Ann. Inst. Statist. Math.
Vol. 44, No. 2, 281–297 (1992)

SOME CONTRIBUTIONS TO SELECTION AND ESTIMATION IN THE NORMAL LINEAR MODEL*

J. H. VENTER AND S. J. STEEL

Department of Statistics and Operations Research, Potchefstroom University, Potchefstroom 2520, South Africa

(Received August 3, 1990; revised June 13, 1991)

Abstract. We study the choice of the quantity α in the FPE_α criterion for selecting a member of a class of normal linear models having an orthogonal structure. Two approaches are discussed, namely fixing the maximal estimation risk at a prescribed level and using minimax regret. Estimation of the risk actually achieved and an illustrative example are also discussed.

Key words and phrases: Model selection, linear model, estimation, FPE_α criterion, minimax regret, risk estimation.

1. Introduction

Consider the co-ordinate free version (Arnold ((1981), p. 55)) of the standard linear model, where we observe a random n-dimensional column vector $\boldsymbol{Y} = [Y_1 Y_2 \cdots Y_n]'$ (prime indicates transpose), which is supposed to have the form

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{e} \tag{1.1}$$

where the n-vector $\boldsymbol{e}$ is $N(\boldsymbol{0}, \sigma^2 I)$-distributed with I the $n \times n$ identity matrix, and where $\boldsymbol{\mu}$ is an unknown parameter vector which is assumed to belong to a **known** linear subspace M of $\mathbb{R}^n$. An extension of this model which incorporates the problem of **model selection** is obtained if in addition we assume that a family $\mathcal{L}$ of linear subspaces of M is given and that $\boldsymbol{\mu}$ may actually belong to or be close to some unknown $L \in \mathcal{L}$. Within this wider framework all aspects of inference may be of interest, but in this paper we restrict attention to estimation of $\boldsymbol{\mu}$. A common practice is to select a subspace L data-dependently and to estimate $\boldsymbol{\mu}$ accordingly and our purpose is to investigate the consequences of such a procedure.

The following notation is required. For any vector $\boldsymbol{x} \in \mathbb{R}^n$, the projection of $\boldsymbol{x}$ onto a subspace L will be denoted by $P_L \boldsymbol{x}$, while the projection of $\boldsymbol{x}$ onto $L^\perp$, the orthogonal complement of L, will be denoted by $Q_L \boldsymbol{x}$. We will write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}'\boldsymbol{y}$ for the usual inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$. The corresponding Euclidean norm of

* This research was supported by the FRD of South Africa.

$\boldsymbol{x}$ is $\|\boldsymbol{x}\| = (\boldsymbol{x}'\boldsymbol{x})^{1/2}$. Results in the book by Arnold (1981) will be used freely. In particular (Arnold ((1981), p. 73)), $P_M \boldsymbol{Y}$ is the component-wise minimum variance unbiased estimator of $\boldsymbol{\mu}$. Among its other desirable properties is that it is minimax with respect to total squared error loss.

The **multiple linear regression** model provides a first motivating example. It assumes that $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{e}$, with $\boldsymbol{e}$ as before and with $X = [\boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_m]$ a known $n \times m$ matrix with columns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$. The i-th row of X contains the values of the m explanatory variables (regressors) when the i-th value of the response, Y_i, is observed, $i = 1, \ldots, n$. We assume that X is of full rank m. Then $\boldsymbol{\mu} = X\boldsymbol{\beta}$ and M is the linear space spanned by the columns of X. **Variable selection** (sometimes also referred to as **subset selection**) is widely discussed in the literature (see e.g. Linhart and Zucchini (1986)). It entails selecting a subset of the regressors rather than using the full set. For example, using regressors $j_1, j_2, \ldots, j_s$ $(s \leq m)$ will be appropriate if one believes $\boldsymbol{\mu}$ to be in the linear subspace spanned by $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, \ldots, \boldsymbol{x}_{j_s}$. If the choice of regressors is completely open, we may introduce the family $\mathcal{L}$ whose members are the 2^m possible linear subspaces spanned by subsets of the columns of X. Often the first column of X consists of 1's to provide for a constant term in the model, and inclusion of this column is required; then $\mathcal{L}$ may be taken as the family with members the 2^{m-1} possible linear subspaces spanned by $\boldsymbol{x}_1$ and some subset of $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$. In either case a common procedure is to select some member, $\hat{L}$ say, of $\mathcal{L}$ data-dependently and to use the corresponding least squares estimator $P_{\hat{L}} \boldsymbol{Y}$ to estimate $\boldsymbol{\mu}$. An important aspect that we will investigate is the risk w.r.t. total squared error loss of such an estimator. Note for future reference that if L_j denotes the 1-dimensional subspace spanned by the j-th column of X, then a typical $L \in \mathcal{L}$ can be expressed as the direct sum, $L = L_{j_1} + L_{j_2} + \cdots + L_{j_s}$, where $j_1, j_2, \ldots, j_s$ index the columns of X that span L. If a constant term is required, we take $j_1 = 1$.

As a second example, consider the **balanced two-factor ANOVA** model where we observe independent random variables Y_{ijk}, where Y_{ijk} is $N(\mu_{ij}, \sigma^2)$-distributed, $i = 1, \ldots, I$; $j = 1, \ldots, J$; $k = 1, \ldots, K$. The Y_{ijk}'s are observations on the response where we study a row factor on I levels and a column factor on J levels, with K observations per cell. We may arrange the Y_{ijk}'s in a column vector $\boldsymbol{Y}$ and the corresponding arrangement of the μ_{ij}'s constitutes $\boldsymbol{\mu}$. Then M is the IJ-dimensional linear space consisting of vectors whose components form IJ blocks of size K each with identical components within blocks. It is well known that $M = L_0 + L_1 + L_2 + L_3$, where L_0 is spanned by the vector with all components 1, L_1 and L_2 are subspaces corresponding to main row and column effects respectively and L_3 to interaction effects. Also L_0, L_1, L_2 and L_3 are mutually orthogonal with $\dim(L_0) = 1$, $\dim(L_1) = I - 1$, $\dim(L_2) = J - 1$ and $\dim(L_3) = (I - 1)(J - 1)$. The family $\mathcal{L}$ can now be taken as $\{L_0, L_0 + L_1, L_0 + L_2, L_0 + L_3, L_0 + L_1 + L_2, L_0 + L_1 + L_3, L_0 + L_2 + L_3, M\}$. Selecting a member from this family entails a decision as to the inclusion of the row, the column and the interaction effects into the model, thereby implying corresponding least squares estimators for the μ_{ij}'s. Once more it is important to investigate the effect of the selection step on the overall risk of the resulting estimator of $\boldsymbol{\mu}$.

