The Annals of Statistics
2000, Vol. 28, No. 6, 1620–1637

ASYMPTOTICALLY MINIMAX REGRET PROCEDURES IN REGRESSION MODEL SELECTION AND THE MAGNITUDE OF THE DIMENSION PENALTY

BY ALEXANDER GOLDENSHLUGER AND EITAN GREENSHTEIN

University of Haifa and Technion

This paper addresses the topic of model selection in regression. We emphasize the case of two models, testing which model provides a better prediction based on n observations. Within a family of selection rules, based on maximizing a penalized log-likelihood under a normal model, we search for asymptotically minimax rules over a class $\mathscr{G}$ of possible joint distributions of the explanatory and response variables. For the class $\mathscr{G}$ of multivariate normal joint distributions it is shown that asymptotically minimax selection rules are close to the AIC selection rule when the models' dimension difference is large. It is further proved that under fairly mild assumptions on $\mathscr{G}$, any asymptotically minimax sequence of procedures satisfies the condition that the difference in their dimension penalties is bounded as the number of observations approaches infinity. The results are then extended to the case of more than two competing models.

1. Introduction. Let $V = (Y, X_1, \dots, X_m),\ m \le \infty$ be a random vector with distribution G. We will refer to $X_1, \dots, X_m$ as the explanatory variables and to Y as the corresponding response variable. Suppose it is desired to construct a predictor for Y based on a subset $X_{l_1}, \dots, X_{l_k},\ k \le m$ of the explanatory variables. Under a squared error prediction loss, an optimal predictor is $\widehat{Y}_0 = E(Y|X_{l_1}, \dots, X_{l_k}) = f(X_{l_1}, \dots, X_{l_k})$. Typically G and, consequently, f are not known, and a main object is to develop a method for estimating f. Such a method usually involves the following steps: assume a model, that is, a collection of conditional distributions $\{G_\omega(Y|X_1, \dots, X_m),\ \omega \in \Omega\}$, estimate ω by $\hat{\omega}$ and let

$$\hat{f}(X_{l_1}, \dots, X_{l_k}) = E_{\hat{\omega}}(Y|X_{l_1}, \dots, X_{l_k}).$$

Given various possible models $\{G^j_\omega(Y|X_1, \dots, X_m),\ \omega \in \Omega_j\},\ j = 1, 2, \dots, J$, and a set of i.i.d. observations $V(t) = (Y(t), X_1(t), \dots, X_m(t));\ t = 1, \dots, n$, an important question is which model to choose. Such a choice determines a predictor as explained above. Thus, in our setting, choosing a model or a predictor method are synonymous. It is well known, and will be seen in the sequel, that the answer to the question, "Which is the most appropriate model (or, equivalently, prediction method) under a given G?" depends on the number of observations n.

Various approaches to model selection yield various criteria. Mallows' C_p [Mallows (1973)] and Akaike's AIC [Akaike (1974)] criteria are motivated by

Received December 1998; revised July 2000.
AMS 2000 *subject classifications*. 62J05, 62C20.
Key words and phrases. Model selection, regression, minimax procedures.

achieving a good prediction. Other related methods, based on cross-validation, were suggested by Stone (1974) and Geisser (1975). Schwarz (1978) analyzed the situation where there is a prior probability on the models, prior distribution on the parameters within the models and a 0–1 loss for selecting a wrong model. Rissanen (1989) gave a criterion motivated by information theory, trying to balance the trade-off between efficient coding using the most appropriate distribution from rich parametric families and the complexity of coding the parameters of such families. Hannan and Quinn (1979) characterize selection methods which are consistent, that is, choose the most parsimonious among the correct models with probability tending to 1 as the number of observations tends to infinity. Most of the results and formulations (rather than heuristics) are derived under the assumption that at least one of the models is correct; that is, for some j_0, $G(\cdot|\cdot) \in \{G^{j_0}_\omega(Y|X_1, \dots, X_m),\ \omega \in \Omega_{j_0}\}$.

Denote $g_\omega(v)$ the conditional density $g_\omega(y|x_1, \dots, x_m)$ under $G_\omega(\cdot|\cdot)$. Let $g^n_\omega(v_1, \dots, v_n) = \prod_{i=1}^n g_\omega(v_i)$. Many of the procedures of model selection amount to choosing the model j_0 that maximize over $j = 1, \dots, J$,

$$\max_{\omega \in \Omega_j} \log(g^n_\omega(v_1, \dots, v_n)) - C_n(j). \tag{1}$$

for an appropriate choice of penalties $C_n(j)$.

Let p_j be the dimension of the parameter set in model j. The value of $C_n(j)$ suggested by Akaike (1974) is $C_n(j) = p_j$, while Schwarz's and Rissanen's selection criteria are determined by $C_n(j) = \frac{1}{2}\log(n)p_j$. Hannan and Quinn (1979) showed that in order to get consistency of a sequence of model selection procedures in some settings the following should hold: lim inf$_{n\to\infty}$(2 log log$(n)p_j)^{-1}C_n(j) > 1$, and $\limsup_{n\to\infty} n^{-1}C_n(j) = 0$. The method of cross-validation and Mallow's C_p were shown to be asymptotically equivalent to the procedure of Akaike in some cases [Stone (1977)]. For comprehensive surveys on model selection see Linhart and Zuchinni (1986), Shibata (1989) and Shao (1997); the last paper is especially relevant since it deals with regression models. Other papers dealing with regression model selection are Oliker (1978), Thompson (1978), Stone (1981), Shibata (1981), Breiman and Freedman (1983), Nishii (1984), Speed and Yu (1993), Foster and George (1994).

The large difference in the magnitude of the values of $C_n(j)$, suggested by different yet very appealing approaches, should still be understood. It motivates many of the above-mentioned papers. The purpose of this work is to give some further insight and perspective to this issue. Our focus on the procedures that are based on penalized log-likelihood is motivated by many papers dealing with such procedures; there are, of course, other appealing types of procedures. In our formulation we do not necessarily assume that one of the models is correct. The assumptions made in a model are, in our view, only a means for determining meaningful and mathematically tractable predictors. We will examine performance of selection procedures with respect to a collection $\mathscr{G}$ of possible joint distributions G, in the spirit of the theory of robust statistics. The reason why one of the models is not simply taken as $\mathscr{G}$ is that the set $\mathscr{G}$

could be large, not finitely parameterized and mathematically intractable to induce meaningful predictors.

We suggest a novel decision theoretic approach to the problem of selecting a model. We emphasize the case of two competing regression models and test which model provides a better prediction based on n observations. Among the tests with penalties $C_n(j)$ as in (1) and the log-likelihood under a normal model, the asymptotically minimax tests (selection procedures) are characterized. In Section 3 we prove that in the case, where $\mathscr{G}$ is the set of all multivariate normal distributions, the minimax procedure is equivalent to Akaike's criterion asymptotically as d and n go to infinity; here d is the difference of the models' dimensions. In Section 4 we show that under fairly mild assumptions on $\mathscr{G}$ any asymptotically minimax rule must satisfy the condition that the difference in penalty terms is bounded as the number of observations approaches infinity. Finally we indicate how the results are extended to the case of several competing models.

