Ann. Inst. Statist. Math.
Vol. 49, No. 2, 321–340 (1997)

LINEAR MODEL SELECTION BASED ON RISK ESTIMATION

J. H. VENTER[1] AND J. L. J. SNYMAN[2]

[1]*Department of Statistics, Potchefstroom University, Potchefstroom 2520, South Africa*
[2]*Department of Statistics, Rand Afrikaans University, Aucklandpark 2006, South Africa*

(Received November 6, 1995; revised April 15, 1996)

Abstract. The problem of selecting one model from a family of linear models to describe a normally distributed observed data vector is considered. The notion of the model of given dimension nearest to the observation vector is introduced and methods of estimating the risk associated with such a nearest model are discussed. This leads to new model selection criteria one of which, called the "partial bootstrap", seems particularly promising. The methods are illustrated by specializing to the problem of estimating the non-zero components of a parameter vector on which noisy observations are available.

Key words and phrases: Linear model selection, risk estimation, little bootstrap estimator, Stein estimation, selection criteria.

1. Introduction

In the standard linear model we observe the n-dimensional random vector $\boldsymbol{Y}$ having the form

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

where $\boldsymbol{e}$ is assumed to be $N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$-distributed with $\boldsymbol{I}_n$ the $n \times n$ identity matrix and where the assumption that we are dealing with a linear model amounts to assuming that $\boldsymbol{\mu} \in M$, a known linear subspace of $\mathcal{R}^n$. In the selection extended form of the linear model we assume that M actually relates to the most comprehensive linear model entertained and that there is a number of smaller linear models which may be more appropriate; to each such smaller linear model corresponds a linear subspace L of M and if the smaller model is true then $\boldsymbol{\mu} \in L$. Thus the selection extended formulation of the linear model assumes that, in addition to M and the knowledge that $\boldsymbol{\mu} \in M$, we are given a family $\mathcal{L}$ of linear subspaces of M and that it may actually be true that $\boldsymbol{\mu} \in L$ for some unknown member $L \in \mathcal{L}$. An example of this formulation is provided by the usual matrix form of linear regression in which $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{X}$ is an $n \times m$ design matrix and $\boldsymbol{\beta}$ an m-vector of regression coefficients. Here M is the column space of $\boldsymbol{X}$, but some of the components of $\boldsymbol{\beta}$ may be zero (the corresponding regressors being redundant) so that actually $\boldsymbol{\mu} \in L$ which is spanned by a subset of the columns of $\boldsymbol{X}$. Venter

and Steel (1992) supply further motivation and illustration of this formulation. In this paper our inference objective is to select a subspace $\hat{L} = \hat{L}(\boldsymbol{Y}) \in \mathcal{L}$ to which the data $\boldsymbol{Y}$ suggests that $\boldsymbol{\mu}$ belongs and to estimate $\boldsymbol{\mu}$ accordingly by an estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\boldsymbol{Y}) \in \hat{L}$; the selected subspace then identifies the selected model whose parameters follow from the estimator $\hat{\boldsymbol{\mu}}$.

We need the following notation. Let $\boldsymbol{x}'$ denote the transpose of $\boldsymbol{x}$ and $\boldsymbol{x}'\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$ the usual inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{R}^n$ and let $\|\boldsymbol{x}\| = (\boldsymbol{x}'\boldsymbol{x})^{1/2}$ be the Euclidean norm of $\boldsymbol{x}$. Further, let $L^{\perp}$ denote the orthogonal complement of the linear subspace L of $\mathcal{R}^n$ and if L is a subspace of the linear subspace K then let $K \mid L$ denote the orthogonal complement of L in K. Also, let $P_L\boldsymbol{x}$ denote the orthogonal projection of $\boldsymbol{x} \in \mathcal{R}^n$ on the subspace L and write $\dim(L)$ for the dimension of L.