Returning to the general set-up, we shall denote $\dim(M)$ by m. Although the

case $m = n$ is of practical interest, it is more difficult and will not be treated here and we assume $m < n$. For this case the standard estimator

$$\hat{\sigma}^2 = \|Q_M \boldsymbol{Y}\|^2/(n-m) \tag{1.2}$$

of σ^2 is available. We use it throughout this paper and concentrate on the problem of estimating $\boldsymbol{\mu}$. It will often be the case that M can be written as the direct sum

$$M = L_0 + L_1 + \cdots + L_k = L_0 + \sum_{j=1}^{k} L_j \tag{1.3}$$

and $\mathcal{L}$ will be a sub-family of the 2^k subspaces of the form $L = L_0 + L_{j_1} + \cdots + L_{j_s} = L_0 + \sum_{j \in J} L_j$, with $J = \{j_1, \ldots, j_s\}$ a subset of $\{1, 2, \ldots, k\}$.

In Section 2 we derive the C_p selection criterion of Mallows (1973) for our general formulation using the co-ordinate free approach. We also show that if the unbiased estimator in this derivation is replaced by a suitable Bayes estimator then the generalized FPE_α criterion of Bhansali and Downham (1977) results. The choice of α in this criterion has been discussed by a number of authors (see e.g. Shibata (1986) and references herein) from various points of view. In Section 3 we consider the special case where M and $\mathcal{L}$ have the structure (1.3) and the L_i's are mutually orthogonal. For this case it is possible to obtain explicit expressions for the risk with respect to total squared error loss of the corresponding estimator $P_{\hat{L}} \boldsymbol{Y}$ and this enables us to find its maximum and minimum risks. We then show how to choose α (or more generally, a number of α's) such that the maximal risk is at a prescribed level. A table is provided to implement this procedure in practice. In Section 4 we discuss the choice of α's by the minimax regret approach suggested by Shibata (1986) in a related problem. A practical example that illustrates application of the various procedures is also discussed. In Section 5 we introduce an estimator for the risk of $P_{\hat{L}} \boldsymbol{Y}$, and we illustrate the use of this estimator within the context of an example. We close with a discussion and some open questions in Section 6.

2. The FPE_α selection criterion

Although the C_p and FPE_α criteria are well known a ready reference to their derivation within the general formulation given here using the co-ordinate free point of view, does not seem available. For completeness we give a brief outline of such a derivation. If we use $P_L \boldsymbol{Y}$ rather than $P_M \boldsymbol{Y}$ as an estimator of $\boldsymbol{\mu}$ (thinking that $\boldsymbol{\mu} \in L \in \mathcal{L}$), the risk w.r.t. total squared error loss is given by

$$E\|P_L \boldsymbol{Y} - \boldsymbol{\mu}\|^2 = \dim(L)\sigma^2 + \|P_{M|L}\boldsymbol{\mu}\|^2 = \dim(L)\sigma^2 + \|Q_L \boldsymbol{\mu}\|^2 \tag{2.1}$$

since $\boldsymbol{\mu} \in M$. Put

$$U = \|Q_L \boldsymbol{Y}\|^2, \quad \lambda^2 = \|Q_L \boldsymbol{\mu}\|^2, \quad q = \dim(L^\perp) = n - \dim(L). \tag{2.2}$$

Then U is $\sigma^2\chi_q^2(\lambda^2/\sigma^2)$-distributed, and (2.1) is estimated unbiasedly by

$$\dim(L)\hat{\sigma}^2 + U - q\hat{\sigma}^2 = \|Q_L \boldsymbol{Y}\|^2 + 2\dim(L)\hat{\sigma}^2 - n\hat{\sigma}^2. \tag{2.3}$$

If we ignore the term that does not depend on L, (2.3) becomes the C_p or FPE_2 criterion, which is the special case $\alpha = 2$ of

$$\mathrm{FPE}_\alpha = \|Q_L \boldsymbol{Y}\|^2 + \alpha\dim(L)\hat{\sigma}^2 \tag{2.4}$$

and L is chosen to minimize this criterion. A motivation of (2.4) can be based on the following considerations. If we suppose for the moment that σ^2 is known, an unbiased estimator of λ^2 is given by $U - q\sigma^2$, and this was used to obtain (2.3) from (2.2). In studying the problem of estimating λ^2 from U, Saxena and Alam (1982) point out that the unbiased estimator is unsatisfactory. It can e.g. be negative while $\lambda^2 \geq 0$. They introduce a family of Bayes estimators of λ^2 of the form

$$\frac{q\sigma^2}{1+c} + \frac{U}{(1+c)^2} \tag{2.5}$$

based on a gamma-type prior for λ^2, with the constant c a parameter of the prior distribution. If this estimator for λ^2 is used in (2.1), $\hat{\sigma}^2$ substituted for σ^2, the term not depending on L is dropped and the expression multiplied by $(1+c)^2$, then (2.4) is obtained.

An alternative form of (2.4) is $\|\boldsymbol{Y}\|^2 - \|P_L \boldsymbol{Y}\|^2 + \alpha\dim(L)\hat{\sigma}^2$ and if we omit $\|\boldsymbol{Y}\|^2$, we obtain the equivalent rule of selecting L to minimize

$$\alpha\dim(L)\hat{\sigma}^2 - \|P_L \boldsymbol{Y}\|^2. \tag{2.6}$$

If M and $\mathcal{L}$ have the structure (1.3) and the L_i's are **mutually orthogonal**, then (2.6) becomes

$$\alpha\hat{\sigma}^2\dim(L_0) - \|P_{L_0}\boldsymbol{Y}\|^2 + \sum_{j\in J}[\alpha\hat{\sigma}^2\dim(L_j) - \|P_{L_j}\boldsymbol{Y}\|^2] \tag{2.7}$$

$$= \hat{\sigma}^2\left[l_0(\alpha - F_0) + \sum_{j\in J} l_j(\alpha - F_j)\right]$$

where

$$l_j = \dim(L_j) \quad \text{and} \quad F_j = \|P_{L_j}\boldsymbol{Y}\|^2/\hat{\sigma}^2 l_j \tag{2.8}$$

is the usual F-statistic for testing the hypothesis that $\boldsymbol{\mu} \in M \mid L_j$. It is evident that the minimizing L, $\hat{L}$ say, will be the direct sum of L_0 and those L_j's which contribute negative terms to the sum in (2.7). Thus with $I(A)$ denoting the indicator function of the event A, we can write

$$\hat{L} = L_0 + \sum_{j=1}^{k} L_j I(F_j > \alpha). \tag{2.9}$$

Shibata (1986) considers the following related structure. Suppose that $M = L_1 + \cdots + L_k$ and $\mathcal{L}$ is restricted to the $k+1$ subspaces of the form $L = L_1 + \cdots + L_r$ with $r = 0, 1, \ldots, k$ (L is taken as the null space if $r = 0$). An example of such a structure is provided by the multiple linear regression problem in which the columns are to be entered in a given fixed sequence. There is no loss of generality in assuming the L_j's mutually orthogonal in this case since they could be replaced by an orthogonal sequence spanning the same sequence of spaces otherwise. Then (2.6) becomes

$$\sum_{j=1}^{r} [\alpha \hat{\sigma}^2 \dim(L_j) - \|P_{L_j} \boldsymbol{Y}\|^2] = \hat{\sigma}^2 \sum_{j=1}^{r} l_j(\alpha - F_j) \tag{2.10}$$

and the selected L is $\hat{L} = L_1 + \cdots + L_{\hat{r}}$ where $\hat{r}$ is that choice of r which minimizes (2.10). It does not seem possible to express $\hat{r}$ in a simple form and this considerably complicates the treatment of the risk of $P_{\hat{L}} \boldsymbol{Y}$ and the determination of good choices of α. By contrast, for $\hat{L}$ of (2.9) it is possible to obtain fairly explicit results as we show in the next section.