2. The minimax criterion.

We formulate the case of two competing models. Let $V(t) = (Y(t), X_1(t), \dots, X_m(t))$ be i.i.d. vectors $V(t) \sim G$, $t = 1, 2, \dots$ and let $\{G^j_\omega,\ \omega \in \Omega_j\}$, $j = \mathrm{I}, \mathrm{II}$ be two competing models. Let $\hat{\omega}^j_n(V(1), \dots, V(n))$, $j = \mathrm{I}, \mathrm{II}$, $n = 1, 2, \dots$ be two sequences of estimators (say MLE) based on the two models. Let $\widehat{Y}^j(n)$; $j = \mathrm{I}, \mathrm{II}$, $n = 1, 2, \dots$ be two sequences of predictors,

$$\widehat{Y}^j(n+1) = E_{\hat{\omega}^j_n}\big[Y(n+1) | X_1(n+1), \dots, X_m(n+1)\big].$$

Define θ^j_n, the expected squared error loss in a prediction of a future observation $Y(n+1)$,

$$\theta^j_n = E_{G^{n+1}}\big[Y(n+1) - \widehat{Y}^j(n+1)\big]^2, \qquad j = \mathrm{I}, \mathrm{II}. \tag{2}$$

Here G^{n+1} is the $(n+1)$th product of the measure G. Denote

$$\theta_n = \theta^{\mathrm{I}}_n - \theta^{\mathrm{II}}_n.$$

The dependence of θ_n on G is suppressed.

From the point of view of prediction, the sequence of model selection problems determined by the observations $V(1), V(2), \dots$ may be thought of as the following sequence of testing problems:

$$H^n_{\mathrm{I}}\colon \theta_n \le 0 \quad \text{versus} \quad H^n_{\mathrm{II}}\colon \theta_n > 0, \qquad n = 1, 2, \dots .$$

Deciding H^n_j, $j = \mathrm{I}, \mathrm{II}$ should be understood as selecting model j at stage n.

Define the loss function

$$L_n(\theta_n, H^n_j) = \begin{cases} l(\theta_n), & \text{if } \theta_n \notin H^n_j, \\ 0, & \text{otherwise,} \end{cases}$$

where $l(\cdot)$ is a symmetric around zero and nondecreasing function on the positive real line. From the class of such functions $l(\cdot)$, the function $l(\theta) = |\theta|$ is of a particular interest. It is consistent with the squared error prediction loss that defines the parameterization θ_n^j, $j = \mathrm{I}, \mathrm{II}$. However, we find it instructive and worthwhile to carry out the general development and, in particular, to consider zero–one loss.

Let $\{\delta_n\} = \{\delta_n(V(1), \dots, V(n))\}$ be a sequence of selection or (equivalently) testing procedures. For $G \in \mathscr{G}$ define

$$R_n(G, \delta_n) = R_n(\theta_n(G), \delta_n) = E_{G^n} L_n(\theta_n, \delta_n),$$

and let $r_n(\delta_n) = \sup_{G \in \mathscr{G}} R_n(G, \delta_n)$. For a given loss function $l(\cdot)$, a sequence of collections, Δ_n, of possible selection rules based on n observations and a collection of distributions $\mathscr{G}$, we define an asymptotically minimax selection selection procedure as follows. Denote $r_n^* = \inf_{\delta_n \in \Delta_n}[r_n(\delta_n)]$.

DEFINITION 1. A sequence of selection procedures $\{\delta_n\}$ is called asymptotically minimax if $\lim_{n\to\infty}[r_n(\delta_n)/r_n^*] = 1$.

We will study collections of procedures, denoted Δ_n^c, that are defined by penalties $C_n(j)$, in the following way: select the model that maximizes over $j = \mathrm{I}, \mathrm{II}$,

$$\max_{\omega \in \Omega_j} \log[g_\omega^n(V(1), \dots, V(n))] - C_n(j).$$

Actually only the difference $C_n = C_n(\mathrm{I}) - C_n(\mathrm{II})$ matters.

Note that our minimax formulation uses regrets rather than the actual prediction risks. An alternative minimax approach would be to select the model that minimizes the maximal over $G \in \mathscr{G}$ mean squared prediction error. This formulation often leads to the trivial selection rule that always chooses the larger model. We refer to Shibata (1986) for some related results on using regrets in regression model selection.

3. The class of multivariate normal distributions. In this section we will assume that $\mathscr{G}$ is the collection of all multivariate normal distributions. A study of prediction in this setting was conducted by Oliker (1978), Thompson (1978) and by Breiman and Freedman (1983).

We assume two possible nested competing regression models

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2), \quad i = k, m, \ k < m,$$

where σ^2 is unknown. We will refer to the models determined by m and k variables as models I and II, respectively. Our goal is to find a sequence of constants $C_n^0(j)$, $j = \mathrm{I}, \mathrm{II}$, $n = 1, 2, \dots$ that determines a sequence of selection procedures which is asymptotically minimax within sequences such that $\delta_n \in \Delta_n^c$.

It may be checked that a minimax value and a minimax procedure under the class of all multivariate normal distributions are the same as under the class of all multivariate normal distributions with independent $X_i, i = 1, \dots, m$ such that $E(X_i) = 0$, $EX_i^2 = 1$, and

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_m X_m + \varepsilon_G.$$

Here $\varepsilon_G \sim N(0, \sigma_G^2)$ is independent of $X_1, \dots, X_m$. For details, see Lemma 2.1 in Breiman and Freedman (1983). We will assume the later class of joint distributions. Denote $\tau^2 = \tau^2(G) = \sum_{j=k+1}^m \alpha_j^2$.

LEMMA 1.

$$\theta_n^{\mathrm{I}} = \frac{n+1}{n}\left(1 + \frac{m}{n-m-2}\right)\sigma_G^2,$$
$$\theta_n^{\mathrm{II}} = \frac{n+1}{n}\left(1 + \frac{k}{n-k-2}\right)(\sigma_G^2 + \tau^2).$$

The proof follows from equation (6) in the paper of Oliker (1978), or from Theorem 1.1 in Breiman and Freedman (1983). Note that we use the parametrization under model I.

From the lemma we obtain

$$\theta_n = \sigma_G^2 \frac{d}{n} - \tau^2\left(1 + \frac{k}{n}\right) + o(n^{-1}), \tag{3}$$

$$\tau^2 \approx \sigma_G^2 \frac{d}{n} \qquad \text{iff, } \theta_n = o(n^{-1}), \tag{4}$$

where $d = m - k$.

First we will consider the case of a 0–1 loss function $l(\theta)$. In this case the minimax test is a rule that attains its maximum for $\theta_n \approx 0$ or, equivalently, when $\tau^2 \approx \sigma_G^2 dn^{-1}$; the minimax value is 1/2. The purpose of the following is to:

1. calculate P_G(rejecting H_{I}^n) for procedures in Δ_n^c;
2. find $C_n(j)$ such that the corresponding procedure satisfies

$$P_{\tau^2 = \sigma_G^2 dn^{-1}}(\text{rejecting } H_{\mathrm{I}}^n) \approx \tfrac{1}{2}.$$

Let X be the random matrix $(X_i(t))$, $t = 1, \dots, n$, $i = 1, \dots, m$, and denote $Y' = (Y(1), \dots, Y(n))$. Let $\overline{Y}_{\mathrm{I}}$ and $\overline{Y}_{\mathrm{II}}$ be the projections of Y on the first m and k columns of X, respectively. Then the ANOVA identity is

$$\|Y - \overline{Y}_{\mathrm{II}}\|^2 = \|Y - \overline{Y}_{\mathrm{I}}\|^2 + \|\overline{Y}_{\mathrm{I}} - \overline{Y}_{\mathrm{II}}\|^2, \qquad j = \mathrm{I}, \mathrm{II}, \tag{5}$$

denoted $U_1^2 = U_2^2 + U_3^2$. Notice that for the two models I, II,

$$\max_{\omega \in \Omega_j} \log(g_\omega^n(V(1), \dots, V(n)))$$
$$= -\frac{n}{2}\log\left(\frac{1}{n}\|Y - \overline{Y}_j\|^2\right) + \text{constant}, \qquad j = \mathrm{I}, \mathrm{II}.$$

Denote $C_n = C_n(\mathrm{I}) - C_n(\mathrm{II})$, and suppose that $C_n = o(n)$ as n goes to infinity. We obtain

$$\begin{aligned} &P_G(\text{rejecting } H_I^n) \\ &\quad = P_G(\log(U_2^2) > \log(U_2^2 + U_3^2) + \frac{2}{n}(C_n(\mathrm{II}) - C_n(\mathrm{I})) \\ &\quad = P_G(U_2^2 \ \exp(2C_n n^{-1}) > U_2^2 + U_3^2) \qquad (6) \\ &\quad = P_G((1 + 2C_n n^{-1})U_2^2 > U_2^2 + U_3^2) + o(1) \\ &\quad = P_G(U_3^2 < \sigma_G^2 2C_n) + o(1). \end{aligned}$$

Conditionally on $(X_1(t), \dots, X_m(t))$, $t = 1, \dots, n$, the distribution of $U_3^2 \sigma_G^{-2}$ is $\mathscr{X}^2$ with $d = m - k$ degrees of freedom and noncentrality parameter γ_n^2; unconditionally γ_n^2 is a random variable $\gamma_n^2 = n\tau^2\sigma_G^{-2}(1 + o_p(1))$. In the case where $\theta_n = o(n^{-1})$, recall from (4) that $\tau^2 = d\sigma_G^2 n^{-1}(1 + o(1))$, and then γ_n^2 converges in probability to d as n approaches infinity.

