The Annals of Statistics
1981, Vol. 9, No. 3, 475–485

ADMISSIBLE SELECTION OF AN ACCURATE AND PARSIMONIOUS NORMAL LINEAR REGRESSION MODEL[1]

BY CHARLES J. STONE

University of California, Los Angeles

Let M_0 be a normal linear regression model and let $M_1, \cdots, M_K$ be distinct proper linear submodels of M_0. Let $\hat{k} \in \{0, \cdots, K\}$ be a model selection rule based on observed data from the true model. Given $\hat{k}$, let the unknown parameters of the selected model $M_{\hat{k}}$ be fitted by the maximum likelihood method. A loss function is introduced which depends additively on two parts: (i) a measure of the difference between the fitted model $M_{\hat{k}}$ and the true model; and (ii) a measure $C_{\hat{k}}$ of the "complexity" of the selected model. A natural model selection rule $\bar{k}$, which minimizes an empirical version of this loss, is shown to be admissible and very nearly Bayes.

1. General discussion and statement of results. Let $\mathcal{X}$ denote an arbitrary set (e.g., $\mathcal{X} = \mathbb{R}^d$ for some $d \geq 0$). Given $x \in \mathcal{X}$, let Y be normally distributed with mean $m(x)$ and standard deviation $\sigma > 0$. The unknown regression function is assumed to belong to a collection M of functions on $\mathcal{X}$. Let $M_0 \subset M$ be a vector space of finite dimension $p \geq 1$, which can be thought of as a normal linear regression model. A subspace of M_0 will be called a submodel. Let K denote a positive integer and let $M_1, \cdots, M_K$ be distinct proper submodels of M_0.

EXAMPLE 1. (Polynomial regression). $\mathcal{X} = \mathbb{R}$; $K \geq 1$; $M_0 = \{b_0 + b_1 x + \cdots + b_K x^K\}$ (the collection of polynomials of degree K or less); $M_k = \{b_0 + b_1 x + \cdots + b_{K-k} x^{K-k}\}$ for $1 \leq k \leq K$.

EXAMPLE 2. (All subsets regression). $\mathcal{X} = \mathbb{R}^d$; $M_0 = \{b_0 + b_1 x_1 + \cdots + b_d x_d\}$, where $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$; $M_1, \cdots, M_{2^d-1}$ are the submodels of M_0 obtained by requiring that $b_j = 0$ for $j \in S$, where S ranges over the nonempty subsets of $\{1, \cdots, d\}$.

Let n be a positive integer, let $x_1, \cdots, x_n \in \mathcal{X}$ be given and let $Y_1, \cdots, Y_n$ be independent normally distributed random variables each having standard deviation σ and such that Y_i has mean $m(x_i)$, $m \in M$ being the true regression function. For simplicity the identifiability assumption (relative to $\{x_1, \cdots, x_n\}$) is made that if $m \in M_0$ and $m(x_1) = \cdots = m(x_n) = 0$, then $m = 0$.

A *model selection rule* $\hat{k}$ is a (possibly randomized) function of $Y_1, \cdots, Y_n$ taking on values in $\{0, \cdots, K\}$. (Note that $x_1, \cdots, x_n$ are regarded as fixed in the theoretical results of this paper; otherwise $\hat{k}$ should be allowed also to depend on these quantities.) There is a large literature on model selection. See Leamer (1978) and review articles and bibliographies by Gaver and Geisel (1974), Ramsey (1974), Hocking (1976), Bancroft and Han (1977), Pereira (1977) and Thompson (1978).

Let $\bar{m}_k$ be the maximum likelihood estimator of the true regression function subject to the constraint of being in M_k. Then $\bar{m}_k$ is also the least squares estimator of m subject to this constraint; that is, it uniquely minimizes $n^{-1} \sum_1^n \{Y_i - \hat{m}(x_i)\}^2$ as $\hat{m}$ ranges over M_k.

Received October, 1979; revised July, 1980.

[1] This research was supported by NSF Grant No. MCS 77-02121.

AMS 1980 *subject classifications.* Primary 62J05; secondary 62C15.

Key words and phrases. Admissibility, normal linear regression model, generalized Bayes, parsimony, complexity.

If $\hat{k}$ is chosen as that value of k which minimizes $n^{-1}\sum_1^n \{Y_i - \bar{m}_k(x_i)\}^2$, then $\hat{k} = 0$ with probability one. To obtain nontrivial model selection rules, choose real-valued "penalties" C_k, $0 < k \leq K$. Let the model selection rule $\bar{k}$ be chosen to minimize $n^{-1}\sum_1^n \{Y_i - \bar{m}_k(x_i)\}^2 + C_k$. (Ties, which occur with probability zero, can be broken, say, by minimizing k.) In the theoretical development below the C_k's are nonrandom, but in applications they as well as the x_i's can be random.