3. Limiting maximal risk

In this section we study the risk of the estimator $P_{\hat{L}} \boldsymbol{Y}$ of $\boldsymbol{\mu}$, with $\hat{L}$ selected by (2.6). In general it seems impossible to obtain analytically tractable expressions for the risk of $P_{\hat{L}} \boldsymbol{Y}$. We therefore restrict our discussion to the case where M and $\mathcal{L}$ have the structure (1.3) and the L_i's are mutually orthogonal, so that $\hat{L}$ is given by (2.9). We consider a further generalization by allowing the possibility that α in (2.9) may vary with L_j so that it may depend on j and $\hat{L}$ is given by

$$\hat{L} = L_0 + \sum_{j=1}^{k} L_j I(F_j > \alpha_j). \tag{3.1}$$

Then the risk of $P_{\hat{L}} \boldsymbol{Y}$ is

$$\begin{aligned} (3.2)\quad & E\|P_{\hat{L}} \boldsymbol{Y} - \boldsymbol{\mu}\|^2 \\ &= E\left\| P_{L_0}(\boldsymbol{Y} - \boldsymbol{\mu}) + \sum_{j=1}^{k} P_{L_j}[(\boldsymbol{Y} - \boldsymbol{\mu}) I(F_j > \alpha_j) - \boldsymbol{\mu} I(F_j \le \alpha_j)] \right\|^2 \\ &= \sigma^2 l_0 + \sum_{j=1}^{k} [E\|P_{L_j}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2 I(F_j > \alpha_j) + \|P_{L_j}\boldsymbol{\mu}\|^2 P(F_j \le \alpha_j)] \\ &= \sigma^2 l_0 + \sum_{j=1}^{k} [\sigma^2 l_j + E(\|P_{L_j}\boldsymbol{\mu}\|^2 - \|P_{L_j}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2) I(F_j \le \alpha_j)] \\ &= \sigma^2 m + \sum_{j=1}^{k} E(\|P_{L_j}\boldsymbol{\mu}\|^2 - \|P_{L_j}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2) I(F_j \le \alpha_j). \end{aligned}$$

Let $\boldsymbol{Z} = (\boldsymbol{Y} - \boldsymbol{\mu})/\sigma$, $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}/\sigma$ and $\lambda_j = \|P_{L_j}\boldsymbol{\mu}\|/\sigma = \|P_{L_j}\tilde{\boldsymbol{\mu}}\|$. Also let B_j denote the subspace of L_j spanned by $P_{L_j}\tilde{\boldsymbol{\mu}}$. Then

$$
\begin{aligned}
(3.3) \qquad \|P_{L_j}\boldsymbol{Y}\|^2 &= \sigma^2\|P_{L_j}\tilde{\boldsymbol{\mu}} + P_{L_j}\boldsymbol{Z}\|^2 \\
&= \sigma^2[\|P_{L_j}\tilde{\boldsymbol{\mu}}\|^2 + 2\langle P_{L_j}\tilde{\boldsymbol{\mu}}, P_{L_j}\boldsymbol{Z}\rangle + \|P_{L_j}\boldsymbol{Z}\|^2] \\
&= \sigma^2[\lambda_j^2 + 2\langle P_{L_j}\tilde{\boldsymbol{\mu}}, \boldsymbol{Z}\rangle + \|P_{B_j}\boldsymbol{Z}\|^2 + \|P_{L_j|B_j}\boldsymbol{Z}\|^2] \\
&= \sigma^2[(U_j + \lambda_j)^2 + W_j]
\end{aligned}
$$

where

$$
(3.4) \qquad W_j = \|P_{L_j|B_j}\boldsymbol{Z}\|^2 \quad \text{and} \quad U_j = \langle P_{L_j}\tilde{\boldsymbol{\mu}}/\lambda_j, \boldsymbol{Z}\rangle
$$

so that $U_j^2 = \|P_{B_j}\boldsymbol{Z}\|^2$. Similarly

$$
(3.5) \qquad \|P_{L_j}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2 = \sigma^2[U_j^2 + W_j].
$$

Also put

$$
(3.6) \qquad V = \hat{\sigma}^2/\sigma^2 \quad \text{and} \quad \nu = n - m.
$$

Then (3.2) becomes

$$
(3.7) \qquad E\|P_{\hat{L}}\boldsymbol{Y} - \boldsymbol{\mu}\|^2 = \sigma^2\left[m + \sum_{j=1}^{k} l_j h(\lambda_j, \alpha_j; l_j, \nu)\right]
$$

with

$$
(3.8) \qquad h(\lambda, \alpha; l, \nu) = \frac{1}{l}E(\lambda^2 - U^2 - W)I[(U + \lambda)^2 + W \le alV]
$$

where

$$
(3.9) \qquad U \text{ is } N(0,1)\text{-}, \quad W \text{ is } \chi^2_{l-1}\text{-} \quad \text{and} \quad \nu V \text{ is } \chi^2_\nu\text{-distributed},
$$

all independently. (3.7) is a generalization of a result of Mallows (1973) who only treats the case of variable selection in multiple regression with σ^2 known.

To compare the risk (3.7) of $P_{\hat{L}}\boldsymbol{Y}$ with that of the minimax estimator $P_M\boldsymbol{Y}$, we need to study the function h of (3.8). In special cases it simplifies. For $l = 1$ we may take $W \equiv 0$ in (3.8). Also, when σ^2 is actually known, we can replace $\hat{\sigma}^2$ by σ^2 in (2.4) and modify $\hat{L}$ similarly. If we take $V \equiv 1$ in (3.8) and call the resulting function $h(\lambda, \alpha; l, \infty)$, then (3.7) with ν replaced by ∞ is seen to be the relevant expression for the risk of the corresponding $P_{\hat{L}}\boldsymbol{Y}$. The simplest case of all is when σ^2 is known and $l = 1$, for which

$$
\begin{aligned}
(3.10) \qquad h(\lambda, \alpha; 1, \infty) &= E(\lambda^2 - U^2)I((U + \lambda)^2 \le \alpha) \\
&= (\lambda^2 - 1)[\Phi(\lambda + \sqrt{\alpha}) - \Phi(\lambda - \sqrt{\alpha})] \\
&\quad + (\lambda + \sqrt{\alpha})\phi(\lambda + \sqrt{\alpha}) - (\lambda - \sqrt{\alpha})\phi(\lambda - \sqrt{\alpha})
\end{aligned}
$$

with ϕ and Φ the $N(0,1)$-density and distribution functions respectively. Note also that

$$h(\lambda,0;l,\nu)=0 \quad \text{and} \quad h(\lambda,\infty;l,\nu)=\lambda^2/l-1. \tag{3.11}$$