Denote by $\bar{m}_d$ the median of a $\mathscr{X}_d^2(d)$ distribution with d degrees of freedom and noncentrality parameter d.

THEOREM 1. *Let Ω_{I} and Ω_{II} be two nested linear models, where d is their dimension difference. Then for the class $\mathscr{G}$ of all multivariate normal distributions,* 0–1 *loss function $l(\cdot)$ and the class of sequences of procedures $\{\delta_n\}$ such that $\delta_n \in \Delta_n^c$, an asymptotically minimax sequence of procedures is determined by penalties $C_n(\mathrm{I}), C_n(\mathrm{II})$ satisfying*

$$C_n = C_n(\mathrm{I}) - C_n(\mathrm{II}) = \frac{\bar{m}_d}{2}.$$

The proof follows from (4), and the discussion following it, upon realizing that for a 0–1 loss the worst case is attained for a distribution G such that $\theta_n(G) \approx 0$, and the corresponding expected loss is 1/2.

Notice that for large values of d, $\bar{m}_d$ is close to $2d$ (the median is close to the mean), and we get the AIC criterion. This asymptotic phenomenon is more general as may be seen from the following Theorem 2. This theorem applies to the class of loss functions satisfying $l(\theta) = O(|\theta|^a)$; $a \ge 0$, as θ approaches infinity, but it is motivated mainly by the case $a = 1$.

In a situation where the number of observations is large, typically, candidate models will have large dimensions, and also the dimension difference will be large. The following theorem indicates that, for the class $\mathscr{G}$ of multivariate normal distributions, an appropriate penalty difference is the models' dimension difference, as in the AIC criterion. Consider a sequence of model selection problems (with different competing models at each stage). Let d^k denote the models' dimension difference in the kth problem, and suppose $\lim_{k\to\infty} d^k = \infty$. For every k, consider a sequence of penalty differences $C_n^{k,M}$ that determines an asymptotically minimax sequence of procedures denoted $\delta_n^{M,k}$. We denote the sequence of procedures induced by the penalties $C_n^k = d^k$ by $\delta_n^{d^k}$.

THEOREM 2. *Suppose $l(\cdot)$ satisfies $l(\theta) = O(|\theta|^a)$, $a \geq 0$, as θ approaches infinity. Suppose that $\mathscr{G}$ consists of all multivariate normal distributions G satisfying $\sigma_G^2 \leq \sigma_0^2 < \infty$ for some σ_0^2. Then*

$$\lim_{k\to\infty} \lim_{n\to\infty} \frac{C_n^{k,M}}{d^k} = \lim_{k\to\infty} \frac{d^k + o(\sqrt{d^k})}{d^k} = 1.$$

Before proving the theorem we present the following lemma.

LEMMA 2. *Let ξ_i, $i = 1, \ldots, d^k$ be i.i.d. normal random variables with mean μ_k and variance 1. Let $U^2 = \sum_{i=1}^{d^k} \xi_i^2$, and $\delta = d^k\mu_k^2 - d^k$; here $d^k\mu_k^2$ is the noncentrality parameter of U^2. Then*:

(i)

$$P(U^2 < 2d^k) \leq \left(1 - \frac{\delta}{2d^k + \delta}\right)^{d^k/2} \exp\left\{\frac{\delta d^k}{2(2d^k + \delta)}\right\}. \tag{7}$$

(ii) Suppose that $\mu_k \to 1$; then there exist constants $c > 0$ and $\alpha > 0$, such that for every d^k and $\delta < cd^k$,

$$P(U^2 < 2d^k) < \exp(-\alpha\delta^2/d^k). \tag{8}$$

PROOF. The first part is proved by applying the Chernoff bounding method to the noncentral chi-square random variable with d^k degrees of freedom and mean $2d^k + \delta$. Specifically, the moment generating function of a noncentral $\mathscr{X}^2$ distribution with d^k degrees of freedom and non-centrality parameter $d^k\mu_k^2$ is given by

$$\phi(s) = e^{-d^k\mu_k^2/2} \sum_{j=0}^{\infty} \left(\frac{d^k\mu_k^2}{2}\right)^j \frac{(1-2s)^{-(d^k+2j)/2}}{j!}.$$

Then the probability $P(U^2 < 2d^k)$ is bounded from above by $\phi(-s)\exp\{s(d^k + d^k\mu_k^2 - \delta)\}$ for any $s > 0$. Choosing $2s = \delta(d^k + d^k\mu_k^2 - \delta)^{-1}$ we obtain (7).

(ii) Inequality (8) follows from (7). See also Lemma 2.2(a) in Breiman and Freedman (1983). □

PROOF OF THEOREM 2. We take $\sigma_G^2 = 1$ in our derivation. Suppose that there exits $\varepsilon > 0$ such that for every k_0 there exists $k > k_0$ for which $\limsup_n (C_n^{k,M}/d^k) > 1 + \varepsilon$; w.l.o.g. assume that $C_n^{k,M} > (1+\varepsilon)d^k$ for large enough k and n. Also, w.l.o.g. there exists a sequence, $C_n^{k,M}$, satisfying the last inequality, such that $d^k = o(n)$; here the rate at which d^k/n approaches zero may be assumed arbitrarily fast. We will get a contradiction to the minimax property of the sequence $C_n^{k,M}$.

Since $d^k = o(n)$, we get by (6) that for the procedures $\delta_n^{d^k}$ one has

$$P_{\theta_n^k}(\text{rejecting } H_{\mathrm{I}}^{nk}) = P_{\theta_n^k}(U_3^2 < 2d^k) + o(1).$$

Here θ_n^k is determined by the sequence of penalties $C_n^k = d^k$ and by the associated sequence of procedures $\delta_n^{d^k}$, in the following way: $\theta_n^k = \theta_n(G_0)$, where $R_n(G_0, \delta_n^{d^k}) = \sup_{G \in \mathscr{G}} R_n(G, \delta_n^{d^k})$. We will show in the sequel, that for large enough B, $|\theta_n^k| < B\sqrt{d^k}/n$ for every n and k, or, equivalently, by (3) we obtain that the corresponding τ_{nk}^2 satisfies $\tau_{nk}^2 = (d^k + b_n^k\sqrt{d^k})/n$ for $b_n^k < B$. As previously discussed, U_3^2 has a noncentral chi-square distribution with random noncentrality parameter; since $d^k = o(n)$, the noncentrality parameter has a degenerate distribution (as $n \to \infty$), and it may be considered fixed, equal to $n\tau_{nk}^2$. Now the expectation of U_3^2, under $|\theta_n^k|$ and under $-|\theta_n^k|$, is $2d^k - b_n^k\sqrt{d^k}$ and $2d^k + b_n^k\sqrt{d^k}$, respectively. Also, the asymptotic variance of U_3^2 under both $|\theta_n^k|$ and $-|\theta_n^k|$ equals $6d^k$. Thus we obtain

$$P_{-|\theta_n^k|}(U_3^2 < 2d^k + c) = P_{|\theta_n^k|}(U_3^2 > 2d^k - c) + o(1).$$