Assume that $m = \dim(M) < n$. The usual unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \|P_{M^{\perp}}\boldsymbol{Y}\|^2/(n-m).$$

The traditional least squares estimator of $\boldsymbol{\mu}$ corresponding to any given $L \in \mathcal{L}$ is $P_L\boldsymbol{Y}$ and $\|P_{L^{\perp}}\boldsymbol{Y}\|^2 = \|\boldsymbol{Y} - P_L\boldsymbol{Y}\|^2$ is the associated residual (or error) sum of squares. A number of criteria for selecting L have been suggested in the literature. Among these are

- the final prediction error criterion which chooses L to minimize

$$FPE_{\alpha} = \|P_{L^{\perp}}\boldsymbol{Y}\|^2 + \alpha\dim(L)\hat{\sigma}^2 \tag{1.1}$$

with α a positive number; see Venter and Steel (1992) for a motivation of this criterion in the present context;

- the generalized cross validation criterion of Craven and Wahba (1979) which chooses L to minimize

$$GCV = \|P_{L^{\perp}}\boldsymbol{Y}\|^2/(n-\dim(L))^2; \tag{1.2}$$

see Eubank (1988) or Venter and Snyman (1995) for further motivation of this criterion;

- a proposal of Akaike (1970, 1973, 1974) which chooses L to minimize

$$AKA = \frac{n+\dim(L)}{n-\dim(L)}\|P_{L^{\perp}}\boldsymbol{Y}\|^2; \tag{1.3}$$

a motivation of this criterion in the present context is given in the appendix of this paper.

A common feature of all such criteria is that if $\dim(L)$ is held fixed then L is chosen to minimize $\|P_{L^{\perp}}\boldsymbol{Y}\|^2$. Henceforth we assume that $\mathcal{L}$ is a finite family. Let $L(\boldsymbol{Y}, p)$ denote that member of $\mathcal{L}$ nearest to $\boldsymbol{Y}$ among all members of $\mathcal{L}$ of the same dimension p, i.e.

$$\|P_{L(\boldsymbol{Y},p)^{\perp}}\boldsymbol{Y}\|^2 = \min\{\|P_{L^{\perp}}\boldsymbol{Y}\|^2 : L \in \mathcal{L} \text{ and } \dim(L) = p\}. \tag{1.4}$$

Then all the criteria listed above in effect restrict attention to selection among only these **nearest subspaces** $L(\boldsymbol{Y},p)$ with corresponding **nearest subspace** (NS) **estimators** $P_{L(\boldsymbol{Y},p)}\boldsymbol{Y}$ of $\boldsymbol{\mu}$ where p varies over the available dimensions of subspaces in $\mathcal{L}$. Thus, in the words of Breiman (1992), the problem reduces to that of selecting the dimension p only. For reference below note that since $\|P_{L^\perp}\boldsymbol{Y}\|^2 = \|P_{M|L}\boldsymbol{Y}\|^2 + \|P_{M^\perp}\boldsymbol{Y}\|^2$ and $\|P_{L^\perp}\boldsymbol{Y}\|^2 = \|\boldsymbol{Y}\|^2 - \|P_L\boldsymbol{Y}\|^2$ for $L \in \mathcal{L}$, we may also characterize $L(\boldsymbol{Y},p)$ as the minimizer of $\|P_{M|L}\boldsymbol{Y}\|^2$ or as the maximizer of $\|P_L\boldsymbol{Y}\|^2$ among the choices of L indicated in (1.4). Also since L is a subspace of M we have $\|P_L\boldsymbol{Y}\|^2 = \|P_L(P_M\boldsymbol{Y})\|^2$ so that $L(\boldsymbol{Y},p)$ is actually only a function of $P_M\boldsymbol{Y}$, i.e. $L(\boldsymbol{Y},p) = L(P_M\boldsymbol{Y},p)$; similarly $P_{L(\boldsymbol{Y},p)}\boldsymbol{Y} = P_{L(P_M\boldsymbol{Y},p)}(P_M\boldsymbol{Y})$ is also only a function of $P_M\boldsymbol{Y}$.