At $\lambda=0$ we have

$$h(0,\alpha;l,\nu)=-\frac{1}{l}E(U^2+W)I(U^2+W\le \alpha l V)=-F_{l+2,\nu}(\alpha l/(l+2)) \tag{3.12}$$

with $F_{r,s}(t)$ the $F_{r,s}$-distribution function. At the other extreme, when $\lambda\to\infty$, then

$$h(\lambda,\alpha;l,\nu)\to 0. \tag{3.13}$$

We have been able to show analytically that for any fixed α with $0<\alpha<\infty$, $h(\lambda,\alpha;1,\infty)$ is unimodal in $\lambda\ge 0$. Its minimum value is given by the $\nu=\infty$ equivalent of (3.12) with $l=1$, viz. $-G_3(\alpha)=1+2\sqrt{\alpha}\phi(\sqrt{\alpha})-2\Phi(\sqrt{\alpha})$ with G_3 the χ^2_3-distribution function. Also, the maximum of $h(\lambda,\alpha;1,\infty)$ is strictly greater than 0 if $\alpha>0$. We studied $h(\lambda,\alpha;l,\nu)$ numerically for many other choices of l and ν and always found it unimodal in λ for $\lambda\ge 0$ and we conjecture that this is generally true. Proceeding on this conjecture, (3.12) gives the minimum value of $h(\lambda,\alpha;l,\nu)$ over λ generally, and there is a unique $\lambda^*=\lambda^*(\alpha;l,\nu)$ with $0<\lambda^*<\infty$ for $0<\alpha<\infty$ such that

$$h^*(\alpha;l,\nu)\equiv h(\lambda^*,\alpha;l,\nu)=\max_{\lambda\ge 0}h(\lambda,\alpha;l,\nu) \tag{3.14}$$

and this maximal value is strictly greater than 0. Figure 1 shows h as a function of λ for the case $\alpha=2$, $\nu=60$ and various values of l and Fig. 2 shows it for $l=5$, $\nu=60$ and various values of α. The unimodality of h as a function of λ is evident. We also see that $h(0,\alpha;l,\nu)$ decreases both when l increases and when α increases. Figure 3 shows h^* as a function of α for various values of l and it is evident that h^* increases as α increases. These conclusions also hold if we vary ν. The numerical work leading to these figures used subroutines DQDAGI and DUVMIF of IMSL.

Applying these findings in (3.1) and (3.7), the maximal risk of $P_{\hat{L}}\boldsymbol{Y}$ is

$$\sigma^2\left[m+\sum_{j=1}^{k}l_j h^*(\alpha_j;l_j,\nu)\right] \tag{3.15}$$

and this occurs at the **least favourable configuration** where $\boldsymbol{\mu}$ is such that

$$\|P_{L_j}\boldsymbol{\mu}\|=\sigma\lambda_j=\sigma\lambda^*(\alpha_j,l_j,\nu) \quad \text{for} \quad j=1,\dots,k. \tag{3.16}$$

Also the minimal risk of $P_{\hat{L}}\boldsymbol{Y}$ is

$$\sigma^2\left[m+\sum_{j=1}^{k}l_j h(0,\alpha_j;l_j,\nu)\right]=\sigma^2\left[m-\sum_{j=1}^{k}l_j F_{l_j+2,\nu}(\alpha_j l_j/(l_j+2))\right] \tag{3.17}$$

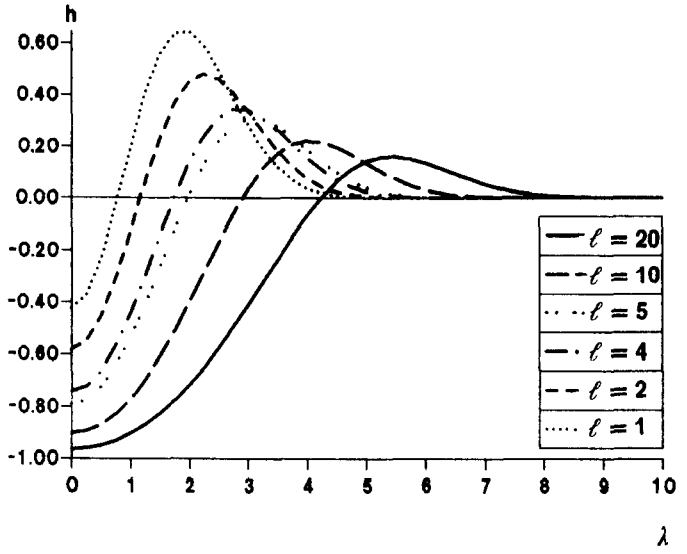

Fig. 1. $h(\lambda, 2; l, 60)$.

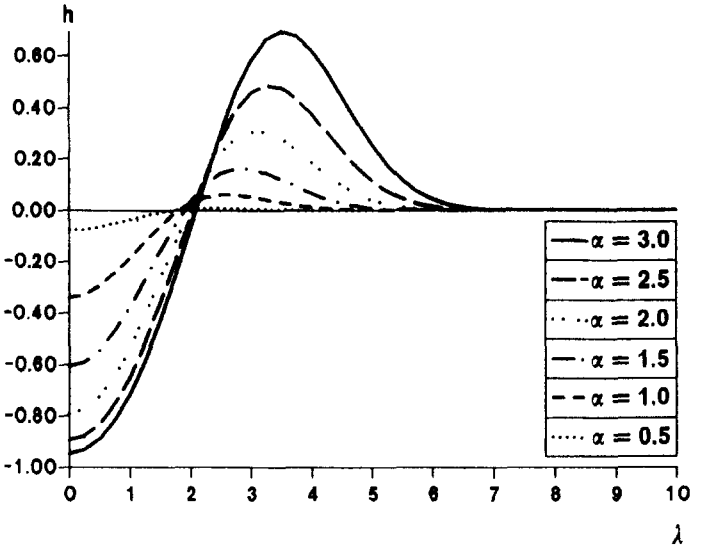

Fig. 2. $h(\lambda, \alpha; 5, 60)$.

and this occurs at the **most favourable configuration** where

$$\|P_{L_j}\mu\| = 0 \quad \text{for} \quad j = 1, \ldots, k. \tag{3.18}$$

The risk of $P_M \boldsymbol{Y}$ is $\sigma^2 m$ and it is convenient to express (3.15) and (3.17) relative to $\sigma^2 m$. Thus the maximal and minimal relative risks of $P_{\hat{L}} \boldsymbol{Y}$ compared to $P_M \boldsymbol{Y}$