The last equation follows by approximating the distribution of a noncentral chi-square random variable with large number of degrees of freedom d^k by a normal distribution. The approximation is by the CLT when representing U_3^2 as a sum of squares of d^k i.i.d. normal random variables, denoted ξ_i, with variance 1 and mean μ_k. This approximation is uniform, since the second and third moments of the i.i.d. terms in the representation of U_3^2 are bounded under θ_n^k. Hence, by the Berry–Esseen theorem the convergence to normality is uniform. By the symmetry of $l(\cdot)$, for large enough d^k and n,

$$\begin{aligned}
R_n\big(-|\theta_n^k|, \delta_n^{M,k}\big) &= l\big(-|\theta_n^k|\big)P\big(\text{rejecting } H_{\mathrm{I}}^{nk} \text{ by } \delta_n^{M,k}\big)\\
&\approx l\big(-|\theta_n^k|\big)P_{-|\theta_n^k|}\big(U_3^2 < 2C_n^{k,M}\big)\\
&> l\big(-|\theta_n^k|\big)P_{-|\theta_n^k|}\big(U_3^2 < (1+\varepsilon)2d^k\big)\\
&\approx l\big(|\theta_n^k|\big)P_{|\theta_n^k|}\big(U_3^2 > 2d^k - 2\varepsilon d^k\big)\\
&> l\big(|\theta_n^k|\big)P_{|\theta_n^k|}\big(U_3^2 > 2d^k\big) \approx R_n\big(G_0, \delta^{d^k}\big).
\end{aligned}$$

Note that since $\tau_{nk}^2 < (d^k + B\sqrt{d^k})/n$, $P_{|\theta_n^k|}(U_3^2 > 2d^k)$ is bounded away from zero, and thus the approximate equality sign is meaningful. Hence we got a contradiction to the minimaxity of $\delta_n^{M,k}$; this follows since $R_n(G_0, \delta_n^{d^k})$, the worst case risk for the procedure $\delta_n^{d^k}$, is smaller than the worst case risk obtained by $\delta_n^{M,k}$. Similarly we show that $\liminf_n(C_n^{k,M}/d^k) > 1-\varepsilon$ for every ε and large enough k. A closer look at the proof, when bounding the normal approximation error using the Berry–Esseen theorem, reveals that $C_n^{k,M} = d^k + o(\sqrt{d^k})$ for a proper choice of $n = n(k)$.

It remains to show that $\tau_{nk}^2 < (d^k + B\sqrt{d^k})/n$, for large enough B. First, we will show it assuming that $\tau_{nk}^2 = (d^k + o(d^k))/n$. Let $\bar{\tau}_{nk}^2 = (d^k + B\sqrt{d^k})/n$, and $\tilde{\tau}_{nk}^2 = (d^k + \sqrt{6d^k})/n$. We will prove that for large enough B, n and k,

$$\tilde{l}(\tilde{\tau}_{nk}^2)P_{\tilde{\tau}_{nk}^2}(U_3^2 < 2d^k) > \tilde{l}(\bar{\tau}_{nk}^2)P_{\bar{\tau}_{nk}^2}(U_3^2 < 2d^k).$$

Here $\tilde{l}(\tau^2) = l(\theta_n(\tau^2))$ is the function $l(\cdot)$ in terms of τ^2. Under the assumption that $\tilde{\tau}^2_{nk} = (d^k + o(d^k))/n$, the asymptotic variance of U_3^2 equals $6d^k$. Hence, by the CLT $P_{\tilde{\tau}^2_{nk}}(U_3^2 < 2d^k) \to \Phi(-1)$. From the last fact we obtain that $\inf_k P_{\tilde{\tau}^2_{nk}}(U_3^2 < 2d^k) = r > 0$. Thus $\tilde{l}(\tilde{\tau}^2)P_{\tilde{\tau}^2}(U_3^2 < 2d^k) > \tilde{l}(\tilde{\tau}^2)r$. The inequality (8) in Lemma 2 for the tail of U_3^2 yields for $\bar{\tau}^2_{nk} = (d^k + B\sqrt{d^k})/n$, $P_{\bar{\tau}^2_{nk}}(U_3^2 < 2d^k) < \exp(-\alpha B)$ with a proper α. Thus we have $\tilde{l}(\bar{\tau}^2)P_{\bar{\tau}^2}(U_3^2 < 2d^k) < \tilde{l}(\bar{\tau})\exp(-\alpha B)$. For polynomial loss the assertion now follows by taking large enough B. The fact that $\tau^2_{nk} = (d^k + o(d^k))/n$, assumed in the proof, may be easily obtained from the tail inequality for U_3^2 as in Lemma 1(i). □

REMARK 1. If we do not assume $\sigma_G^2 \le \sigma_0^2 < \infty$ in the last theorem, then $r_n^* \equiv \infty$ under any sequence of selection rules determined by a sequence C_n. Thus any such a procedure and, in particular, AIC procedures, is (trivially) asymptotically minimax. We may avoid this extra assumption and difficulty if we consider normalized squared prediction errors, for example, $(Y - \widehat{Y})^2/\sigma_Y^2$, and define θ_n accordingly.

The above results indicate that the proposed minimax criterion behaves similarly to the AIC criterion when the dimension's difference d is large. It may also be seen from the following numerical study. Table 1 displays the values of $\lim_n C_n$ that correspond to asymptotically minimax rules under a 0–1 loss and under an absolute value loss, for various values of models' dimension differences d. In the case of the 0–1 loss, the values in the table represent the median of the corresponding $\mathscr{X}_d^2(d)$ distribution divided by two. For the absolute value loss, the values in the table are obtained by numerically minimizing the function

$$R(c) = \max_\theta |\theta| \bigl[P\{\mathscr{X}_d^2(d+\theta) < 2c\} 1_{\{\theta<0\}} + \bigl(1 - P\{\mathscr{X}_d^2(d-\theta) \ge 2c\}\bigr) 1_{\{\theta \ge 0\}} \bigr]. \tag{9}$$

TABLE 1

The limiting differences of penalties corresponding to asymptotically minimax procedures

$\boldsymbol{d}$	$\boldsymbol{l(\theta) = 1}$	$\boldsymbol{l(\theta) = \lvert\theta\rvert}$	$\boldsymbol{d}$	$\boldsymbol{l(\theta) = 1}$	$\boldsymbol{l(\theta) = \lvert\theta\rvert}$
1	0.54	0.48	9	8.51	8.10
2	1.53	1.38	10	9.45	9.30
3	2.52	2.25	20	19.50	19.20
4	3.54	3.24	40	39.00	38.40
5	4.50	4.20	50	48.75	48.00
6	5.49	5.22	55	53.63	54.45
7	6.51	6.10	60	59.40	59.40
8	7.44	7.20	80	79.20	79.20

As the proof of Theorem 2 suggests, it is sufficient to maximize the right-hand side of (9) over $\theta \in [-d, d]$. Finally, in our implementation we minimize $R(c)$ over $c \in [0, 3d]$. Notice that for small values of d, $\lim_n C_n$ is significantly smaller than the corresponding values suggested by the AIC; this is contrary to the criticism suggesting that the values corresponding to the AIC are too small.

4. General class of distributions. In the main result of this section, Theorem 3, we give conditions on $\mathscr{G}$ and $l(\cdot)$ under which, for any asymptotically minimax sequence of procedures $\{\delta_n\}$, the corresponding sequence of differences $\{C_n(\mathrm{I}) - C_n(\mathrm{II})\}$ is bounded. Theorem 4 gives conditions for the existence of asymptotically minimax sequence of procedures with bounded penalty differences.