The criterion of minimizing Mallows' C_p statistic yields the rule $\bar{k}$ with $C_k = 2n^{-1}\sigma^2 \dim M_k$ if σ is known and $C_k = 2n^{-1}\hat{\sigma}^2 \dim M_k$ if σ is unknown and $\hat{\sigma}$ is an estimate of it. Mallows (1973) discusses but does not recommend model selection rules based on minimizing C_p. Akaike's Information Criterion (AIC) (see Akaike (1973, 1974) and the large more recent related literature by Akaike and others) leads to the same criterion as minimizing C_p if σ is known. If σ is unknown, however, AIC leads to minimizing

$$\log\left[\frac{1}{n}\sum_1^n \{Y_i - \bar{m}_k(x_i)\}^2\right] + \frac{2}{n}\dim M_k.$$

Selection rules based on C_p and AIC appear to be reasonable in problems such as polynomial regression in which there is at most one model of any given dimension. But in problems such as all subsets regression, especially when there is a large number n of cases and many independent variables, it is desirable to consider penalties of the form $C_k = c \dim M_k$, where c is significantly larger than the value $2n^{-1}\hat{\sigma}^2$ suggested by C_p.

One promising way of choosing c when n is large is to use cross-validation (see M. Stone (1974) and Geisser (1975)) to get a reliable estimator $\hat{R}(c)$ of

$$\sigma^2 + \frac{1}{n}\sum_1^n \{\bar{m}_{\bar{k}(c)}(x_i) - m(x_i)\}^2,$$

where $\bar{k}(c)$ is the rule $\bar{k}$ when $C_k = c \dim M_k$, and then to choose $\hat{c}$ to minimize $\hat{R}(c)$. The model selection rule $\bar{k}(\hat{c})$ and an analogous rule for classification problems have been made computationally feasible and used successfully to prune regression and classification trees grown by AID and similar algorithms (see Breiman and Stone (1978)).

The purpose of this paper is to formulate and prove some optimality properties for the model selection rule $\bar{k}$. In addition to being worthwhile in itself, the results give some theoretical justification to the many efforts to find model selection rules which work well in practice.

The results will be formulated in a slightly more general context. It is assumed that $\mathcal{X}$ is a measure space and that the functions in M are measurable. Let μ be a probability measure on $\mathcal{X}$, which is assumed to be regular in the sense that

(i) $\int m^2(x)\mu\,(dx) < \infty$ for $m \in M$

and

(ii) $m \in M_0$ and $\int m^2(x)\mu\,(dx) = 0$ together imply that $m = 0$.

Let an estimate $\hat{m}$ of the true regression function be used to predict Y for an x chosen at random according to μ by setting $\hat{Y} = \hat{m}(x)$. Then for given $\hat{m}$, the mean square error of prediction is

$$\sigma^2 + \int \{\hat{m}(x) - m(x)\}^2\mu(dx).$$

Consider the inner product $(\ ,\ )$ on M defined by

$$(m_1, m_2) = \int m_1(x)m_2(x)\mu(dx)$$

and the corresponding norm $\|\ \|$ defined by $\|m\| = \sqrt{(m, m)}$. Then

$$\|\hat{m} - m\|^2 = \int \{\hat{m}(x) - m(x)\}^2 \mu(dx)$$

is a measure of the inaccuracy of the estimate $\hat{m}$ of m. For $0 \leq k \leq K$ let P_k denote the orthogonal projection of M_0 onto M_k relative to the indicated inner product. (Note that P_0 is the identity transformation on M_0). Then $P_k m$ uniquely minimizes $\|\hat{m} - m\|$ *as* $\hat{m}$ ranges over M_k.

Let $\bar{m}_0$ be the maximum likelihood estimator of the true regression function based on $Y_1, \cdots, Y_n$ and subject to the constraint of being in M_0. Within M_k a natural estimator of the true regression function is given by $\bar{m}_k = P_k \bar{m}_0$. If, for example, μ is the empirical distribution μ_0 of $x_1, \cdots, x_n$ defined by $\mu_0(A) = n^{-1} \sum_1^n I_A(x_i)$, where I_A is the indicator function of A, then

$$\|\hat{m} - m\|^2 = \frac{1}{n} \sum_1^n \{\hat{m}(x_i) - m(x_i)\}^2;$$

also

$$\frac{1}{n} \sum_1^n \{Y_i - \hat{m}(x_i)\}^2 = \frac{1}{n} \sum_1^n \{Y_i - \bar{m}_0(x_i)\}^2 + \|\hat{m} - \bar{m}_0\|^2 \qquad \text{for} \quad \hat{m} \in M_k,$$

so $\bar{m}_k$ reduces to its previous definition in this special case.

Consider the loss function for $k \in \{0, \cdots, K\}$ defined by

$$L(m, k) = \|\bar{m}_k - m\|^2 + C_k. \tag{1.1}$$

As mentioned above, $\|\bar{m}_k - m\|^2$ is a measure of the inaccuracy of the estimate $\bar{m}_k$ of m. The penalty C_k can be interpreted as the "payment for using a complicated function" (see Kiefer (1968)), the cost of measuring those independent variables required to compute $\bar{m}_k(x)$ (see Lindley (1968)), or the complexity of the model M_k (see Demster (1971)). The interpretation of C_k as a measure of the complexity of the model M_k (e.g., proportional to dim M_k) corresponds to the principle known as Occam's Razor, which emphasizes the desirability of selecting accurate and parsimonious models of reality. (For various admonitions in the statistical literature to follow this principle see page 8 of Blalock (1961), Tukey (1961) and Box (1976).) A closely related principle in hypothesis testing emphasizes the desirability of considering "material" or "substantive" significance as opposed to mere statistical significance (see Hodges and Lehmann (1954)). Anderson (1962) formulated the problem of choosing the degree of polynomial regression along the lines of hypothesis testing. He motivated the problem, however, in terms of the principle of economy described above. No particular form or interpretation of the penalties C_k, $0 \leq k \leq K$, is required for the theoretical results below.