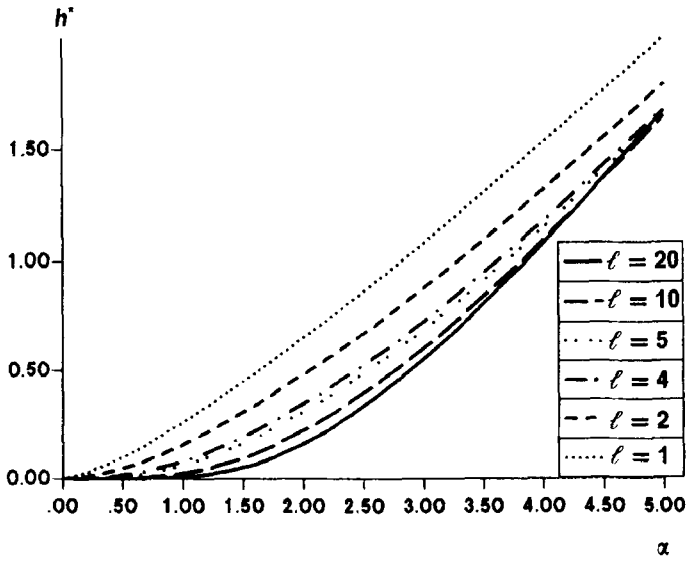

Fig. 3. $h^*(\alpha; l, 60)$.

are

$$1 + \frac{1}{m}\sum_{j=1}^{k} l_j h^*(\alpha_j; l_j, \nu) \quad \text{and} \quad 1 - \frac{1}{m}\sum_{j=1}^{k} l_j F_{l_j+2,\nu}(\alpha_j l_j/(l_j+2)) \tag{3.19}$$

respectively. The term $(1/m)\sum_{j=1}^{k} l_j h^*(\alpha_j; l_j, \nu)$ is the largest fraction by which the risk of $P_{\hat{L}}\boldsymbol{Y}$ can exceed that of $P_M \boldsymbol{Y}$ and we will refer to it as the **exceedance** of $P_{\hat{L}}\boldsymbol{Y}$. The term $(1/m)\sum_{j=1}^{k} l_j F_{l_j+2,\nu}(\alpha_j l_j/(l_j+2))$ is the largest gain in relative risk possible by using $P_{\hat{L}}\boldsymbol{Y}$ rather than $P_M \boldsymbol{Y}$ and it will be referred to as the **gain** of $P_{\hat{L}}\boldsymbol{Y}$.

We now consider the choice of the α_j's. One reasonable goal is to choose the α_j's to limit the exceedance of $P_{\hat{L}}\boldsymbol{Y}$. Thus if ϵ is some prescribed number (say $\epsilon = 0.25$) we may require that

$$\frac{1}{m}\sum_{j=1}^{k} l_j h^*(\alpha_j; l_j, \nu) = \epsilon. \tag{3.20}$$

If we insist that all the α_j's be the same (in the spirit of Section 2), then equation (3.20) can be solved for α with $\alpha_j = \alpha$ for all j. The resulting α will be denoted by $\alpha^* = \alpha^*(\epsilon; l_1, \ldots, l_k, \nu)$. Another approach which is simpler to implement in practice is to allow the α_j's to differ and to take

$$\alpha_j = \tilde{\alpha}_j = \tilde{\alpha}\left(\frac{m\epsilon}{m - l_0}; l_j, \nu\right) \tag{3.21}$$

where $\tilde{\alpha}(\eta; l, \nu)$ is the solution of the equation

$$h^*(\alpha; l, \nu) = \eta. \tag{3.22}$$

Then (3.20) is satisfied. For the case $l_0 = 0$, (3.21) reduces to

$$\alpha_j = \tilde{\alpha}(\epsilon; l_j, \nu) \tag{3.23}$$

and if we make this choice even for $l_0 > 0$, we will be acting conservatively in the sense that the exceedance is strictly below ϵ. Table 1 gives the values of $\tilde{\alpha}(\epsilon; l, \nu)$ for a range of ϵ-, l- and ν-values.

As an **example** consider the two way balanced ANOVA data of Table D on page 140 of Scheffé (1959). In that case $n = 90$, $m = 30$, $\nu = 60$ and the error sum of squares is 112.67 yielding $\hat{\sigma}^2 = 1.88$. A constant term for the overall mean effect is presumed included in the model so that $l_0 = 1$. With subspaces L_1 and L_2 corresponding to main effects of "sources" and "cylinders" respectively and L_3 to interaction effects, columns 2 and 3 of Table 2 show the dimensions (degrees of freedom) l_j and the observed statistics F_j. We calculated numerically the values of $\alpha^* = \alpha^*(\epsilon; 4, 5, 20, 60)$ and found the values of $\tilde{\alpha}_j = \tilde{\alpha}(\epsilon; l_j, 60)$ from Table 1 and the first two entries under each ϵ for each L_j block show these values. Comparing the F_j-values with these α_j-values we see that L_1 and L_2 are included in the selected model by either approach for the range of ϵ-values shown (and even for a much wider range). Regarding L_3 (interaction terms), selection based on the α^*-values includes these terms for $\epsilon = 0.05$ while selection based on the $\tilde{\alpha}_j$-values excludes interaction terms for $\epsilon = 0.05$; for larger ϵ-values, both approaches exclude the interaction terms. Since the α^*-value is a type of average of the $\tilde{\alpha}_j$-values and the $\tilde{\alpha}_j$-values increase with increasing l_j, we see that selection based on the α^*-values will more easily include high dimensional subspaces. Since high dimensional subspaces correspond to many parameters, selection based on the $\tilde{\alpha}_j$-values will tend to lead to more parsimonious models, which is desirable. The original C_p criterion corresponds to taking $\alpha = 2$ everywhere, has an exceedance of 0.212 and also selects the model which excludes interaction terms only.

While there is a large degree of consensus on the model to be selected for this data using either the "constant" α^*-values of the "variable" $\tilde{\alpha}_j$-values, the same might not be true for other examples and the question would be which is preferable. One possible answer is to prefer the procedure with the largest gain. For each of the α^*- and $\tilde{\alpha}_j$-values, the gains are shown in the first two rows of the GAIN block in Table 2. Evidently there is little to choose between the α^*- and the $\tilde{\alpha}_j$-values from this point of view. We note in passing that the gain corresponding to the C_p criterion is 0.874.

In practice a model is often selected by means of hypothesis testing. If $F(\gamma; l, \nu)$ is the $100\gamma\%$ upper critical value of the $F_{l,\nu}$-distribution, the subspace L_j is included in the selected model if $F_j \geq F(\gamma; l_j, \nu)$ on the basis of a γ-level test. Evidently this is equivalent to the estimation approach to model selection with "variable" α_j's now given by $\tilde{\alpha}_j = F(\gamma; l_j, \nu)$. The exceedance associated with this procedure is

$$\frac{1}{m} \sum_{j=1}^{k} l_j h^*(F(\gamma; l_j, \nu); l_j, \nu) \tag{3.24}$$

which is a function of γ. This function is strictly decreasing from ∞ at $\gamma = 0$ to 0 at $\gamma = 1$ so that we can find the choice of $\gamma = \gamma(\epsilon; l_1, \ldots, l_k, \nu)$ which makes

Table 1. Values of $\tilde{\alpha}(\epsilon; l; \nu)$.