The derivations of the criteria listed above assumed that L was under consideration regardless of $\boldsymbol{Y}$. If we go one step further and suppose that the model dimension p is under consideration regardless of $\boldsymbol{Y}$ and for given p $L(\boldsymbol{Y},p)$ is to be selected, then the derivations of such criteria are no longer valid since $L(\boldsymbol{Y},p)$ now depends on $\boldsymbol{Y}$ and reconsideration is necessary. Let

$$R_p - E\|P_{L(\boldsymbol{Y},p)}\boldsymbol{Y} - \boldsymbol{\mu}\|^2 \tag{1.5}$$

be the risk of $P_{L(\boldsymbol{Y},p)}\boldsymbol{Y}$ with respect to squared error loss. Then one way to choose p is to find an estimator $\hat{R}_p$ of R_p for each available dimension p and to choose $\hat{p}$ to minimize $\hat{R}_p$. The final subspace selected would then be $L(\boldsymbol{Y},\hat{p})$ and the final estimator of $\boldsymbol{\mu}$ is $P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y}$. The problem then becomes that of estimating the risk R_p. Usually $P_{L(\boldsymbol{Y},p)}\boldsymbol{Y}$ is a very complicated function of $\boldsymbol{Y}$. The user may also wish to estimate the risk $E\|P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y} - \boldsymbol{\mu}\|^2$ of the final estimator $P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y}$ which is an even more complicated function of $\boldsymbol{Y}$. Therefore we need a general method of estimating risks of the form $E\|\boldsymbol{\delta}(\boldsymbol{Y}) - \boldsymbol{\mu}\|^2$ in a way which makes minimal assumptions on the estimator $\boldsymbol{\delta}(\boldsymbol{Y})$ of $\boldsymbol{\mu}$.

In Section 2 we study this issue in a canonical risk estimation case separately from the selection problem. We introduce a family of estimators which was motivated originally by the work of Breiman (1992) who introduced the so-called "little bootstrap" risk estimators in the context of variable selection in multiple linear regression. These risk estimators are generalized substantially here and their connections with Stein unbiased as well as Bayes risk estimation are pointed out. In Section 3 we return to the linear model selection problem and apply the results of Section 2 obtaining a new family of selection criteria relevant to linear model selection. To illustrate these criteria Section 4 specializes to the problem of simultaneously selecting and estimating the non-zero components of a parameter vector on which a noisy observation vector is available. Numerical work enables us to identify a member (referred to as the "partial bootstrap") of this family of criteria which has corresponding estimators whose final risk behaviour is quite appealing.

The discussion above and the application of the results of Section 2 in Sections 3 and 4 restrict attention to least squares (projection) estimators of $\boldsymbol{\mu}$. The method of Section 2 can also be applied to non-least squares (e.g. ridge-type) estimators. Section 5 closes with remarks on such possibilities and other matters.

2. A family of risk estimators

In this section we study the following canonical **risk estimation** problem: Given an observed m-dimensional vector $\boldsymbol{T}$ assumed to be $N_m(\boldsymbol{\theta}, \sigma^2 I_m)$-distributed and given also an estimator $\boldsymbol{h}(\boldsymbol{T})$ of $\boldsymbol{\theta}$, we wish to estimate the risk $R(\boldsymbol{\theta}) = E\|\boldsymbol{h}(\boldsymbol{T}) - \boldsymbol{\theta}\|^2$ using the observed data $\boldsymbol{T}$.