The risk function for a model selection rule $\hat{k}$ is given by

$$R(m, \hat{k}) = E_m L(m, \hat{k}), \qquad m \in M,$$

where $E_m(P_m)$ denotes expectation (probability) when m is the true regression function. Corresponding to $\hat{k}$ and a prior probability distribution ρ on M is the Bayes risk

$$r(\rho, \hat{k}) = \int R(m, \hat{k}) \rho(dm).$$

Let $r(\rho)$ denote the infimum of $r(\rho, \hat{k})$ as $\hat{k}$ ranges over all model selection rules.

Let $m_1, \cdots, m_p$ be an orthonormal basis of M_0 with respect to the inner product defined for $m_1, m_2 \in M_0$ as $\sum_1^n m_1(x_i) m_2(x_i)$. Let dm refer to the infinite measure on M_0 induced by Lebesgue measure on $\mathbb{R}^p$ and the map $(b_1, \cdots, b_p) \to b_1 m_1 + \cdots + b_p m_p$. (Observe that dm is independent of the particular choice of the orthogonal basis $m_1, \cdots, m_p$.) For $A \subset M_0$ set $|A| = \int_A dm$ and, for $c \in \mathbb{R}$, set $cA = \{cm : m \in A\}$. Given $A \in M_0$ with $0 < |A| < \infty$, let ρ_A denote the uniform probability distribution on A defined by $\rho_A(dm) =$

$|A|^{-1}I_A(m)dm$. The set A is said to be "a compact convex polyhedron in M_0 containing the origin as an interior point" if it is the image under the above map of such a set in $\mathbb{R}^p$, in which case $0 < |A| < \infty$.

Consider now the specific model selection rule $\bar{k}$ defined to be a value of $k \in \{0, \cdots, K\}$ which minimizes $L(\bar{m}_0, k) = \|\bar{m}_k - \bar{m}_0\|^2 + C_k$. (Note that $\bar{k}$ reduces to its previous definition in the special case $\mu = \mu_0$.) *The rule $\bar{k}$ is generalized Bayes.* To see this let ρ_0 be the improper prior distribution on M_0 defined by $\rho_0(dm) = dm$. The corresponding posterior distribution of m given $Y_1, \cdots, Y_n$ is the same as the distribution of $\bar{m}_0 + \sum_1^p \xi_j m_j$, where $\xi_1, \cdots, \xi_p$ are independent $N(0, \sigma^2)$ random variables. It follows easily from this representation that $\bar{k}$ is generalized Bayes with respect to ρ_0. For similar results see Lindley (1968), Brooks (1973), Halpern (1973), and Faden and Rausser (1976).

Two model selection rules $\hat{k}$ and k^* are said to be *equivalent* if $P_m(\hat{k} = k^*) = 1$ for some $m \in M$ or, equivalently, for all $m \in M$. Written as $\hat{k}(Y_1, \cdots, Y_n)$ and $k^*(Y_1, \cdots, Y_n)$, these rules are equivalent if and only if $\hat{k}(y_1, \cdots, y_n) = k^*(y_1, \cdots, y_n)$ almost everywhere on $\mathbb{R}^n$. If $\hat{k}$ is equivalent to k^*, then $R(m, \hat{k}) = R(m, k^*)$ for all $m \in M$.

A model selection rule k^* is said to be *admissible* if there is no model selection rule $\hat{k}$ such that $R(m, \hat{k}) \leq R(m, k^*)$ for all $m \in M$ with strict inequality holding for some $m \in M$. A sufficient condition for k^* to be admissible is that for every rule $\hat{k}$ which is not equivalent to k^* there is an $m \in M$ such that $R(m, \hat{k}) > R(m, k^*)$. A necessary and sufficient condition for this is that for every rule $\hat{k}$ which is not equivalent to k^* there is a prior probability distribution ρ on M such that $r(\rho, \hat{k}) > r(\rho, k^*)$. Thus (ii) of Theorem 1 below implies that *$\bar{k}$ is admissible* ((i) implies that $\bar{k}$ is "very nearly Bayes"). Since $\bar{k}$ does not depend on the choice of M or σ (except that $M \supset M_0$), these quantities can be regarded as either known or unknown. If $C_0 = \cdots = C_k$, then $\bar{k} = 0$ is admissible.

THEOREM 1. *There is a compact convex polyhedron A in M_0 containing the origin as an interior point such that* (i) *for some $\delta > 0$*

$$r(\rho_{cA}, \bar{k}) - r(\rho_{cA}) = o(e^{-\delta c^2}) \qquad as \quad c \to \infty;$$

and (ii) *if $\hat{k}$ is any model selection rule which is not equivalent to $\bar{k}$, then*

$$\liminf_{c\to\infty} c^p[r(\rho_{cA}, \hat{k}) - r(\rho_{cA}, \bar{k})] > 0.$$

Theorem 1 will be reduced to a more convenient form in Section 2, which will be proven in Section 3.