	ϵ	$\nu = 15$	$\nu = 30$	$\nu = 60$	$\nu = 100$	$\nu = \infty$
$l = 1$	0.05	0.302	0.306	0.308	0.309	0.310
	0.10	0.491	0.497	0.500	0.501	0.503
	0.25	0.955	0.962	0.966	0.968	0.970
	0.50	1.619	1.625	1.628	1.629	1.631
$l = 2$	0.05	0.501	0.512	0.518	0.520	0.523
	0.10	0.745	0.758	0.765	0.768	0.773
	0.25	1.300	1.314	1.321	1.324	1.329
	0.50	2.045	2.052	2.055	2.057	2.059
$l = 3$	0.05	0.655	0.673	0.683	0.687	0.693
	0.10	0.926	0.946	0.957	0.962	0.969
	0.25	1.517	1.536	1.545	1.549	1.555
	0.50	2.282	2.288	2.290	2.291	2.291
$l = 4$	0.05	0.774	0.799	0.813	0.818	0.827
	0.10	1.060	1.086	1.100	1.106	1.115
	0.25	1.665	1.686	1.697	1.702	1.708
	0.50	2.432	2.434	2.435	2.435	2.434
$l = 5$	0.05	0.869	0.900	0.917	0.924	0.935
	0.10	1.161	1.193	1.210	1.217	1.228
	0.25	1.771	1.795	1.808	1.812	1.820
	0.50	2.534	2.535	2.534	2.533	2.530
$l = 10$	0.05	1.146	1.201	1.231	1.244	1.265
	0.10	1.444	1.495	1.522	1.534	1.551
	0.25	2.045	2.077	2.091	2.096	2.102
	0.50	2.779	2.773	2.762	2.755	2.741
$l = 15$	0.05	1.282	1.352	1.392	1.410	1.436
	0.10	1.575	1.638	1.672	1.686	1.707
	0.25	2.161	2.198	2.212	2.216	2.219
	0.50	2.875	2.864	2.846	2.834	2.811
$l = 20$	0.05	1.362	1.445	1.492	1.512	1.544
	0.10	1.651	1.723	1.762	1.778	1.801
	0.25	2.226	2.266	2.280	2.283	2.284
	0.50	2.927	2.913	2.889	2.874	2.840
$l = 25$	0.05	1.416	1.507	1.561	1.583	1.619
	0.10	1.700	1.779	1.822	1.839	1.864
	0.25	2.267	2.309	2.324	2.326	2.323
	0.50	2.959	2.943	2.915	2.896	2.854

Table 2. Example.

L_j	l_j	F_j	α	$\epsilon =$	0.05	0.10	0.25	0.50
L_1	4	8.31	α^*		1.235	1.549	2.123	2.778
			$\tilde{\alpha}$		0.813	1.100	1.697	2.435
			$\bar{\alpha}$		1.290	1.769	2.764	3.987
L_2	5	5.94	α^*		1.235	1.549	2.123	2.778
			$\tilde{\alpha}$		0.917	1.210	1.808	2.534
			$\bar{\alpha}$		1.281	1.706	2.575	3.629
L_3	20	1.29	α^*		1.235	1.549	2.123	2.778
			$\tilde{\alpha}$		1.492	1.762	2.280	2.889
			$\bar{\alpha}$		1.203	1.424	1.846	2.329
	GAIN		α^*		0.571	0.751	0.891	0.939
			$\tilde{\alpha}$		0.628	0.742	0.865	0.928
			$\bar{\alpha}$		0.560	0.730	0.897	0.952

the exceedance equal to ϵ for any prescribed ϵ. In our example numerical work shows that for ϵ equal to 0.05, 0.10, 0.25 and 0.50 the corresponding values of γ are equal to 0.284, 0.147, 0.0355 and 0.0062 respectively. The corresponding values of the $\bar{\alpha}_j$'s are shown as the third entry under each ϵ in Table 2 and the gain of the corresponding estimator can be calculated as before and is shown in the last row of the GAIN block. In terms of exceedance and gain it is evident that the three procedures are very similar. However, notice that the $\bar{\alpha}_j$-values decrease with increasing l_j whereas the $\tilde{\alpha}_j$-values increase. This is readily seen to be generally true. Thus the hypothesis testing approach will also tend to select less parsimonious models than the estimation approach based on the $\tilde{\alpha}_j$-values (as is illustrated by the $\epsilon = 0.05$ case).

If all the subspaces L_i are of the same dimension, as for example in the case of a regression model when every $l_i = 1$, the three approaches introduced above for selecting $\hat{L}$ are readily seen to be equivalent. In such a case the α-values corresponding to some given value of ϵ can be obtained from Table 1.

4. Minimax regret

In this section we consider choosing the α_j's by the minimax regret approach of Shibata (1986) and Hosoya (1983) applied to the structure (1.3) with the orthogonality assumption as before. Since $M \in \mathcal{L}$ any $\boldsymbol{\mu} \in M$ must be in some $L \in \mathcal{L}$. Let $J(\boldsymbol{\mu})$ be the smallest subset of $\{1, 2, \ldots, k\}$ such that $\boldsymbol{\mu} \in L(\boldsymbol{\mu}) = L_0 + \sum_{j \in J(\boldsymbol{\mu})} L_j$. If $L(\boldsymbol{\mu})$ were known we could have used $P_{L(\boldsymbol{\mu})} \boldsymbol{Y}$ to estimate $\boldsymbol{\mu}$ entailing a risk

$$E\|P_{L(\boldsymbol{\mu})} \boldsymbol{Y} - \boldsymbol{\mu}\|^2 = \sigma^2 \dim(L(\boldsymbol{\mu})) = \sigma^2 \left[l_0 + \sum_{j \in J(\boldsymbol{\mu})} l_j \right]. \tag{4.1}$$

Then the **regret** of $P_{\hat{L}} \boldsymbol{Y}$ is defined as

$$\delta R(\boldsymbol{\mu}, \alpha_1, \ldots, \alpha_k) = E\|P_{\hat{L}} \boldsymbol{Y} - \boldsymbol{\mu}\|^2 - E\|P_{L(\boldsymbol{\mu})} \boldsymbol{Y} - \boldsymbol{\mu}\|^2. \tag{4.2}$$

Table 3. Minimax regret values and choices of α.