Let $V = (Y, X_1, \dots, X_m)$ be a random vector with distribution G, where $X_1, \dots, X_m$ are the explanatory variables, and Y is the corresponding response variable. Assume that Y and X_i are centered random variables, $E_G|Y|^2 < \infty$ and $E_G|X_i|^2 < \infty$, $\forall\, i = 1, \dots, m$. Without loss of generality we will assume that $X_1, \dots, X_m$ are linearly independent with respect to G, that is, $\sum_{i=1}^m a_i X_i = 0$, G-a.s. only if all a_i's are equal to zero. Let $\mathscr{L}(X_1, \dots, X_m)$ be the linear subspace spanned by $(X_1, \dots, X_m)$. It is well known that for a given G there exists an orthonormal system $(\eta_1^G, \dots, \eta_m^G)$ such that $\mathscr{L}(X_1, \dots, X_i) = \mathscr{L}(\eta_1^G, \dots, \eta_i^G)$ for every $i \le m$; then Y admits the standard orthogonal decomposition

$$Y = \sum_{i=1}^{m} \alpha_i \eta_i^G + \varepsilon_G, \tag{10}$$

where $E_G[\varepsilon_G] = 0$, $E_G|\eta_i^G|^2 = 1$, $E_G[\eta_i^G \varepsilon_G] = 0$, $i = 1, \dots, m$ and $E_G[\varepsilon_G^2] = \sigma_G^2$. Note that $E_G[\varepsilon_G | \eta_1, \dots, \eta_m]$ is not necessarily zero here. In what follows we will assume that $(X_1, \dots, X_m)$ is an orthonormal system with respect to the underlying distribution G, and in (10) we will write X_i instead of η_i^G. In particular, in all the assumptions to follow X_i, $i = 1, \dots, m$ should be understood as an orthonormal system. In this section we will consider classes of distributions $\mathscr{G}$ satisfying

$$\text{(A)} \qquad \sup_{G \in \mathscr{G}} E_G|X_i|^{16} < \infty, \quad i = 1, \dots, m, \qquad \sup_{G \in \mathscr{G}} E_G|\varepsilon_G|^{16} < \infty.$$

The following two competing normal regression models are assumed:

$$Y = \sum_{i=1}^{m} \beta_i X_i + \varepsilon, \tag{I}$$

$$Y = \sum_{i=1}^{k} \beta_i X_i + \varepsilon, \tag{II}$$

where ε is a normal zero mean random variable with unknown variance, and $d = m - k > 0$. The assumed normal models give rise to the following predictions:

$$\widehat{Y}^{\mathrm{I}}(n+1) = \hat\beta_{\mathrm{I}}'\phi_{\mathrm{I}}(n+1), \qquad \widehat{Y}^{\mathrm{II}}(n+1) = \hat\beta_{\mathrm{II}}'\,\phi_{\mathrm{II}}(n+1),$$

where $\phi_{\mathrm{I}}(t) = (X_1(t), \dots, X_m(t))$, $\phi_{\mathrm{II}}(t) = (X_1(t), \dots, X_k(t))$, and $\hat\beta_{\mathrm{I}}, \hat\beta_{\mathrm{II}}$ are the standard least squares estimates based on models I and II, respectively,

$$\hat\beta_j = \left(\sum_{t=1}^{n} \phi_j(t)\phi_j'(t)\right)^{-1}\left(\sum_{t=1}^{n} \phi_j(t)Y(t)\right), \qquad j = \mathrm{I}, \mathrm{II}. \tag{11}$$

We introduce some notation used in what follows. Let $\mathscr{M}$ be the set of all possible regression models based on the explanatory variables $X_1, \dots, X_m$. Let $\phi_j(t)$, $j \in \mathscr{M}$ be the subvector of $(X_1(t), \dots, X_m(t))$ corresponding to model j. Define also $\alpha_{\mathrm{I}} = (\alpha_1, \dots, \alpha_m)'$, $\alpha_{\mathrm{II}} = (\alpha_1, \dots, \alpha_k)'$ and $W_j(n) = \sum_{t=1}^{n} \phi_j(t)\phi_j'(t)$, $j \in \mathscr{M}$. In words, $W_j(n)$ is the "$X'X$" matrix corresponding to model j and based on n observations. We also write $W_{\mathrm{I}}(n)$ and $W_{\mathrm{II}}(n)$ for the matrices corresponding to models I and II, respectively. Our current goal is to evaluate the mean square prediction errors of $\widehat{Y}^j(n+1)$, $j =$I, II.

In order to derive an analog of Lemma 1 for the least squares estimates under a general class of distributions $\mathscr{G}$, we need an invertibility property of the matrices $W_j(n)$ as in the following condition (B):

(B) There exists an integer number N_0 such that

$$\sup_{n\ge N_0,\, G\in\mathscr{G}} E_G\big(\|[W_j(n)/n]^{-1}\|^8 \mathbf{1}\{\lambda_{\min}[W_j(n)] > 0\}\big) < \infty, \qquad j \in \mathscr{M}. \tag{12}$$

Here $\|\cdot\|$ stands for the standard Euclidean matrix norm, and $\lambda_{\min}[\cdot]$ is the minimal eigenvalue of a matrix.

Roughly, condition (B) assumes that the expected value of $\|[W_j(n)/n]^{-1}\|^8$ is bounded uniformly over the class of distributions $\mathscr{G}$ for all models $j \in \mathscr{M}$. The matrices $W_j(n)$, $j = \mathrm{I}, \mathrm{II}$ may be singular with positive probability for every n. If $W_j(n)$, is not of full rank then the corresponding estimate (11) is not unique. Note, however, that a predictor $\widehat{Y}^j(n+1)$ is well defined here; it is the projection of Y on a largest linear subspace spanned by the columns of the corresponding regression matrix. In this case the reasoning below should be applied to the models of smaller dimensions. That is why we require (12) for all submodels $j \in \mathscr{M}$.

One way to obtain condition (B) is by introducing a "no-inference zone." When the norm of the matrix $W_j(n)^{-1}$ is too large [$W_j(n)$ is close to singularity], there is no inference. An elaborate and technical study of condition (B) may be found in Goldenshluger and Greenshtein (1998).

LEMMA 3. *Let $\mathscr{G}$ satisfy conditions* (A) *and* (B); *then there exists N_0 such that*

$$\sup_{n\ge N_0,\, G\in\mathscr{G}} E_G\big(n^2\|\hat\beta_j - \alpha_j\|^4 \mathbf{1}\{\lambda_{\min}[W_j(n)] > 0\}\big) < \infty, \qquad j = \mathrm{I}, \mathrm{II}.$$

PROOF. We have

$$\hat{\beta}_{\mathrm{I}} - \alpha_{\mathrm{I}} = (n^{-1} W_{\mathrm{I}}(n))^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} \phi_{\mathrm{I}}(t) \varepsilon_G(t) \right), \tag{13}$$

$$E_G \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \phi_{\mathrm{I}}(t) \varepsilon_G(t) \right\|^8 = E_G \left[\sum_{i=1}^{m} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_i(t) \varepsilon_G(t) \right)^2 \right]^4. \tag{14}$$

Note that $\{X_i(t)\varepsilon_G(t)\}_{t=1,\dots,n}$ is a sequence of independent zero mean random vectors. It can be checked by direct calculation that under the moment conditions (A),

$$\sup_{n, G \in \mathscr{G}} E_G \left(n^{-1/2} \sum_{t=1}^{n} X_i(t) \varepsilon_G(t) \right)^8 < \infty, \qquad i = 1, \dots, m.$$

Therefore it follows from (14) that

$$\sup_{n, G \in \mathscr{G}} E_G \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \phi_{\mathrm{I}}(t) \varepsilon_G(t) \right\|^8 < \infty,$$

and using (13) and condition (B) we complete the proof. □

The next lemma establishes expressions for the mean square prediction errors of $\widehat{Y}^j(n+1)$, $j = \mathrm{I,II}$ defined by (2). For simplicity in proofs in the sequel we will consider cases where the probability of singularity of $W_{\mathrm{I}}(n)$ and $W_{\mathrm{II}}(n)$, $n \geq m$ is zero. To treat the singularity, one should project on subspaces of lower dimension.