How this applies in the selection context of the previous section will be pointed out in Section 3. For now we assume that $\boldsymbol{\theta}$ is an arbitrary unknown vector in $\mathcal{R}^m$ but until further notice we assume that σ^2 is known and therefore omit it as an argument in the various risk function expressions. Regarding the estimator $\boldsymbol{h}(\boldsymbol{T})$ of $\boldsymbol{\theta}$ we assume that the various expectations below in which $\boldsymbol{h}(\boldsymbol{T})$ appears exist and that it is homogeneous of degree 1, i.e.

$$\boldsymbol{h}(a\boldsymbol{x}) = a\boldsymbol{h}(\boldsymbol{x}) \quad \text{for all } a > 0 \quad \text{and all } \boldsymbol{x} \in \mathcal{R}^m.$$

Let $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{h}(\boldsymbol{x})$ so that $\boldsymbol{g}(\boldsymbol{x})$ is also homogeneous of degree 1. Then the estimation risk of $\boldsymbol{h}(\boldsymbol{T})$ is

$$\begin{aligned} R(\boldsymbol{\theta}) &= E\|\boldsymbol{T} - \boldsymbol{\theta} - \boldsymbol{g}(\boldsymbol{T})\|^2 \\ &= E\|\boldsymbol{T} - \boldsymbol{\theta}\|^2 + E\|\boldsymbol{g}(\boldsymbol{T})\|^2 - 2E(\boldsymbol{T} - \boldsymbol{\theta})'\boldsymbol{g}(\boldsymbol{T}) \\ &= m\sigma^2 + \gamma(\boldsymbol{\theta}) - 2\psi(\boldsymbol{\theta}) \end{aligned} \tag{2.1}$$

where

$$\gamma(\boldsymbol{\theta}) = E\|\boldsymbol{g}(\boldsymbol{T})\|^2 \quad \text{and} \quad \psi(\boldsymbol{\theta}) = E(\boldsymbol{T} - \boldsymbol{\theta})'\boldsymbol{g}(\boldsymbol{T}).$$

If $\boldsymbol{g}$ satisfies the almost differentiability condition of Stein (1981) then we have

$$R(\boldsymbol{\theta}) = E\{m\sigma^2 + \|\boldsymbol{g}(\boldsymbol{T})\|^2 - 2\sigma^2 \operatorname{tr}(\boldsymbol{G}(\boldsymbol{T}))\}$$

where $\boldsymbol{G}(\boldsymbol{T})$ is the $m \times m$ matrix whose (i,j)-th element is $\partial g_i(\boldsymbol{T})/\partial T_j$ with $g_i(\boldsymbol{T})$ the i-th component of $\boldsymbol{g}(\boldsymbol{T})$. Hence, in this case

$$\hat{R}_S = m\sigma^2 + \|\boldsymbol{g}(\boldsymbol{T})\|^2 - 2\sigma^2 \operatorname{tr}(\boldsymbol{G}(\boldsymbol{T})) \tag{2.2}$$

is an unbiased estimator of $R(\boldsymbol{\theta})$. However, the almost differentiability condition will often not be satisfied in the selection applications so that (2.2) will then not be available. Even if it should be available in a particular case (2.2) may have relatively poor mean squared error behavior as is illustrated by the study of Venter and Steel (1990). These reasons prompt us to look for alternative risk estimators.

A family of estimators, indexed by real numbers $t > 0$, is given by

$$\hat{R}(\boldsymbol{T}, t) = m\sigma^2 + \|\boldsymbol{g}(\boldsymbol{T})\|^2 - 2\psi\left(\frac{1}{t}\boldsymbol{T}\right). \tag{2.3}$$

Here $\|\boldsymbol{g}(\boldsymbol{T})\|^2$ estimates $\gamma(\boldsymbol{\theta})$ in (2.1) unbiasedly and we shall motivate the use of $\psi(\frac{1}{t}\boldsymbol{T})$ to estimate $\psi(\boldsymbol{\theta})$ and discuss possible choices of t. Following Breiman

