac{\nu s^2}{2\sigma^2} [1 + mz/(\nu \mu_m s^2)]\right\} \sum_{p=0}^{\infty} c(m, p) m^p z^p \sigma^{-2p},$$

where $c_3 = c_1 [2^{\nu/2} \Gamma(\nu/2)]^{-1}$. Changing the variables z and s^2 to $u = z/(\mu_m s^2)$ and $y = s^2$, and integrating out y , we obtain equation (21).

3. Comparison of Certain Model Selection Criteria

The Bayesian predictive density of the MSEF that is given in (15) may be used as a criterion for selecting linear models. For each model we may draw the predictive density, and choose the model that has the mass of its density closest to zero. Or, we may choose the model that minimizes an expected loss. The choice of a quadratic loss function leads to the mean of the MSEF as the selection criterion.

The Bayesian criterion above belongs to the class of model selection criteria that are based on measures of how well each model

explains data. Akaike's (1974) information criterion (AIC) and Efron's (1984) confidence interval for the mean squared errors also belong to this class. The AIC may be given by

$$\text{AIC} = n \log \hat{\sigma}^2 + 2k \quad (22)$$

where $\hat{\sigma}^2$ is the sum of squared residuals divided by the sample size n , and k is the number of regression coefficients.

Efron's confidence interval for the mean squared errors may be interpreted as an inferential procedure for the C_p that is suggested by Mallows (1973). Let two linear regression models be given by

$$\begin{aligned} \text{Model A: } y &= X_A \beta_A + \epsilon \\ \text{Model B: } y &= X_B \beta_B + \epsilon \end{aligned} \quad (23)$$

where y is an $(n \times 1)$ vector of observations on the dependent variable; X_i is an $(n \times k_i)$ matrix of observations on the k_i explanatory variables of model i ($i=A,B$) and β_i is a $(k_i \times 1)$ vector of regression coefficients of model i ($i=A,B$), and ϵ is an $(n \times 1)$ vector of error terms. The unbiased estimator of the difference of the mean squared errors (MSE) of models A and B, $\Delta = \text{MSE}_B - \text{MSE}_A$, is given by

$$\hat{\Delta} = (|y_{B0}|^2 - |y_{A0}|^2) + 2(d_B - d_A) \bar{\sigma}^2 \quad (24)$$

where $|y_{B0}|^2 = \hat{\beta}_A' X_A' M_B X_A \hat{\beta}_A$, $|y_{A0}|^2 = \hat{\beta}_B' X_B' M_A X_B \hat{\beta}_B$, $M_i = I - X_i (X_i' X_i)^{-1} X_i'$, $i=A,B$, $\bar{\sigma}^2 = y' [I - XX'] y$, $X = [X_A, X_B]$, and d_i is the dimension of model i .

Efron decomposes the MSE (Δ) and its estimate ($\hat{\Delta}$) in a symmetric coordinate system and proposes to compute confidence intervals in the symmetric coordinate system. The computation of confidence intervals is suggested either by parametric bootstrapping or by non-parametric bootstrapping.

In contrast to the class of model selection criteria based on measures of 'goodness of fit', Cox's tests of separate families [Cox (1962)] are based on the translation of non-nested models into hypothesis testing on parameters. Pesaran (1974, 1982) proposes the N-tests. Davidson and MacKinnon (1981) suggest the J-tests. The N-tests are given by

$$N_0 = \frac{n}{2} \log \left(\frac{\hat{\sigma}_B^2 / \hat{\sigma}_{BA}^2}{\hat{\sigma}_A^2} \right) / \left\{ \frac{\hat{\sigma}_A^2}{\hat{\sigma}_{BA}^2} \hat{\beta}'_A X'_A M_{A B} M_{A B} X_A \hat{\beta}_A \right\}^{1/2} \quad (25)$$