LEMMA 4. *Let $\mathscr{G}$ be a class of distributions satisfying conditions* (A) *and* (B); *then*

(i)

$$\theta_n^{\mathrm{I}} = \sigma_G^2 + \frac{q_{\mathrm{I}}}{n} + o_G(n^{-1}), \qquad n \to \infty, \tag{15}$$

$$\theta_n^{\mathrm{II}} = \sigma_G^2 + \tau^2 + \frac{q_{\mathrm{II}}}{n} + o_G(n^{-1}), \qquad n \to \infty, \tag{16}$$

where $\tau^2 = \tau^2(G) = \sum_{i=k+1}^{m} \alpha_i^2$, and

$$q_{\mathrm{I}} = q_{\mathrm{I}}(G) = \sum_{i=1}^{m} E_G[X_i^2 \varepsilon_G^2],$$

$$q_{\mathrm{II}} = q_{\mathrm{II}}(G) = \sum_{i=1}^{k} E_G\Big[X_i^2 \Big(\varepsilon_G + \sum_{i=k+1}^{m} \alpha_l X_l\Big)^2\Big].$$

(ii) *There exists a constant $\tilde{\kappa}$ such that $|o_G(n^{-1})| \leq \tilde{\kappa} n^{-1}$, $\forall G \in \mathscr{G}$.*

PROOF. (i) The mean square prediction error of $\widehat{Y}^{\mathrm{I}}(n+1)$ is

$$\begin{aligned}\theta_n^{\mathrm{I}} &= E_G\big[(\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}})'\phi_{\mathrm{I}}(n+1)-\varepsilon_G(n+1)\big]^2\\ &= \sigma_G^2 - 2E_G\big[(\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}})'\phi_{\mathrm{I}}(n+1)\varepsilon_G(n+1)\big] + E_G\|\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}}\|^2\\ &= \sigma_G^2 + E_G\|\hat{\beta}_{\mathrm{I}}-\alpha_I\|^2.\end{aligned}$$

The third equality follows from conditioning on the σ-algebra $\mathscr{F}_n$ generated by $Y(t), X_{\mathrm{I}}(t), \dots, X_m(t)$, $t=1,\dots,n$, and from the fact that $\phi_{\mathrm{I}}(n+1)$, and $\varepsilon_G(n+1)$ are uncorrelated and independent of $\mathscr{F}_n$. Further, we have

$$\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}} = \left(\frac{1}{n}\sum_{t=1}^{n}\phi_{\mathrm{I}}(t)\phi_{\mathrm{I}}'(t)\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}\phi_{\mathrm{I}}(t)\varepsilon_G(t)\right).$$

By the law of large numbers for every $G \in \mathscr{G}$, $n^{-1}\sum_{t=1}^{n}\phi_{\mathrm{I}}(t)\phi_{\mathrm{I}}'(t) \to^p I_m$ as $n\to\infty$, where I_m is the $m\times m$ identity matrix. Observe that $\{\xi_{\mathrm{I}}(t)\} = \{\phi_{\mathrm{I}}(t)\varepsilon_G(t)\}$, $t=1,\dots,n$ is a sequence of independent zero mean m-vectors all having the same distribution. In addition, $E_G\|\xi_{\mathrm{I}}(t)\|^2 < \infty$ for every $G\in\mathscr{G}$, so by the multidimensional central limit theorem we have that $n^{-1/2}\sum_{t=1}^{n}\xi_{\mathrm{I}}(t) \to^d \mathscr{N}_m(0, Q_{\mathrm{I}})$, where $Q_{\mathrm{I}} = Q_{\mathrm{I}}(G) = E_G[\phi_{\mathrm{I}}\phi_{\mathrm{I}}'\varepsilon_G^2]$. Hence $\sqrt{n}(\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}}) \to^d \mathscr{N}_m(0, Q_{\mathrm{I}})$, and

$$n\|\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}}\|^2 \overset{d}{\to} Z'Z, \qquad Z\sim\mathscr{N}_m(0,Q_{\mathrm{I}}). \tag{17}$$

It follows from Lemma 3 that $\{n\|\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}}\|^2\}$ is a sequence of uniformly integrable random variables. This fact along with (17) implies convergence of expectations; that is, $nE_G\|\hat{\beta}_{\mathrm{I}}-\alpha_{\mathrm{I}}\|^2 \to \mathrm{tr}(Q_{\mathrm{I}})$ as $n\to\infty$. Hence, (15) follows if we note that $q_{\mathrm{I}} = \mathrm{tr}(Q_{\mathrm{I}})$. The equality (16) follows from the same considerations for the model II.

Statement (ii) of the lemma follows immediately from Lemma 3. □

From the lemma we obtain

$$\theta_n = -\tau^2 + \frac{q_{\mathrm{I}}-q_{\mathrm{II}}}{n} + o_G(n^{-1}), \tag{18}$$

where $|o_G(n^{-1})| \le \tilde{\kappa}n^{-1}$ for some $\tilde{\kappa}$ and every $G\in\mathscr{G}$. When X_i, $i=1,\dots,m$ are independent rather than uncorrelated, and X_i's are independent of ε_G, this equality takes the form [cf. (3)],

$$\theta_n = -\tau^2\Big(1+\frac{k}{n}\Big) + \frac{d\sigma_G^2}{n} + o_G(n^{-1}).$$

Denote

$$U_3^2 = U_1^2 - U_2^2 = \sum_{t=1}^{n}(Y(t)-\hat{\beta}_{\mathrm{II}}'\phi_{\mathrm{II}}(t))^2 - \sum_{t=1}^{n}(Y(t)-\hat{\beta}_{\mathrm{I}}'\phi_{\mathrm{I}}(t))^2,$$

as in (5). The dependence of these variables on n is suppressed. Let

$$U = U_3^2 - 2C_n\Big(\frac{U_2^2}{n}-\sigma_G^2\Big) - U_2^2\Big(\exp(2C_nn^{-1}) - 1 - 2C_nn^{-1}\Big).$$

An argument similar to (6) shows that for every n,

$$P_G\{\delta_n \text{chooses model I}\} = P_G\{U \geq 2\sigma_G^2 C_n\}.$$

In the next lemma we establish some useful properties of the random variable U.

LEMMA 5. *Suppose $\mathscr{G}$ satisfies conditions* (A), (B), *and let $\mathscr{G}_n(K)$ denote the set of all distributions from $\mathscr{G}$ such that $\tau^2(G) = \tau^2 \leq Kn^{-1}$ for a given K. The random variable U admits the following representation $U = \zeta_n + \eta_n$ with the following properties*:

(i) $\sup_{n\geq N_0, G\in\mathscr{G}_n(K)} E_G|\zeta_n| < \infty$, *for some N_0 and every K.*
(ii) $\sup_{G\in\mathscr{G}} E_G|\eta_n| = o(C_n)$, $n \to \infty$.