(1992), let $\boldsymbol{U}$ be $N_m(\mathbf{0}, \sigma^2 \boldsymbol{I}_m)$-distributed independently of $\boldsymbol{T}$ and define $\tilde{\boldsymbol{T}} = \boldsymbol{T} + t\boldsymbol{U}$. Put

$$
\begin{aligned}
\hat{\psi}(\boldsymbol{T}, t) &= t^{-1} E[\boldsymbol{U}' \boldsymbol{g}(\tilde{\boldsymbol{T}}) \mid \boldsymbol{T}] \\
&= t^{-2} E[(\tilde{\boldsymbol{T}} - \boldsymbol{T})' \boldsymbol{g}(\tilde{\boldsymbol{T}}) \mid \boldsymbol{T}] \\
&= E\left[\left(\frac{1}{t}\tilde{\boldsymbol{T}} - \frac{1}{t}\boldsymbol{T} \right)' \boldsymbol{g}\left(\frac{1}{t}\tilde{\boldsymbol{T}} \right) \,\middle|\, \boldsymbol{T} \right] \\
&= \psi\left(\frac{1}{t}\boldsymbol{T} \right)
\end{aligned} \tag{2.4}
$$

where we used homogeneity in the second last line and the definition of ψ in the last line. (2.4) yields various alternative expressions for $\hat{\psi}(\boldsymbol{T}, t) \equiv \psi(\frac{1}{t}\boldsymbol{T})$ and enables us to argue that it is approximately unbiased for $\psi(\boldsymbol{\theta})$ if t is chosen small. Notice that $E[\boldsymbol{U} \mid \tilde{\boldsymbol{T}}] = t(\tilde{\boldsymbol{T}} - \boldsymbol{\theta})/(1 + t^2)$ and that the marginal distribution of $\tilde{\boldsymbol{T}}$ is $N_m(\boldsymbol{\theta}, (1 + t^2)\sigma^2 \boldsymbol{I}_m)$. Hence

$$
\begin{aligned}
E\hat{\psi}(\boldsymbol{T}, t) &- t^{-1} E[\boldsymbol{U}' \boldsymbol{g}(\tilde{\boldsymbol{T}})] - t^{-1} E\{E[\boldsymbol{U}' \mid \tilde{\boldsymbol{T}}] \boldsymbol{g}(\tilde{\boldsymbol{T}})\} \\
&= E\left(\frac{\tilde{\boldsymbol{T}}}{\sqrt{1 + t^2}} - \frac{\boldsymbol{\theta}}{\sqrt{1 + t^2}} \right)' \boldsymbol{g}\left(\frac{\tilde{\boldsymbol{T}}}{\sqrt{1 + t^2}} \right) \\
&= \psi\left(\frac{\boldsymbol{\theta}}{\sqrt{1 + t^2}} \right).
\end{aligned} \tag{2.5}
$$

Therefore, if $\psi(\boldsymbol{\theta})$ is continuous as a function of $\boldsymbol{\theta}$ and t is small, then $E\hat{\psi}(\boldsymbol{T}, t) \approx \psi(\boldsymbol{\theta})$. In general $\hat{\psi}(\boldsymbol{T}, t)$ is not defined for $t = 0$ but under suitable regularity conditions on $\boldsymbol{g}$ (including differentiability) $\hat{\psi}(\boldsymbol{T}, t)$ approaches the Stein unbiased estimator of $\psi(\boldsymbol{\theta})$ as $t \to 0$. To see this, use Taylor expansions in the first line of (2.4) to get

$$
\begin{aligned}
\hat{\psi}(\boldsymbol{T}, t) &- t^{-1} E\left[\sum_{j=1}^{m} \boldsymbol{U}_j g_j(\boldsymbol{T} + t\boldsymbol{U}) \mid \boldsymbol{T} \right] \\
&= t^{-1} E\left[\sum_{j=1}^{m} \boldsymbol{U}_j g_j(\boldsymbol{T}) + t \sum_{j=1}^{m} \sum_{i=1}^{m} \boldsymbol{U}_i \boldsymbol{U}_j \frac{\partial}{\partial T_i} g_j(\boldsymbol{T}) + o(t) \mid \boldsymbol{T} \right] \\
&\to \sigma^2 \sum_{i=1}^{m} \frac{\partial}{\partial T_i} g_i(\boldsymbol{T}) = \sigma^2 \operatorname{tr}(\boldsymbol{G}(\boldsymbol{T})) \quad \text{as} \quad t \to 0.
\end{aligned}
$$

Thus in a sense (2.4) extends Stein unbiased risk estimators to approximately unbiased risk estimators applicable when the differentiability requirement of Stein does not necessarily hold. From the point of view of unbiasedness we will want to take t small, but this may be accompanied by a large variance and a bias-variance tradeoff may require taking t away from zero.