Instead of insisting that $\bar{m}_k$ be used once the model M_k is selected, one can consider more general procedures for choosing $\hat{m} \in M_k$ to estimate m. This leads to admissibility problems for procedures of the form $(\hat{m}, \hat{k})$, where $\hat{m} \in M_{\hat{k}}$. An argument suggested in part by the proof of Theorem 4.2 of James and Stein (1961) can be used to show that $(\bar{m}_{\bar{k}}, \bar{k})$ is admissible in this more general context if $\dim M_0 \leq 2$ (for the special case of this result when $\dim M_0 = 1$ see Meeden and Arnold (1980)). When $\dim M_0 \geq 3$, however, Stein type considerations presumably lead to the inadmissibility of $(\bar{m}_{\bar{k}}, \bar{k})$ unless the class of competing procedures is sufficiently reduced by invariance requirements to lead to an equivalent admissibility problem in a one- or two-dimensional setting.

I wish to thank Larry Brown and Arthur Cohen for a number of valuable comments on previous versions of this paper.

2. Canonical form. In this section Theorem 1 will be reduced to a more convenient form, which will first be described.

Let $1 \leq p \leq n$, let "$\cdot$" denote the usual inner product on $\mathbb{R}^p$, and let $|\ |$ be the corresponding norm given by $|v| = \sqrt{v \cdot v}$. Let H be a positive definite symmetric $p \times p$ matrix and define the inner product $(,)$ and norm $\|\ \|$ on $\mathbb{R}^p$ by $(v_1, v_2) = v_1 \cdot Hv_2$ and $\|v\| = \sqrt{(v, v)}$. If H is the $p \times p$ identity matrix, then $(\ ,\ )$ and $\|\ \|$ reduce to "$\cdot$" and $|\ |$.

Let K be a positive integer. Set $V_0 = \mathbb{R}^p$ and let V_k, $1 \leq k \leq K$, be distinct proper

subspaces of V_0. For $0 \le k \le K$ let P_k denote the orthogonal projection (relative to the inner product $(\ ,\ )$) of V_0 onto V_k and let C_k be a real-valued constant.

Let $Z_1, \cdots, Z_p, W_1, \cdots, W_{n-p}$ be independent normally distributed random variables each having standard deviation $\sigma > 0$ and such that $EW_1 = \cdots EW_{n-p} = 0$. Let Z be the random vector in $\mathbb{R}^p$ having coordinates $Z_1, \cdots, Z_p$ and let W be the random vector in $\mathbb{R}^{n-p}$ having coordinates $W_1, \cdots, W_{n-p}$. Set $\bar{v}_k = P_k Z$ for $0 \le k \le K$. Observe that $\bar{v}_0 = P_0 Z = Z$ is the maximum likelihood estimator of the true mean $v \in V_0$ of Z based on Z, W. (If H is the identity matrix, then for $1 \le k \le K$, $\bar{v}_k$ is the maximum likelihood estimator of v under the constraint of being in V_k.) The loss associated with the true mean v of Z and the value $k \in \{0, \cdots, K\}$ is defined to be

$$L(v, k) = \| v - \bar{v}_k \|^2 + C_k.$$

A model selection rule $\hat{k}$ is a randomized $\{0, \cdots, K\}$-valued function of Z, W. The risk function for such a rule is defined by

$$R(v, \hat{k}) = E_v L(v, \hat{k}), \qquad v \in V_0.$$

The quantities $r(\rho, \hat{k})$ and $r(\rho)$ are also defined as in Section 1. Two model selection rules $\hat{k}$ and k^* are said to be equivalent if $P_v(\hat{k} = k^*) = 1$ for some $v \in V_0$ or, equivalently, for all $v \in V_0$. Let the model selection rule $\bar{k}$ be chosen to minimize

$$\| \bar{v}_k - \bar{v}_0 \|^2 + C_k = \| P_k Z - Z \|^2 + C_k.$$

Let dv refer to Lebesgue measure on $V_0 = \mathbb{R}^p$ and set $|A| = \int_A dv$ for $A \subset V_0$. Let ρ_A denote the uniform probability distribution on A defined by $\rho_A(dv) = |A|^{-1} I_A(v) dv$.

Let B denote the closed unit ball $\{v : |v| \le 1\}$ in V_0. For each $u \in \partial B$ there is an $\epsilon \in (0, 1)$ such that if $1 \le k \le K$ and $u \notin V_k$, then

$$u \cdot v < 1 - \epsilon, \qquad v \in V_k \cap B.$$

For such a choice of ϵ, $\{v : u \cdot v > 1 - \epsilon\}$ is an open neighborhood of u. The collection of all such neighborhoods forms an open covering of the compact set ∂B, which therefore has a finite subcovering consisting of I open neighborhoods corresponding as above to $u_i \in \partial B$ and $\epsilon_i \in (0, 1)$ for $1 \le i \le I$: Thus

$$\partial B \subset \bigcup_{i=1}^{I} \{v : u_i \cdot v > 1 - \epsilon_i\}.$$

Set

$$A = \{v \in V_0 : u_i \cdot v \le 1 - \epsilon_i \quad \text{for} \quad 1 \le i \le I\}.$$

Then A is a closed convex polyhedron in V_0 which contains the origin as an interior point and is disjoint from ∂B. Thus A is contained in the interior of B and hence A is a compact convex polyhedron. The next result is Theorem 1 applied to this set A and the model of the present section.