PROOF. Denote $\nu_G = \sum_{i=k+1}^m \alpha_i X_i$; then we have

$$\begin{aligned} U_3^2 = &\Bigg[(\alpha_{\mathrm{II}} - \hat{\beta}_{\mathrm{II}})'\Bigg(\sum_{t=1}^n \phi_{\mathrm{II}}(t)\phi_{\mathrm{II}}'(t)\Bigg)(\alpha_{\mathrm{II}} - \hat{\beta}_{\mathrm{II}}) \\ &\qquad -(\alpha_{\mathrm{I}} - \hat{\beta}_{\mathrm{I}})'\Bigg(\sum_{t=1}^n \phi_{\mathrm{I}}(t)\phi_{\mathrm{I}}'(t)\Bigg)(\alpha_{\mathrm{I}} - \hat{\beta}_{\mathrm{I}})\Bigg] \\ &+2\Bigg[(\alpha_{\mathrm{II}} - \hat{\beta}_{\mathrm{II}})'\sum_{t=1}^n \phi_{\mathrm{II}}(t)(\varepsilon_G(t) + \nu_G(t)) - (\alpha_{\mathrm{I}} - \hat{\beta}_{\mathrm{I}})'\sum_{t=1}^n \phi_{\mathrm{I}}(t)\varepsilon_G(t)\Bigg] \\ &+\sum_{t=1}^n\Big[(\varepsilon_G(t) + \nu_G(t))^2 - \varepsilon_G^2(t)\Big] \\ = &\,\mathscr{T}_1(n) + \mathscr{T}_2(n) + \mathscr{T}_3(n). \end{aligned}$$

Using the same ideas as in the proof of Lemma 3, one can show that there exists N_0 such that $\sup_{n\geq N_0 G\in\mathscr{G}}(E_G|\mathscr{T}_i(n)|) < \infty, i = 1, 2$. In addition, $E_G[\mathscr{T}_3(n)] = n\tau^2$, and therefore $\sup_{n\geq N_0 G\in\mathscr{G}_n(K)} E_G[U_3^2] < \infty$. Now we define $\eta_n = 2C_n(U_2^2 n^{-1} - \sigma_G^2)$ and

$$\zeta_n = U_3^2 = U_2^2\Big(\exp(2C_n n^{-1}) - 1 - 2C_n n^{-1}\Big).$$

The statement of the lemma follows from the fact that $\sup_{G\in\mathscr{G}} E_G|U_2^2 n^{-1} - \sigma_G^2| = o(1)$ as $n \to \infty$. □

Our current goal is to characterize asymptotically minimax selection procedures for a general class $\mathscr{G}$ of possible joint distributions of the explanatory and response variables. Recall that the risk of a sequence of procedures δ_n is defined by

$$r_n(\delta_n) = \sup_{G\in\mathscr{G}} E_{G^n} L_n(\theta_n, \delta_n) = \sup_{G\in\mathscr{G}}\Big[l(\theta_n) P_G\{\delta_n \text{ makes a wrong decision}\}\Big].$$

Let $\mathscr{G}_n^{\mathrm{I}} = \{G \in \mathscr{G}\colon \theta_n(G) > 0\}$; $\mathscr{G}_n^{\mathrm{II}} = \{G \in \mathscr{G}\colon \theta_n(G) > 0\}$. Let δ_n be a sequence of selection rules associated with the sequence of differences of dimension penalties C_n. Define

$$R_n^{\mathrm{I}}(\delta_n) = \sup_{G\in\mathscr{G}_n^{\mathrm{I}}}\Big[l(\theta_n(G))P_G\Big\{U < 2\sigma_G^2 C_n\Big\}\Big],$$
$$R_n^{\mathrm{II}}(\delta_n) = \sup_{G\in\mathscr{G}_n^{\mathrm{II}}}\Big[l(\theta_n(G))P_G\Big\{U \geq 2\sigma_G^2 C_n\Big\}\Big].$$

Due to Lemma 4 and (18) there exist some positive constants K_1, K_2 and N_0 such that for all $n \geq N_0$,

$$\tau^2(G) \geq K_1/n \quad \text{implies} \quad G \in \mathscr{G}_n^{\mathrm{I}}, \tag{19}$$

$$G \in \mathscr{G}_n^{\mathrm{II}} \quad \text{implies} \quad \theta_n(G) = |\theta_n(G)| \leq K_2/n. \tag{20}$$

Asymptotically minimax selection procedures are determined essentially by properties of the sets $\mathscr{G}_n^{\mathrm{I}}$ and $\mathscr{G}_n^{\mathrm{II}}$. The next theorem characterizes asymptotically minimax selection rules for general classes of joint distributions $\mathscr{G}$.

THEOREM 3. *Suppose $\mathscr{G}$ satisfies conditions* (A) *and* (B), *and for every $c > 0$ the set $\mathscr{G}$ contains at least one distribution G such that $\tau^2(G) \in (0, c)$. If $l(k_1/n)(l(k_2/n))^{-1} \not\to 1$ as $n \to \infty$ for all $k_1 \neq k_2$, then for every sequence of asymptotically minimax selection rules $\delta_n \in \Delta_n^c$, the difference in the dimension penalties $C_n = C_n(\mathrm{I}) - C_n(\mathrm{II})$ is bounded, that is,*

$$\limsup_{n\to\infty} C_n < \infty. \tag{21}$$

PROOF. In the proof w.l.o.g we set $\sigma_G^2 = 1$. Suppose there exists a sequence $\bar{\delta}_n$ of asymptotically minimax rules with penalties, $\overline{C}_n$, such that $\lim_{n\to\infty} \overline{C}_n = \infty$. By definition and Lemma 5,

$$R_n^{\mathrm{I}}(\bar{\delta}_n) = \sup_{\{G\colon\, \theta_n(G)\leq 0\}} l(\theta_n(G))P_G\{\zeta_n + \eta_n < 2\overline{C}_n\},$$
$$R_n^{\mathrm{II}}(\bar{\delta}_n) = \sup_{\{G\colon\, \theta_n(G)> 0\}} l(\theta_n(G))P_G\{\zeta_n + \eta_n \geq 2\overline{C}_n\}.$$

It follows from the premise of the theorem that $\mathscr{G}$ contains a sequence of distributions $\{G_j\}$ with monotone decreasing $\tau^2(G_j) = \tau_j^2 \to 0$, $j \to \infty$. Fix K_1 and K_2 satisfying (19) and (20), and let K_1 be large enough so that $\tau^2(G) > K_1/n$ implies $|\theta_n(G)| > 2K_2/n$. Define the subsequence $n_j = [K_1/\tau_j^2]+1$; here $[\cdot]$ denotes the integer part. Note that $n_j \to \infty$ as $j \to \infty$. For the subsequence of time instants $\{n_j\}$, the sets $\mathscr{G}_{n_j}^{\mathrm{I}}$ contain at least one distribution, G_j, such that $|\theta_n(G_j)| > 2K_2 n_j^{-1}$. Now $R_n^{\mathrm{I}}(\bar{\delta}_{n_j})$ may be bounded from below by

$$\begin{aligned} R_{n_j}^{\mathrm{I}}(\bar{\delta}_{n_j}) &\geq l(\theta_n(G_j))P_{G_j}\{\zeta_n + \eta_n < 2\overline{C}_n\} \\ &\geq l(2K_2 n_j^{-1})(1 - o(1)), \qquad n_j \to \infty. \end{aligned}$$

The last inequality is by Markov inequality combined with Lemma 5; notice that $\tau^2(G_j) < Kn_j^{-1}$ for a proper K. Thus, we obtain that

$$r_{nj}(\bar{\delta}_{n_j}) \geq l(2K_2 n_j^{-1})(1 - o(1)), \qquad n_j \to \infty. \tag{22}$$

Now let $\tilde{\delta}_n$ be the sequence of selection rules which always chooses model I. This sequence of rules corresponds to the choice $C_n \equiv 0$. The risk of such a procedure will be

$$r_n(\tilde{\delta}_n) = R_n^{\mathrm{II}}(\tilde{\delta}_n) \leq l(K_2 n^{-1}), \qquad n \to \infty. \tag{23}$$

Comparing (23) with (22) and taking into account the properties of $l(\cdot)$, we obtain $\limsup_{n\to\infty} r_n(\tilde{\delta}_n)[r_n(\bar{\delta}_n)]^{-1} < 1$, which contradicts the asymptotic minimax property of $\bar{\delta}_n$; thus (21) follows. □