An intuitively appealing choice is $t = 1$, for then we are estimating $\psi(\boldsymbol{\theta})$ by its "plug-in" or parametric bootstrap estimator $\hat{\psi}(\boldsymbol{T}, 1) = \psi(\boldsymbol{T})$. The estimator

$\hat{\psi}(\boldsymbol{T}, t)$ also has a Bayes connection which suggests even larger choices of t. Suppose $\boldsymbol{\theta}$ is given the $N_m(\boldsymbol{0}, \sigma^2\tau^2\boldsymbol{I}_m)$ prior where $\tau^2 > 0$. Then it is easily seen that the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{T}$ is $N_m(\lambda\boldsymbol{T}, \lambda\sigma^2\boldsymbol{I}_m)$ where $\lambda = \tau^2/(1+\tau^2)$. Now we may express $\psi(\boldsymbol{\theta})$ as

$$\psi(\boldsymbol{\theta}) = E[\boldsymbol{Z}'\boldsymbol{g}(\boldsymbol{\theta} + \boldsymbol{Z})]$$

where $\boldsymbol{Z}$ is $N_m(\boldsymbol{0}, \sigma^2\boldsymbol{I}_m)$-distributed independently of $\boldsymbol{T}$ and $\boldsymbol{\theta}$. Consequently the Bayes estimator of $\psi(\boldsymbol{\theta})$ becomes

$$E[\psi(\boldsymbol{\theta}) \mid \boldsymbol{T}] = E[\boldsymbol{Z}'\boldsymbol{g}(\boldsymbol{\theta} + \boldsymbol{Z}) \mid \boldsymbol{T}] = E[(\boldsymbol{V} - \boldsymbol{\theta})'\boldsymbol{g}(\boldsymbol{V}) \mid \boldsymbol{T}]$$

where $\boldsymbol{V} = \boldsymbol{\theta} + \boldsymbol{Z}$. It is readily established that

$$E[(\boldsymbol{V} - \boldsymbol{\theta}) \mid \boldsymbol{V}, \boldsymbol{T}] = \frac{1}{1+\lambda}(\boldsymbol{V} - \lambda\boldsymbol{T})$$

so that we obtain

$$\begin{aligned} \boldsymbol{E}[\psi(\boldsymbol{\theta}) \mid \boldsymbol{T}] &= \boldsymbol{E}\{\boldsymbol{E}[(\boldsymbol{V} - \boldsymbol{\theta})' \mid \boldsymbol{V}, \boldsymbol{T}]\boldsymbol{g}(\boldsymbol{V}) \mid \boldsymbol{T}\} \\ &= \frac{1}{1+\lambda}\boldsymbol{E}\{(\boldsymbol{V} - \lambda\boldsymbol{T})'\boldsymbol{g}(\boldsymbol{V}) \mid \boldsymbol{T}\} \\ &\quad - \boldsymbol{E}\left\{\left(\frac{\boldsymbol{V}}{\sqrt{1+\lambda}} - \frac{\lambda\boldsymbol{T}}{\sqrt{1+\lambda}}\right)'\boldsymbol{g}\left(\frac{\boldsymbol{V}}{\sqrt{1+\lambda}}\right) \,\middle|\, \boldsymbol{T}\right\}. \end{aligned}$$