where $\hat{\sigma}_i^2 = y' M_i y / n$ ($i=A, B$), and $\hat{\sigma}_{BA}^2 = \hat{\sigma}_A^2 + (\hat{\beta}'_A X'_A M_{A B} X_A \hat{\beta}_A) / n$. The N_0 test is computed using model A in (23) as the null hypothesis. By using model B as the null hypothesis, one obtains the N_1 test the formula of which is given by interchanging subscripts A and B in (25). The J test by Davidson and MacKinnon is the t-test on parameter λ in

$$y = X_A b_A + \lambda (X_B \hat{\beta}_B) + u \quad (26)$$

where $b_A = (1-\lambda)\beta_A$. Again, a symmetric test can arise by testing λ in

$$y = \lambda(X_A \hat{\beta}_A) + X_B b_B + u \quad (27)$$

where $b_B = (1-\lambda)\beta_B$.

The N- and J- tests give rise to cases where one either rejects or accepts both models. Table 1 gives four possible cases.

Table 1: Four Cases of N- and J- Tests

| | $N_0 (J_0) - \text{Test}$ | |
|---------------------------|--|--|
| $N_1 (J_1) - \text{Test}$ | <u>Case 1</u> Accept Model A Reject Model B (p_1) | <u>Case 2</u> Reject Model A Accept Model B (p_2) |
| | <u>Case 3</u> Reject Model A Reject Model B (p_3) | <u>Case 4</u> Accept Model A Accept Model B (p_4) |

Note: p_i is the probability of case i , $(i=1, \dots, 4)$.

As is obvious from equation (25), the N-tests can only be defined if $X_A' M_B M_A M_B X_A \neq 0$. Sufficient conditions for making this quantity zero are $M_B X_A = 0$ or $M_B X_A = X_A$. $M_B X_A = 0$ occurs if the columns of X_A are linear combination of the columns of X_B , and $M_B X_A = X_A$ occurs if $X_B' X_A = 0$ (i.e. when the explanatory variables of the two models

are orthogonal.) As for the J-tests, they cannot be defined if linear dependence exists between X_A and X_B .

Let us make sampling experiments to compare the powers of the Bayesian MSEF criterion, the AIC, Efron's confidence interval, N-tests, and J-tests. In evaluating these tests we need to develop a measure of nearness of competing two models. Pesaran (1982) introduces a sequence of local alternatives

$$X_B = X_A C + n^{-\frac{1}{2}} D + o(n^{-\frac{1}{2}}) \quad (28)$$

where C and D are $k_A \times k_B$ and $n \times k_B$ nonzero matrices of constants, and $D'M_A D/n$ exists. Pesaran uses the local alternatives (28) so that he can derive asymptotic non-null distributions of the test statistics.

Instead of (28), the measure of nearness of two models may be given by the measure of correlation among non-overlapping explanatory variables of the two models. Let models A and B be written as

$$y = X_1 \beta_{A1} + X_2 \beta_{A2} + \epsilon$$

$$y = X_1 \beta_{B1} + Z \beta_{B2} + \epsilon$$

so that $X_A = [X_1, X_2]$ and $X_B = [X_1, Z]$. The non-overlapping explanatory variables of the two models are X_2 and Z , and the measure of nearness of the two models may be given by

$$\rho_{AB}^2 = \text{Min}(\lambda_i^2)$$

where λ_i^2 is the square of the i -th nonzero canonical correlation coefficient between X_2 and Z . ρ_{AB}^2 is bounded between 0 and 1, and if $\rho_{AB}^2 = 1$, the models A and B can be thought to be identical, whereas $\rho_{AB}^2 = 0$ indicates that the two models are farthest apart.

Sampling experiments are made by specifying the two models as

$$\text{Model A: } y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t$$

$$\text{Model B: } y_t = \gamma_0 + \gamma_1 x_{t1} + \gamma_2 z_{t2} + \varepsilon_t$$

Hence, the models A and B have $(1, x_{t1})$ as the common variables, whereas x_{t2} and z_{t2} are uncommon variables. As in Pesaran's (1982) experiments, x_{ti} 's are drawn from $N(0,1)$, and z_{t2} is generated by

$$z_{t2} = \lambda_2 x_{t2} + v_{t2}, \quad v_{t2} \sim N(0,1).$$

λ_2 is controlled by the correlation between z_{t2} and x_{t2} :

$$\lambda_2 = \rho_2 / (1 - \rho_2^2)^{1/2}$$

where $\rho_2 = \text{Corr}(x_{t2}, z_{t2})$.