The condition imposed on $\mathscr{G}$ in the above theorem is essential for boundedness of C_n. It ensures that the sets $\mathscr{G}_n^{\mathrm{I}}$ and $\mathscr{G}_n^{\mathrm{II}}$ are not empty eventually, and $\mathscr{G}_n^{\mathrm{II}}$ contains distributions with τ^2 arbitrarily close to zero. Note that if $\mathscr{G}_n^{\mathrm{II}}$ is empty starting from some n, then the sequence of procedures δ_n which always selects model I ($C_n \equiv 0$) is asymptotically minimax, and the theorem holds trivially. On the other hand, if $\mathscr{G}$ contains only distributions with $\tau^2(G) = 0$ then the sequence of rules which identically choose model II ($C_n \equiv \infty$) is asymptotically minimax, and for such $\mathscr{G}$ the theorem does not hold. It should be stressed that the condition that the sets $\mathscr{G}_n^{\mathrm{I}}$, $\mathscr{G}_n^{\mathrm{II}}$ are not empty eventually is not sufficient for boundedness of C_n. Indeed, suppose that $\mathscr{G}$ contains only two distributions with $\tau^2(G_1) = 0$ and with $\tau^2(G_2) = c$ for some constant $c > 0$. Here $\mathscr{G}_n^{\mathrm{I}}$ and $\mathscr{G}_n^{\mathrm{II}}$ are not empty for large n, but $\mathscr{G}_n^{\mathrm{I}}$ is not rich enough and contains only a distribution with "large" $\tau^2(G)$. Different choices of such distributions may lead to asymptotically minimax procedures with bounded as well as unbounded C_n; the boundedness is determined by the tail behavior of the corresponding distributions of $\zeta_n + \eta_n$.

Theorem 3 asserts that if $\mathscr{G}$ contains distributions with $\tau^2(G)$ arbitrarily close to zero, then *every* asymptotically minimax rule corresponds to a bounded sequence of penalty differences C_n. It turns out that under more relaxed assumptions on $\mathscr{G}$ it can be shown that there *exists* an asymptotically minimax rule with bounded C_n (see Theorem 4 below).

Consider loss functions $l(\cdot)$ satisfying

$$l(k_1/n) = O(l(k_2/n)) \qquad \forall\, k_1, k_2, \text{ as } n \to \infty. \tag{24}$$

Our main interest is again in the cases $l(\theta) = |\theta|$ and $l(\theta) = 1$.

DEFINITION 2. Given a loss $l(\cdot)$ satisfying (24) and a class $\mathscr{G}$ satisfying conditions (A) and (B), we say that $\mathscr{G}$ and $l(\cdot)$ determine a *difficult* selection problem if the induced minimax sequence of values r_n^* satisfies $r_n^* = O(l(n^{-1}))$, $n \to \infty$.

Observe that for any $\mathscr{G}$ satisfying conditions (A) and (B), and $l(\cdot)$ satisfying (24), the minimax sequence of risks is of magnitude less than or equal to $O(l(n^{-1}))$ (compare with the sequence of rules which identically chooses model (I); thus when equality holds, the problem is indeed difficult. Notice further that for any given $l(\cdot)$, the set of all classes $\mathscr{G}$ that determine difficult problems is larger than the set of all classes $\mathscr{G}$ satisfying the conditions of Theorem 3.

THEOREM 4. *Suppose $\mathscr{G}$ and $l(\cdot)$ determine a difficult problem. Then there exists a sequence of minimax procedures with a bounded difference of penalties C_n.*

PROOF. Obviously r_n^* may be achieved by a sequence of equalizer rules $\delta_n \in \Delta_n^c$ satisfying $R_n^{\mathrm{I}}(\delta_n) \approx R_n^{\mathrm{II}}(\delta_n)$, $\forall\, n$. Now suppose that such a rule is associated with $C_n \to \infty$. In this case an argument similar to that of Theorem 3 implies that $R_n^{\mathrm{II}}(\delta_n) \le o(1)l(n^{-1})$. Since $\delta_n$is an equalizer, $R_n^{\mathrm{I}}(\delta_n)$ also satisfies the same inequality, and this is in contradiction with the assumption that the pair $\mathscr{G}$ and $l(\cdot)$ determines a difficult selection problem. The theorem is proved. □

5. Discussion. We will discuss now our results in light of the results obtained by Schwarz (1978). There is a fundamental difference since under the Schwarz approach $C_n \to \infty$, while under our approach C_n are usually bounded. A key difference is that in the development of Schwarz the prior is fixed throughout the asymptotics, while we, by taking minimax procedures for every n, consider implicitly statistical problems with an increasing difficulty which is scaled with n. Our approach, considering more difficult problems as n increases, is common in asymptotic theory.

The interplay between the likelihood function in normal models and the squared error prediction loss is important for establishing our results. Similar types of results for general prediction error losses and general competing models (which induce competing collections of predictors), would require adjustments of the selection methods. A possible approach is a criterion that is based on the performance of the "empirically best" predictor in each of the competing collections of predictors, together with a dimension penalty. See further development in this direction in Greenshtein (2000).

Finally we will briefly indicate how to extend the results to the case of more then two competing models. In order to simplify notations we consider only the generalization under $l(\theta) = |\theta|$. Let $\{\Omega_j\}$ be a set of nested competing models, which induce the parameters θ_n^j, $j = 1, \dots, J$ for each $G \in \mathscr{G}$. Denote $\theta_n^o = \min_j \theta_n^j$. Define the loss function

$$L_n((\theta_n^1, \dots, \theta_n^J), i) = \theta_n^i - \theta_n^0$$

for choosing model i. The extension of Δ_n^c to more than two models, by introducing a vector of dimension penalties, is obvious. We define

$$R_n(G, \delta_n) = R_n((\theta_n^1, \dots, \theta_n^J), \delta_n) = E_{G^n} L_n((\theta_n^1, \dots, \theta_n^J), \delta_n).$$

Our results are readily generalized to this setting.

Acknowledgments. The second author is grateful to Yosi Rinott for many helpful and stimulating discussions. Thanks also to Paul Feigin for helpful discussions.

REFERENCES

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

BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136.

FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975.

GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

GOLDENSHLUGER, A. and GREENSHTEIN, E. (1998). Asymptotic minimax procedures in regression model selection and the magnitude of the dimension penallty. Technical report, Technion–Israel Institute of Technology.

GREENSHTEIN, E. (2000). Predictor-selection. Another look on model-selection and estimation. Technical report, Technion–Israel Institute of Technology.

HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.

LINHART, H. and ZUCHINI, W. (1986). *Model selection*. Wiley, New York.

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

NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12** 758–765.

OLIKER, V. I. (1978). On the relationship between the sample size and the number of variables in a linear regression model. *Comm. Statist. A* **7** 509–516.

RISSANEN, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Books, Singapore.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

SHIBATA, R. (1986). Selection of the number of regression variables: a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.* **38** 459–474.

SHIBATA, R. (1989). Statistical aspects of model selection. *From Data to Model* (J. C. Willems, ed.) 215–240. Springer, New York.

SPEED, T. and YU, B. (1993). Model selection and prediction: normal regression. *Ann. Inst. Statist. Math.* **45** 35–54.

STONE, C. J. (1981). Admissible selection of an accurate and parsimonious normal linear regression model. *Ann. Statist.* **9** 475–485.

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

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Royal Statist. Soc. Ser. B* **39** 44–47.

THOMPSON, M. L. (1978). Selection of variables in multiple regression. *Internat. Statist. Rev.* **46** 1–49 and 129–146.

DEPARTMENT OF STATISTICS
UNIVERSITY OF HAIFA
HAIFA 31905
ISRAEL
E-mail: goldensh@rstat.haifa.ac.il

FACULTY OF INDUSTRIAL ENGINEERING
AND MANAGEMENT
TECHNION–ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA 32000
ISRAEL
E-mail: eitang@ie.technion.ac.il