l	$\ddot{\alpha}(l,\nu)$			$h^*(\ddot{\alpha},l,\nu)$		
	$\nu=15$	$\nu=60$	$\nu=\infty$	$\nu=15$	$\nu=60$	$\nu=\infty$
1	1.887	1.878	1.875	0.607	0.601	0.599
2	1.934	1.897	1.887	0.457	0.442	0.437
3	1.922	1.880	1.864	0.363	0.356	0.348
4	1.914	1.859	1.838	0.325	0.299	0.290
5	1.907	1.834	1.813	0.290	0.260	0.248
10	1.881	1.773	1.726	0.203	0.158	0.140
15	1.870	1.737	1.673	0.168	0.114	0.093
20	1.863	1.714	1.637	0.148	0.090	0.066
25	1.859	1.698	1.610	0.136	0.074	0.049

By (3.7) and (4.1) this becomes

$$\delta R(\boldsymbol{\mu},\alpha_1,\ldots,\alpha_k) = \sigma^2\left[\sum_{j\in J(\boldsymbol{\mu})} l_j h(\lambda_j,\alpha_j;l_j,\nu) + \sum_{j\notin J(\boldsymbol{\mu})} l_j(1+h(0,\alpha_j;l_j,\nu))\right]. \tag{4.3}$$

To maximize the regret will respect to $\boldsymbol{\mu}\in M$, we first maximize with respect to $\boldsymbol{\mu}$ restricted such that $J(\boldsymbol{\mu})=J$ and then vary J over all subsets of $\{1,2,\ldots,k\}$. Clearly

$$\max_{\{\boldsymbol{\mu}:J(\boldsymbol{\mu})=J\}} \delta R(\boldsymbol{\mu},\alpha_1,\ldots,\alpha_k) = \sigma^2\sum_{j=1}^{k} l_j[I(j\in J)h^*(\alpha_j;l_j,\nu) + I(j\notin J)(1+h(0,\alpha_j;l_j,\nu))]. \tag{4.4}$$

It is readily seen that the equation

$$h^*(\alpha;l,\nu) = 1 + h(0,\alpha;l,\nu) \tag{4.5}$$

has a unique solution $\alpha=\ddot{\alpha}(l,\nu)$. For the choices $\ddot{\alpha}_j=\ddot{\alpha}(l_j,\nu)$, (4.4) becomes

$$\sigma^2\sum_{j=1}^{k} l_j h^*(\ddot{\alpha}_j;l_j,\nu) = \sigma^2\sum_{j=1}^{k} l_j[1-F_{l_j+2,\nu}(\ddot{\alpha}_j l_j/(l_j+2))] \tag{4.6}$$

which does not depend on J, and for any other choice of the α_j's we can choose J to make (4.4) larger than (4.6). Hence (4.6) is the value of the minimax regret and the $\ddot{\alpha}_j$-values are the minimax regret choices of the α_j's. For purposes of comparison it is convenient to divide (4.6) by $\sigma^2 m$ and speak of the relative minimax regret. Comparing (4.6) and (3.19) we see that the relative minimax regret is equal to the exceedance and is also related in a simple way to the gain corresponding to the minimax regret choice of the α_j's. Table 3 gives the values of $\ddot{\alpha}(l,\nu)$ and $h^*(\ddot{\alpha}(l,\nu);l,\nu)$ for some choices of l and ν. We note that $\ddot{\alpha}(l,\nu)$ does not vary much with l and ν and are fairly close to the C_p choice of $\alpha=2$. As a result the corresponding exceedances are rather large for small l.

Returning to our example, we get $\ddot{\alpha}_1 = 1.859$, $\ddot{\alpha}_2 = 1.834$ and $\ddot{\alpha}_3 = 1.714$ so that the selected model includes only the main effects which is in accordance with the model selected by most of the previously discussed criteria. The minimax relative regret is 0.143 which is also the exceedance and seems reasonably small.

5. Estimating the risk in the orthonormal case

Consider the estimator $P_{\hat{L}}\boldsymbol{Y}$ with $\hat{L}$ given by (3.1). We assume that the constants α_j in (3.1) have been chosen beforehand, e.g. to limit the exceedance of the resulting estimator to some given amount ϵ. The relative risk of $P_{\hat{L}}\boldsymbol{Y}$, given by

$$1 + \frac{1}{m}\sum_{j=1}^{k} l_j h(\lambda_j, \alpha_j; l_j, \nu) \tag{5.1}$$

is between the two limits in (3.19). Close to the least favourable configuration (3.16) the relative risk (5.1) will be close to the upper limit in (3.19), while the lower limit in (3.19) is approached if we are close to the most favourable configuration (3.18). In most cases, however, the true situation will be some intermediate configuration. For instance, the observed F_j-values in the example in Section 3 seem to indicate that λ_1 and λ_2, corresponding to the main effects, are significantly different from zero, while λ_3, corresponding to the interaction effects, is close to zero, so that it appears unlikely that we are close to the least or the most favourable configuration. To obtain an indication of the actual accuracy of $P_{\hat{L}}\boldsymbol{Y}$ as an estimator of μ we may estimate the relative risk (5.1) of $P_{\hat{L}}\boldsymbol{Y}$. A detailed study of the problem of finding a good estimator for (5.1) will not be done here; we will only give a brief outline of a simple approach and apply it to our example.

Since (5.1) is a linear combination of the $h(\lambda_j, \alpha_j; l_j, \nu)$ and the l_j are known, estimating (5.1) can be accomplished by estimating the individual terms. Considering such a term in general, let L be a given l-dimensional subspace of M and put $\lambda = \|P_L\mu\|/\sigma$ and assume that α is a known positive constant. We wish to find an estimator of $h = h(\lambda, \alpha; l, \nu)$ given by (3.8) based on the statistic $F = \|P_L\boldsymbol{Y}\|^2/l\hat{\sigma}^2$ which has a non-central $F_{l,\nu}(\lambda^2)$-distribution. Since $EF = (1+\lambda^2/l)\nu/(\nu-2)$ and the factor $\nu/(\nu-2)$ will usually be close to 1, $l(F-1)$ is an approximately unbiased estimator of λ^2. However, it may be negative. Truncating and taking its square root, a simple estimator of λ is

$$\hat{\lambda} = \sqrt{l(F-1)^+} \tag{5.2}$$

with $x^+ = \max(0, x)$. A corresponding estimator of h is

$$\hat{h} = h(\hat{\lambda}, \alpha; l, \nu). \tag{5.3}$$

To appraise this estimator consider its MSE, $E[\hat{h} - h]^2$. The minimax value of the problem of estimating h provides a norm with which the MSE of $\hat{h}$ can be compared to see whether the accuracy of $\hat{h}$ is within reach of what is possible for

this problem. To calculate this minimax value, consider a discrete prior placing probability π_i on the point $\lambda = \lambda_i$ for $i = 0, 1, \ldots$ where $\sum \pi_i = 1$ and we shall take $0 = \lambda_0 < \lambda_1 < \cdots$. The corresponding Bayes estimator of h is

$$\hat{h}_B(F) = \frac{\sum \pi_i h(\lambda_i, \alpha; l, \nu) f(F \mid \lambda_i, l, \nu)}{\sum \pi_i f(F \mid \lambda_i, l, \nu)} \tag{5.4}$$