The model selection criteria are also influenced by the 'fit' of the true model as measured by the coefficient of determination of the

true model (model A in our experiments), R^2 , and by the relative sizes of β_1 and β_2 . In our experiments, we set R^2 at .5 ($R^2=.5$), and in Table 2 we set β_1 and β_2 to be (1.0, .5), respectively, whereas in Table 3 the values of β_1 and β_2 are switched: (.5, 1.0). The constant term ρ_0 is set at 1.0 in both tables. The number of replications for each value of ρ_2^2 is 500.

The following observations can be made from Tables 2 and 3:

- (1) As the sample sizes increase the powers of all the criteria tend to increase for given values of ρ_2^2 .
- (2) Comparing Table 3 with Table 2, we see that the powers in Table 3 are larger than those in Table 2.
- (3) The N-test tends to perform better than the J-test. For $\rho_2^2=0.1$ or 0, the powers of the N-test decline. This is due to the fact that for low values of ρ_2^2 , the nonoverlapping variables x_{t2} and z_{t2} tend to be orthogonal, and this brings the N-test closer to the case in which it is not defined (i.e. $X_A'X_B=0$).
- (4) The powers of the Bayesian MSEF criterion tend to dominate those of the other criteria, especially for the cases of sample size 20. For larger sample sizes, the AIC performs as good as the Bayesian MSEF criterion.

Table 2: Empirical Powers of Model Selection
 Criteria [$R^2 = .5$, $(\beta_1, \beta_2) = (1.0, .5)$]

| | δ | N-Test | | J-Test | | Difference of Predictive Means (4) | Difference of AIC's (5) | Efron's 90% CI (6) |
|----------------|----------|-------------|-----------------|-------------|-----------------|---|----------------------------|-----------------------|
| | | $p_1^{(2)}$ | $1-\beta^{(3)}$ | $p_1^{(2)}$ | $1-\beta^{(3)}$ | | | |
| n=20 | | | | | | | | |
| $\rho_2^2=1.0$ | 0 | ND (7) | ND (7) | ND (7) | ND (7) | 1 (8) | 1 (8) | 0 (8) |
| .9 | 2 | .216 | .218 | .054 | .084 | .84 | .662 | 0 |
| .7 | 6 | .402 | .402 | .15 | .17 | .932 | .782 | .002 |
| .5 | 10 | .466 | .470 | .138 | .156 | .964 | .828 | .024 |
| .3 | 14 | .668 | .702 | .318 | .336 | .978 | .859 | .028 |
| .1 | 18 | .680 | .724 | .308 | .334 | .988 | .872 | .078 |
| 0 | 20 | .722 | .898 | .418 | .452 | .982 | .848 | .154 |
| n=60 | | | | | | | | |
| $\rho_2^2=1.0$ | 0 | ND | ND | ND | ND | 1 | 1 | 0 |
| .9 | 6 | .262 | .272 | .182 | .200 | .710 | .752 | 0 |
| .7 | 18 | .600 | .604 | .392 | .414 | .876 | .884 | .004 |
| .5 | 30 | .770 | .778 | .638 | .652 | .936 | .936 | .114 |
| .3 | 42 | .880 | .946 | .772 | .824 | .972 | .982 | .182 |
| .1 | 54 | .548 | 1 | .902 | .928 | .990 | .990 | .574 |
| 0 | 60 | .812 | 1 | .888 | .936 | .994 | .990 | .642 |
| n=100 | | | | | | | | |
| $\rho_2^2=1.0$ | 0 | ND | ND | ND | ND | 1.0 | 1.0 | 0 |
| .9 | 10 | .406 | .418 | .288 | .310 | .790 | .800 | 0 |
| .7 | 30 | .700 | .712 | .570 | .592 | .920 | .916 | .040 |
| .5 | 50 | .906 | .946 | .834 | .866 | .972 | .962 | .450 |
| .3 | 70 | .940 | 1.0 | .928 | .974 | .992 | .994 | .702 |
| .1 | 90 | .906 | 1.0 | .946 | .996 | 1.0 | 1.0 | .812 |
| 0 | 100 | .566 | 1.0 | .952 | .998 | 1.0 | 1.0 | .962 |

Notes: For each value of ρ_2 , the number of replications is 500.