THEOREM 1′: (i) *For some* $\delta > 0$,

$$r(\rho_{cA}, \bar{k}) - r(\rho_{cA}) = o(e^{-\delta c^2}) \qquad \text{as} \quad c \to \infty.$$

(ii) *If* $\hat{k}$ *is any model selection rule which is not equivalent to* $\bar{k}$, *then*

$$\liminf_{c \to \infty} c^p [r(\rho_{cA}, \hat{k}) - r(\rho_{cA}, \bar{k})] > 0.$$

In order to reduce Theorem 1 to Theorem 1′, let the basis $m_1, \cdots, m_p$ of M_0 and real numbers m_{ji}, $1 \le j, i \le n$, be such that $m_{ji} = m_j(x_i)$ for $1 \le j \le p$ and $1 \le i \le n$ and

$$\sum_{i=1}^{n} m_{ji} m_{li} = 1 \qquad \text{if} \quad j = l,$$

$$= 0 \qquad \text{if} \quad j \ne l.$$

For $1 \leq k \leq K$ let V_k denote the collection of points $v = (v_1, \cdots, v_p) \in V_0$ such that $v_1 m_1 + \cdots + v_p m_p \in M_k$ or, equivalently, such that

$$v_j = \sum_{i=1}^n m_j(x_i) m(x_i), \qquad 1 \leq j \leq p,$$

for some $m \in M_k$. Let $H = (H_{jl})$ denote the positive definite symmetric matrix defined by

$$H_{jl} = \int m_j(x) m_l(x) \mu(dx), \qquad 1 \leq j, l \leq p.$$

Finally set

$$Z_j = \sum_{i=1}^n m_j(x_i) Y_i, \qquad 1 \leq j \leq p,$$

and

$$W_j = \sum_{i=1}^n m_{j+p,i} Y_i, \qquad l \leq j \leq n - p.$$

3. Proof of Theorem 1'. It will first be shown that (i) implies (ii). For $z \in V_0$, $\bar{k}(z)$ is a value of k which minimizes $\| P_k z - z \|^2 + C_k$. For almost every z there is a unique such k. Ties, which occur for z in a set of Lebesgue measure zero, can be broken, say, by minimizing k. Thus $\bar{k}(z)$ is well defined for all z.

Let $\hat{k}$ be any model selection rule. It can be thought of as a randomized function of Z having probability $\pi_k(z)$ of taking on the value $k \in \{0, \cdots, K\}$ when $Z = z$. Suppose $\hat{k}$ is not the equivalent to $\bar{k}$. Then there is a compact subset D of V_0 such that

$$\int_D \{\sum_{k=0}^K \pi_k(z)(\| P_k z - z \|^2 + C_k) - \| P_{\bar{k}(z)} z - z \|^2 - C_{\bar{k}(z)}\} \, dz > 0. \tag{3.1}$$

Set $\rho_c = \rho_{cA}$, let $\rho_c(\cdot \mid z)$ denote the corresponding posterior density given that $Z = z$, let f_c denote the corresponding marginal density of Z and let N denote the normal density on V_0 given by

$$N(z) = \frac{1}{(\sigma \sqrt{2\pi})^p} e^{-|z|^2/2\sigma^2}.$$

Then

$$f_c(z) = \frac{1}{|cA|} \int I_{cA}(v) N(z - v) \, dz$$

and

$$\rho_c(v \mid z) = \frac{I_{cA}(v) N(z - v)}{\int I_{cA}(v) N(z - v) \, dz}.$$

Note that $|cA| = c^p |A|$. It is easily seen that

$$\lim_{c \to \infty} c^p |A| f_c(z) = 1 \tag{3.2}$$

and

$$\lim_{c \to \infty} \int \| P_k z - v \|^2 \rho_c(v \mid z) \, dv = \| P_k z - z \|^2 + \int \| v \|^2 N(v) \, dv, \qquad k \in \{0, \cdots, K\}, \tag{3.3}$$

both limits being uniform for z in the compact set D. Set

$$G_c(z, k) = \int \| P_k z - v \|^2 \rho_c(v \mid z)\, dv + C_k.$$

Let k_c be the Bayes model selection rule corresponding to ρ_c, defined so that

$$G_c(z, k_c(z)) = \min_{0 \le k \le K} G_c(z, k).$$

Then $r(\rho_c, k_c) = r(\rho_c)$. Now

$$r(\rho_c, \hat{k}) - r(\rho_c, \bar{k}) = \int \{\textstyle\sum_{k=0}^{K} \pi_k(z) G_c(z, k) - G_c(z, \bar{k}(z))\}\, f_c(z)\, dz,$$

from which it follows that

$$r(\rho_c, \hat{k}) - r(\rho_c, \bar{k}) \ge \int_D \{\textstyle\sum_{k=0}^{K} \pi_k(z) G_c(z, k) - G_c(z, \bar{k}(z))\}\, f_c(z)\, dz - \{r(\rho_c, \bar{k}) - r(\rho_c)\}.$$

Consequently, by (3.2) and (3.3),

$$\begin{aligned} &\liminf_{c \to \infty} c^d \{r(\rho_c, \hat{k}) - r(\rho_c, \bar{k})\} \\ &\quad \ge \int_D \{\textstyle\sum_{k=0}^{K} \pi_k(z)(\| P_k z - z \|^2 + C_k) - \| P_{\bar{k}(z)} z - z \|^2 - C_{\bar{k}(z)}\}\, dz \\ &\qquad\qquad - \limsup_{c \to \infty} c^d \{r(\rho_c, \bar{k}) - r(\rho_c)\}. \end{aligned}$$

Thus (i) implies (ii), as desired.