where $f(x \mid \lambda, l, \nu)$ is the non-central $F_{l,\nu}(\lambda^2)$-density of F. The MSE of $\hat{h}_B$ is

$$\int_0^\infty [\hat{h}_B(x) - h(\lambda, \alpha; l, \nu)]^2 f(x \mid \lambda, l, \nu) dx \tag{5.5}$$

and if we can choose the π_i's and λ_i's such that (5.5) as a function of λ achieves equal global maxima in each of the points $\lambda = \lambda_i$, then (5.4) is the minimax estimator of h and this maximum of (5.5) is the minimax value for the problem of estimating h (see e.g. Lehmann ((1983), pp. 249–250)). Numerical studies showed that for $l \leq 2$ three point priors satisfying this requirement can be calculated while for $l > 2$ two point priors suffice. Figure 4 shows how the MSE's of $\hat{h}$ and the minimax estimator compare in the illustrative case $l = 5$, $\alpha = 2$ and $\nu = \infty$. For this case we find that $\lambda_1 = 2.55$ and $\pi_1 = 0.519$. While the MSE of $\hat{h}$ is larger than that of the minimax estimator for small values of λ, the converse is true for large values of λ. The minimax estimator itself has the unsatisfactory property that it does not approach 0 as $F \to \infty$ and its MSE does not approach 0 as $\lambda \to \infty$. Overall we feel that $\hat{h}$ provides a simple and reasonably accurate first estimator of (5.1).

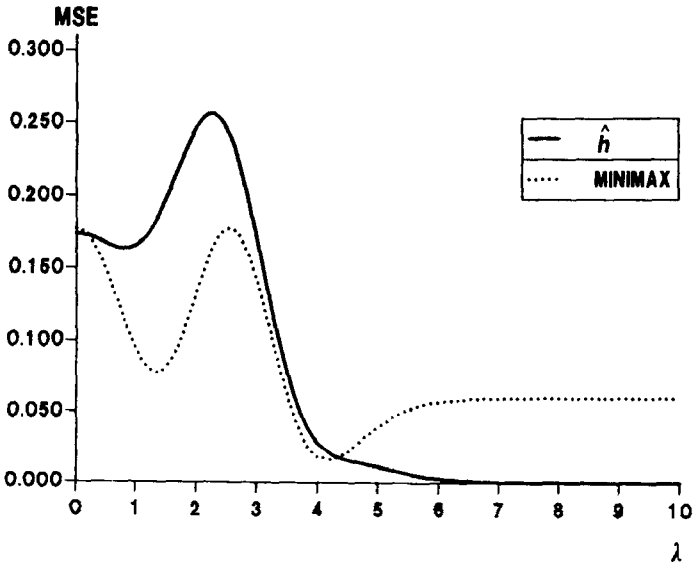

Fig. 4. Comparison of MSE of $\hat{h}$ with MSE of minimax estimator ($l = 5$, $\alpha = 2$, $\nu = \infty$).

To conclude we apply these results to the example of Section 3. For the minimax regret choice of the α_j's, the estimated overall relative risk is found to be 0.666. For the different cases reported in Table 2, the estimated overall relative risks are given in Table 4. The estimated relative risks in the third row of Table 4 first decrease as ϵ increases but increase when ϵ becomes large. This trend is also true for the first two rows with minima occurring between $\epsilon = 0.25$ and $\epsilon = 0.5$. Clearly these estimated relative risks indicate that substantial risk reduction is obtained by the selection rules studied here over a large range of ϵ values.

Table 4. Estimated relative risks.

α	$\epsilon =$	0.05	0.10	0.25	0.50
α^*		0.833	0.714	0.581	0.575
$\tilde{\alpha}$		0.732	0.648	0.567	0.559
$\bar{\alpha}$		0.846	0.761	0.658	0.689

6. Concluding discussion

(i) The strategy in Sections 2 and 3 to design the selection rule so that the maximal risk of the estimator is $1 + \epsilon$ times the minimax value has often been used in the literature (see e.g. Hodges and Lehmann (1952), Bickel (1984) etc.). An advantage of doing this is that different estimators are made comparable in terms of an important global property, viz. their maximum risks, and that they can then be compared sensibly in terms of other properties. However, in practice it might be difficult to commit oneself to a particular value of ϵ in order to obtain a uniquely specified choice of the α_j's. Fortunately as illustrated in our example, a wide range of ϵ-values may lead to the same conclusions, so that this need not be a critical issue. Of course, similar issues occur often; the choice of a significance level in hypothesis testing is another example.

(ii) In some problems no proper subspace M of $\mathbb{R}^n$ which is known to contain μ is available, i.e. effectively $M = \mathbb{R}^n$, $m = n$ and the usual estimator $\hat{\sigma}^2$ is not available. Therefore it is desirable to extend the results obtained in this paper to such cases and even to attempt to improve on these results when m is close to n by e.g. basing an estimator for σ^2 on $\|Q_{\hat{L}} \boldsymbol{Y}\|^2$.

(iii) Throughout this paper we assume that once a model (subspace $\hat{L}$) has been selected the estimator of μ is the projection of $\boldsymbol{Y}$ onto $\hat{L}$. This is the least squares estimator associated with $\hat{L}$ and is commonly used in practice. Other estimators may however also be considered such as an appropriate Stein type estimator which shrinks $P_M \boldsymbol{Y}$ towards $P_{\hat{L}} \boldsymbol{Y}$. Research on this possibility is in progress.

(iv) A major limitation of the results obtained in Sections 3 and 4 is the orthogonality assumption regarding the subspaces L_j in the assumed structure (1.3). Thus the results obtained can be applied directly to the variable selection problem in multiple regression only if the columns of the X-matrix are orthogonal. If this is not the case, one possible solution is to transform to principal components

and to select a model in terms of the principal components of the original variables (columns of the X-matrix) rather than in terms of the variables themselves (see e.g. Subsection 7.4 of Linhart and Zucchini (1986)). If the user insists on a model selected in terms of the original variables this solution will not be acceptable.

Acknowledgements

We would like to thank the referees for constructive criticism which led to substantial improvement of the original paper.

References

Arnold, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*, Wiley, New York.

Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion, *Biometrika*, **64**, 547–551.

Bickel, P. J. (1984). Parametric robustness: small biases can be worthwhile, *Ann. Statist.*, **12**, 864–879.

Hodges, J. L. and Lehmann, E. L. (1952). The use of previous experience in reaching statistical decisions, *Ann. Math. Statist.*, **23**, 396–407.

Hosoya, Y. (1983). Information criteria and tests in time-series models, *Time Series Analysis: Theory and Practice 5* (ed. O. D. Anderson), 39–52, North-Holland, Amsterdam.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

Linhart, H. and Zucchini, W. (1986). *Model Selection*, Wiley, New York.

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

Saxena, K. M. L. and Alam, K. (1982). Estimation of the non-centrality parameter of a chi squared distribution, *Ann. Statist.*, **10**, 1012–1016.

Scheffé, H. (1959). *The Analysis of Variance*, Wiley, New York.

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