(1) $\rho_2 = \text{Corr}(x_{t2}, z_{t2})$, and δ is the measure of the distance of two models, D, in equation (28), and it is given by $\delta = \lim_{n \rightarrow \infty} D'M_A D / \lim_{n \rightarrow \infty} (X_B' X_B / n)$. In our experimental design becomes $\delta = n(1-\rho_2^2)$.

(2) p_1 is the probability of accepting model A and rejecting model B.

(3) β is the probability of Type II errors, and it is given by $\beta = p_2 + p_3$ in Table 1.

(Notes on Table 2 continued)

- (4) The predictive mean is computed by $E(\text{MSEF}|\cdot) = \int zp(z|\text{data})dz$ for each model, and the difference is $E(\text{MSEF}_A|\cdot) - E(\text{MSEF}_B|\cdot)$.
- (5) The difference of the AIC's is $AIC_A - AIC_B$.
- (6) Efron's 90% confidence interval (CI) is computed by assuming that the sample estimate, $\bar{\sigma}^2$ is true (hence $d_E = \infty$ in Efron's notation), and by the Edgeworth expansions for the parametric bootstrap distribution without resorting to Monte Carlo.
- (7) For $\rho_2^2 = 1.0$ the N- and J- tests are not defined.
- (8) For $\rho_2^2 = 1.0$, the difference of the predictive means and the difference of the AIC's both become 1 by construction. Efron's CI becomes zero by construction.
- (9) For all the sample sizes, the period of prediction, m , is set at 10.

Table 3: Empirical Powers of Model Selection
Criteria [$R^2=.5$, $(\beta_1, \beta_2)=(.5, 1.0)$]

| | δ | N-Test | | J-Test | | Difference of Predic- tive Means | Difference of AIC's | Efron's 90% CI |
|----------------|----------|--------|-----------|--------|-----------|--|------------------------|-------------------|
| | | p_1 | $1-\beta$ | p_1 | $1-\beta$ | | | |
| n=20 | | | | | | | | |
| $\rho_2^2=1.0$ | 0 | ND | ND | ND | ND | 1 | 1 | 0 |
| .9 | 2 | .308 | .412 | .110 | .128 | .888 | .822 | 0 |
| .7 | 6 | .784 | .808 | .474 | .504 | .972 | .930 | .076 |
| .5 | 10 | .862 | .914 | .694 | .718 | .996 | .976 | .248 |
| .3 | 14 | .884 | .962 | .746 | .876 | 1.0 | .986 | .316 |
| .1 | 18 | .872 | .966 | .812 | .852 | 1.0 | .996 | .456 |
| 0 | 20 | .846 | .998 | .906 | .962 | 1.0 | 1.0 | .696 |
| n=60 | | | | | | | | |
| $\rho_2^2=1.0$ | 0 | ND | ND | ND | ND | 1.0 | 1.0 | 0 |
| .9 | 6 | .600 | .608 | .502 | .516 | .920 | .896 | .002 |
| .7 | 18 | .946 | .978 | .926 | .958 | .992 | .992 | .428 |
| .5 | 30 | .906 | .998 | .914 | .996 | .998 | .998 | .808 |
| .3 | 42 | .964 | 1.0 | .960 | .998 | 1.0 | 1.0 | .900 |
| .1 | 54 | .698 | 1.0 | .970 | 1.0 | 1.0 | 1.0 | .998 |
| 0 | 60 | .876 | 1.0 | .948 | 1.0 | 1.0 | 1.0 | 1.0 |
| n=100 | | | | | | | | |
| $\rho_2^2=1.0$ | 0 | ND | ND | ND | ND | 1 | 1 | 0 |
| .9 | 10 | .858 | .874 | .814 | .836 | .922 | .942 | .174 |
| .7 | 30 | .954 | .998 | .952 | .998 | .998 | .998 | .774 |
| .5 | 50 | .956 | 1.0 | .948 | 1.0 | 1.0 | 1.0 | .978 |
| .3 | 70 | .952 | 1.0 | .966 | 1.0 | 1.0 | 1.0 | .998 |
| .1 | 90 | .936 | 1.0 | .950 | 1.0 | 1.0 | 1.0 | 1.0 |
| 0 | 100 | .714 | 1.0 | .954 | 1.0 | 1.0 | 1.0 | 1.0 |