To complete the proof of Theorem 1′, (i) will now be verified. Let $\mu_c(z) = \int v \rho_c(v \mid z)\, dv$ denote the mean of the posterior density $\rho_c(\cdot \mid z)$. Then

$$r(\rho_c, \bar{k}) - r(\rho_c) = \int F_c(z) f_c(z)\, dz, \tag{3.4}$$

where

$$F_c(z) = \| P_{\bar{k}(z)} z - \mu_c(z) \|^2 + C_{\bar{k}(z)} - \| P_{k_c(z)} z - \mu_c(z) \|^2 - C_{k_c(z)}.$$

By the definition of $\bar{k}$

$$\| P_{\bar{k}(z)} z - z \|^2 + C_{\bar{k}(z)} \le \| P_{k_c(z)} z - z \|^2 + C_{k_c(z)}.$$

Therefore

$$\begin{aligned} F_\rho(z) &\le \{\| P_{\bar{k}(z)} z - \mu_c(z) \|^2 - \| P_{\bar{k}(z)} z - z \|^2\} \\ &\quad - \{\| P_{k_c(z)} z - \mu_c(z) \|^2 - \| P_{k_c(z)} z - z \|^2\} \\ &= 2(P_{k_c(z)} z - z, \mu_c(z) - z) - 2(P_{\bar{k}(z)} z - z, \mu_c(z) - z). \end{aligned}$$

Relative to the inner product "$\cdot$" on $\mathbb{R}^p$, let Q_k and U_k denote respectively the orthogonal projection from V_0 onto V_k and onto the orthogonal complement of V_k in V_0. Then Q_0 is the identity transformation on V_0 and $U_0 = 0$. Now

$$\mu_c(z) - z - U_k(\mu_c(z) - z) = Q_k(\mu_c(z) - z) \in V_k,$$

so

$$(P_k z - z, \mu_c(z) - z) = (P_k z - z, U_k(\mu_c(z) - z)).$$

Thus by Schwarz's inequality

$$|(P_k z - z, \mu_c(z) - z)| \le \| P_k z - z \| \, \| U_k(\mu_c(z) - z) \| \le \| z \| \, \| U_k(\mu_c(z) - z) \|.$$

Let η be a positive constant such that

$$\eta^{-1}|v| \leq \|v\| \leq \eta|v|, \qquad v \in \mathbb{R}^p.$$

Then

$$F_c(z) \leq 2\eta^2|z|\{|U_{k_c(z)}(\mu_c(z) - z)| + |U_{\bar{k}(z)}(\mu_c(z) - z)|\}. \tag{3.5}$$

Three preparatory lemmas will be obtained in order to show that (i) follows from (3.4) and (3.5). For $\delta > 0$ set

$$A^\delta = \{v \in V_0 : |v - a| \leq \delta \quad \text{for some} \quad a \in A\}$$

and

$$cA^\delta = \{cv : v \in A^\delta\} = \{v \in V_0 : |v - a| \leq c\delta \quad \text{for some} \quad a \in cA\}.$$

LEMMA 1. *For every $\delta > 0$ there is a $\delta' > 0$ such that*

$$\int_{V_0 \setminus cA^\delta} |z|^2 f_c(z)\, dz = o(e^{-\delta' c^2}) \qquad \text{as} \quad c \to \infty.$$

PROOF. Observe that

$$\begin{aligned}
\int_{V_0 \setminus cA^\delta} |z|^2 f_c(z)\, dz &= \frac{1}{|cA|} \int_{cA} dv \int_{V_0 \setminus cA^\delta} |z|^2 N(z - v)\, dz \\
&\leq \frac{1}{|cA|} \int_{cA} dv \int_{|z-v| > c\delta} |z|^2 N(z - v)\, dz \\
&= \frac{1}{|cA|} \int_{cA} dv \int_{|z| > c\delta} (|z|^2 + |v|^2) N(z)\, dz \\
&= o(e^{-\delta' c^2})
\end{aligned}$$

for some $\delta' > 0$, as desired.

For $\gamma > 0$ let $B_\gamma(z)$ denote the ball $\{v : |v - z| \leq \gamma\}$ in V_0.

LEMMA 2. *Let $0 < \delta < \epsilon$. Then for c sufficiently large*

$$|\mu_c(z) - z| \leq c\epsilon, \qquad z \in cA^\delta.$$

PROOF. Choose γ and η with $\delta < \gamma < \eta < \epsilon$. Now A is convex and $|A| > 0$, from which it follows easily that for some $a > 0$

$$|A \cap B_\gamma(z)| \geq a, \qquad z \in A^\delta.$$

Thus

$$|cA \cap B_{c\gamma}(z)| \geq ac^p, \qquad z \in cA^\delta,$$

and hence

$$\begin{aligned}
\int_{cA \cap B_{c\eta}(z)} e^{-|z-v|^2/2\sigma^2}\, dv &\geq \int_{cA \cap B_{c\gamma}(z)} e^{-|z-v|^2/2\sigma^2}\, dv \\
&\geq ac^p e^{-c^2\gamma^2/2\sigma^2}, \qquad z \in cA^\delta.
\end{aligned}$$