See footnotes below Table 2.

- (5) Efron's 90% confidence interval (CI) appears to be too conservative, and for small sample sizes, the powers are substantially lower than the other criteria.

For the N- and J- tests we presented two measures of power, p_1 and $1-\beta$, respectively. The probability of Type II error is β , and $1-\beta$ is the conventional concept of power in a nested hypothesis. As Pesaran (1974) states, however, for a non-nested hypothesis, a suitable concept of power is the probability of making correct decision, which is p_1 . Pesaran (1982) uses $1-\beta$ as the measure of power in his experiments.

Efron (1984) suggests a non-parametric bootstrapping procedure in addition to a parametric bootstrapping procedure. Since the non-parametric bootstrapping procedure requires considerable computational time in generating empirical powers, we did not carry it out in our experiments. Confidence intervals that are generated by non-parametric bootstrapping tend to be larger than those by parametric bootstrapping, and their powers are in general lower than those by parametric bootstrapping.

In our sampling experiments, we set the prediction period, m , at 10. We varied m at different values, and the results are comparable to those of $m=10$.

References

- Akaike, H. (1974) "A New look at the statistical model identification," IEEE Transactions in Automatic Control, AC-19, 716-723
- Bhattacharyya, A. (1945) "A Note on the distribution of the sum of chi-squares," Sankhya, 7, 27-28
- Bhattacharyya, B.C. (1943) "On an aspect of the Pearson system of curves, and a few analogies," Sankhya, 6, 415-448
- Davidson, R. and J.G. MacKinnon (1981) "Several tests for model specification in the presence of alternative hypotheses," Econometrica, 49, 781-793
- Efron, B. (1984) "Comparing non-nested linear models," Journal of the American Statistical Association, 79, 791-803
- Gradshteyn, I.S. and I.M. Ryzhik (1980) Tables of integrals, series, and products (Academic Press, New York)
- Gurland, J. (1953) "Distribution of quadratic forms and ratios of quadratic forms," Annals of Mathematical Statistics, 24, 416-427
- Hotelling, H. (1940) "The selection of variates for use in prediction with some comments on the general problem of nuisance parameters," Annals of Mathematical Statistics, 11, 271-283
- Hotelling, H. (1948) "Some new methods for distributions of quadratic forms," (Abstract), Annals of Mathematical Statistics, 19, 119
- Jhonson, N.L. and S. Kotz (1970) Continuous univariate distributions-2 (John Wiley and Sons, New York)
- Mallows, C. (1973) "Some comments on C_p " Technometrics, 15, 661-675
- McCarthy, M.D. (1939) "On the application of the z-test to randomized blocks," Annals of Mathematical Statistics, 10, 337-359
- Neumann, J. von. (1941) "Distribution of the ratio of the mean square successive difference to variance," Annals of Mathematical Statistics, 12, 367-395
- Pesaran, M.H. (1974) "On the general problem of model selection," the Review of Economic Studies, 41, 153-171
- Pesaran, M.H. (1982) "Comparison of local power of alternative tests of non-nested regression models," Econometrica, 50, 1287-1304
- Zellner, A. (1971) An introduction to Bayesian inference in econometrics, (John Wiley and Sons, New York)