Observe that for $j = 0, 1$

$$\int_{cA\setminus B_{c\eta}(z)} |v - z|^j e^{-|z-v|^2/2\sigma^2}\, dv = o(e^{-c^2\gamma^2/2\sigma^2})$$

$$= o\left(\int_{cA\cap B_{c\eta}(z)} e^{-|z-v|^2/2\sigma^2}\, dv\right)$$

uniformly over $z \in cA^\delta$. Observe also that

$$\frac{\int_{cA\cap B_{c\eta}(z)} |v - z|\, e^{-|z-v|^2/2\sigma^2}\, dv}{\int_{cA\cap B_{c\eta}(z)} e^{-|z-v|^2/2\sigma^2}\, dv} \le c\eta, \qquad z \in cA^\delta.$$

Observe finally that for all z,

$$\mu_c(z) - z = \frac{\int_{cA} (v - z)e^{-|z-v|^2/2\sigma^2}\, dv}{\int_{cA} e^{-|z-v|^2/2\sigma^2}\, dv}.$$

The desired conclusion follows easily from these three observations.

LEMMA 3. *For some* $\delta > 0$

$$\max\{|U_k(\mu_c(z) - z)| : z \in cA^\delta, 1 \le k \le K \quad \text{and} \quad |U_k z| \le c\delta\} = o(e^{-\delta c^2}) \qquad \text{as} \quad c \to \infty.$$

PROOF. Recall from Section 2 that

$$A = \{v \in V_0 : u_i \cdot v \le 1 - \epsilon_i \quad \text{for} \quad 1 \le i \le I\} \subset \{v \in V_0 : |v| < 1\},$$

where $0 < \epsilon_i < 1$, $u_i \in V_0$ and $|u_i| = 1$ for $1 \le i \le I$. Also if $1 \le i \le I$, $1 \le k \le K$, $u_i \notin V_k$, $v \in V_k$ and $|v| \le 1$, then $u_i \cdot v < 1 - \epsilon_i$.

Suppose $1 \le k \le K$ and let $z \in \partial A \cap V_k$ be fixed. Let i_l, $1 \le l \le L$, be values of i such that $1 \le i \le I$ and $u_i \cdot z = 1 - \epsilon_i$. Then $u_{i_l} \in V_k$ for $1 \le l \le L$ and there is an open neighborhood $\mathcal{N}$ of z such that

$$\mathcal{N} \cap A = \mathcal{N} \cap \{v : u_{i_l} \cdot v \le 1 - \epsilon_{i_l} \quad \text{for} \quad 1 \le l \le L\}.$$

By compactness there is a $\delta > 0$ such that if $z \in A^\delta$ and $|U_k z| \le \delta$, then

$$A \cap B_{2\delta}(z) = A_k(z) \cap B_{2\delta}(z),$$

where $A_k(z)$ is of the form

$$A_k(z) = \{v : u_{i_l} \cdot v \le 1 - \epsilon_{i_l} \quad \text{for} \quad 1 \le l \le L\};$$

here L and $u_{i_l} \in V_k$, $1 \le l \le L$, depend on z. If $z \in cA^\delta$ and $|U_k z| \le c\delta$, then

$$cA \cap B_{2c\delta}(z) = cA_k(c^{-1}z) \cap B_{2c\delta}(z).$$

Consequently, by an argument similar to that used in the proof of the previous lemma, if $\delta > 0$ is sufficiently small, then uniformly over $1 \le k \le K$, $z \in cA^\delta$ and $|U_k z| \le c\delta$,

$$U_k(\mu_c(z) - z) = \frac{\int_{cA} U_k(v - z)e^{-|z-v|^2/2\sigma^2}\, dv}{\int_{cA} e^{-|z-v|^2/2\sigma^2}\, dv}$$

$$= \frac{\int_{cA_k(c^{-1}z)} U_k(v - z)e^{-|z-v|^2/2\sigma^2}\, dv}{\int_{cA_k(c^{-1}z)} e^{-|z-v|^2/2\sigma^2}\, dv} + o(e^{-\delta c^2}).$$

It is clear from the form of $A_k(z)$ that

$$\int_{cA_k(c^{-1}z)} U_k(v - z)e^{-|z-v|^2/2\sigma^2}\, dv = 0.$$

Thus the conclusion of the lemma is valid.

The proof of Theorem 1′ will now be completed. By (3.4) it must be shown that for some $\delta > 0$

$$\int F_c(z)f_c(z)\, dz = o(e^{-\delta c^2}).$$

It is easily seen that $\mu_c(z) = O(|z|)$ as $|z| \to \infty$ uniformly in c and hence that $F_c(z) = O(|z|^2)$ as $|z| \to \infty$ uniformly in c. Thus by Lemma 1 for every $\delta > 0$ there is a $\delta' > 0$ such that

$$\int_{V_0 \setminus cA^\delta} F_c(z)f_c(z)\, dz = o(e^{-\delta c^2}).$$

By (3.5) it suffices to verify that for some $\delta > 0$

$$\sup_{z \in cA^\delta} |U_{\bar{k}(z)}(\mu_c(z) - z)| = o(^{-\delta c^2}) \tag{3.6}$$

and

$$\sup_{z \in cA^\delta} |U_{k_c(z)}(\mu_c(z) - z)| = o(e^{-\delta c^2}). \tag{3.7}$$

To verify (3.6) observe first that if $\bar{k}(z) = 0$, then

$$U_{\bar{k}(z)}(\mu_c(z) - z) = U_0(\mu_c(z) - z) = 0.$$

Suppose $\bar{k}(z) \neq 0$. Then

$$|U_{\bar{k}(z)}z|^2 = |Q_{\bar{k}(z)}z - z|^2 \leq |P_{\bar{k}(z)}z - z|^2 \leq \eta^2 \|P_{\bar{k}(z)}z - z\|^2 \leq \eta^2(C_0 - C_{\bar{k}(z)}).$$

Consequently $|U_{\bar{k}(z)}z| \leq c\delta$ for c sufficiently large, so (3.6) follows from Lemma 3.

To prove (3.7) observe that if $k_c(z) = 0$, then $U_{k_c(z)}(\mu_c(z) - z) = 0$. Suppose $z \in cA^\delta$ and $k_c(z) \neq 0$. Then

$$\begin{aligned}\|P_{k_c(z)}z - \mu_c(z)\|^2 &\leq \|z - \mu_c(z)\|^2 + C_0 - C_{k_c(z)} \\ &\leq \eta^2 |z - \mu_c(z)|^2 + C_0 - C_{k_c(z)}.\end{aligned}$$

Choose $\epsilon > 0$. By Lemma 2 it can be assumed that $|\mu_c(z) - z| \leq c\epsilon$. Therefore

$$\begin{aligned}|U_{k_c(z)}z| &\leq |Q_{k_c(z)}z - z| \\ &\leq |P_{k_c(z)}z - z| \\ &\leq |P_{k_c(z)}z - \mu_c(z)| + |\mu_c(z) - z| \\ &\leq \eta \|P_{k_c(z)}z - \mu_c(z)\| + |\mu_c(z) - z| \\ &\leq \eta(\eta^2\epsilon^2c^2 + C_0 - C_{k_c(z)})^{1/2} + c\epsilon.\end{aligned}$$

Thus (3.7) also follows from Lemma 3. This completes the proof of Theorem 1′.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. Second International Symposium on Information Theory*. 267–281, (B. N. Petrov and F. Csáki, eds.). Akadémiai Kaidó, Budapest.

AKAIKE, H. (1974). A new look at statistical model identification. *IEEE Trans. Automatic Control* **AC-19** 716–723.

ANDERSON, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.* **33** 255–265.

BANCROFT, T. A. and HAN, C.-P. (1977). Inference based on conditional specification: A note and a bibliography. *Internat. Statist. Rev.* **45** 117–127.

BLALOCK, H. M. JR. (1961). *Causal Inferences in Nonexperimental Research.* University of North Carolina, Chapel Hill.

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799.

BREIMAN, L. and STONE, C. J. (1978). Parsimonious binary classification trees. Technical Note TSC-CSD-TN-004, Technology Service Corporation, Santa Monica, California.

BROOKS, R. J. (1973). The choice of variables for prediction in curvilinear multiple regression. *Ann. Statist.* **1** 506–516.

DEMPSTER, A. P. (1971). Model searching and estimation in the logic of inference (with discussion) 56-81. In *Foundations of Statistical Inference*, (V. P. Godambe and D. A. Sprott, eds.). Holt, Rinehart and Winston, Toronto.

FADEN, A. M. and RAUSSER, G. C. (1976). Econometric policy model construction: The post-Bayesian approach. *Ann. Economic and Social Measurement* **5** 349–362.

GAVER, K. M. and GEISEL, M. S. (1974). Discriminating among alternative models: Bayesian and non-Bayesian methods. 49–77. In *Frontiers in Econometrics*, (P. Zarembka, ed.). Academic, New York.

GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **74** 153–160.

HALPERN, E. F. (1973). Polynomial regression from a Bayesian approach. *J. Amer. Statist. Assoc.* **68** 137–143.

HOCKING, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32** 1–49.

HODGES, J. L., JR. and LEHMANN, E. L. (1954). Testing the approximate validity of statistical hypotheses. *J. Roy. Stat. Soc. Ser. B.* **16** 261–268.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 361–379.

KIEFER, J. (1968). Pages 139–142 in *The Future of Statistics: Proceedings of a Conference on the Future of Statistics Held at the University of Wisconsin, Madison, Wisconsin, June 1967*, (P. G. Watts, ed.). Academic, New York.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inferences with Non-experimental Data.* Wiley, New York.

LINDLEY, D. V. (1968). The choice of variables in multiple regression (With discussion). *J. Roy. Statist. Soc. Ser. B.* **30** 31–66.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MEEDEN, G. and ARNOLD, B. C. (1979). The admissibility of a preliminary test estimator when the loss incorporates a complexity cost. *J. Amer. Statist. Assoc.* **74** 872–874.

PEREIRA, B. DE B. (1977). Discriminating among several models: A Bibliography. *Internat. Statist. Rev.* **45** 163–172.

RAMSEY, J. B. (1974). Classical model selection through specification error tests. 13–47. In *Frontiers in Econometrics.* (P. Zarembka, ed.). Academic, New York.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. (With discussion). *J. Roy. Stat. Soc. Ser. B.* **36** 111–147.

THOMPSON, M. L. (1978). Selection of variables in multiple regression: Part I. A review and evaluation. *Internat. Statist. Rev.* **46** 1–19. Part II. Chosen procedures, computations and examples. *Ibid.* **46** 129–146.

TUKEY, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics* **3** 191–219.

CHARLES J. STONE
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, LOS ANGELES
LOS ANGELES, CALIFORNIA 90024