Also, since the conditional distribution of $\boldsymbol{V}/\sqrt{1+\lambda}$ given $\boldsymbol{T}$ is $N_m(\lambda\boldsymbol{T}/\sqrt{1+\lambda}, \sigma^2\boldsymbol{I}_m)$, it now follows from the definition of $\psi(\boldsymbol{\theta})$ that the Bayes estimator of $\psi(\boldsymbol{\theta})$ is

$$E[\psi(\boldsymbol{\theta}) \mid \boldsymbol{T}] = \psi\left(\frac{\lambda}{\sqrt{1+\lambda}}\boldsymbol{T}\right) = \psi\left(\frac{1}{t}\boldsymbol{T}\right) = \hat{\psi}(\boldsymbol{T}, t)$$

where $t = \sqrt{1+\lambda}/\lambda$. Notice that as τ^2 varies from 0 to ∞, λ varies from 0 to 1 and t from ∞ to $\sqrt{2}$. Thus, for $t > \sqrt{2}$ $\hat{\psi}(\boldsymbol{T}, t)$ is a Bayes estimator, the smallest allowed choice $t = \sqrt{2}$ corresponding to the vague prior choice $\tau^2 = \infty$. At the other extreme the choice $\tau^2 = 0$ reflects certainty that $\boldsymbol{\theta} = \boldsymbol{0}$ and this corresponds to $t = \infty$ and $\hat{\psi}(\boldsymbol{T}, \infty) = \psi(\boldsymbol{0})$ as the estimator of $\psi(\boldsymbol{\theta})$ which is sensible if it is indeed true that $\boldsymbol{\theta}$ is close to $\boldsymbol{0}$.

Breiman (1992) introduced $\hat{\psi}(\boldsymbol{T}, t)$ as in the second line of (2.4) in the variable selection context and also argued its approximate unbiasedness. The present treatment is more general and the systematic use of conditional expectations makes it somewhat more transparent. That $\hat{\psi}(\boldsymbol{T}, t) = \psi(\frac{1}{t}\boldsymbol{T})$ as well as the Stein and Bayes connections of $\hat{\psi}(\boldsymbol{T}, t)$ seem not to have been noticed before. Simulation results lead Breiman to suggest using $t = 0.6$ and he mentions that good results were obtained with t as large as 1 but felt that "its theoretical justification is weak" (Breiman (1992), p. 745). The results here go some way towards justifying

using larger values of t. Numerical evidence in Section 4 below also points in this direction.

The function ψ can often not be expressed in a simple form due to the complexity of $\boldsymbol{g}$. We then need to compute $\hat{\psi}(\boldsymbol{T}, t)$ by simulation using the first line of (2.4). The use of antithetic variables often improves this calculation. Since $-\boldsymbol{U}$ is distributed as $\boldsymbol{U}$, we may also write

$$\hat{\psi}(\boldsymbol{T}, t) = \frac{1}{2t} E[\boldsymbol{U}'\{\boldsymbol{g}(\boldsymbol{T} + t\boldsymbol{U}) - \boldsymbol{g}(\boldsymbol{T} - t\boldsymbol{U})\} \mid \boldsymbol{T}].$$

Consequently, to compute $\hat{\psi}(\boldsymbol{T}, t)$ we generate a large number B of independent copies $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(B)}$ of $\boldsymbol{U}$ and approximate $\hat{\psi}(\boldsymbol{T}, t)$ by

$$\frac{1}{2t} \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{U}^{(b)\prime}\{\boldsymbol{g}(\boldsymbol{T} + t\boldsymbol{U}^{(b)}) - \boldsymbol{g}(\boldsymbol{T} - t\boldsymbol{U}^{(b)})\}. \tag{2.6}$$

Until now we assumed σ^2 known; we now turn to the case where σ^2 is unknown but an estimator $\hat{\sigma}^2$ is available, independent of $\boldsymbol{T}$. Then se define $\boldsymbol{U}$ to be $N_m(\boldsymbol{0}, \hat{\sigma}^2 \boldsymbol{I}_m)$-distributed in the definition of $\tilde{\boldsymbol{T}}$ and take $\hat{\psi}(\boldsymbol{T}, t)$ as in the first line of (2.4). Also we replace σ^2 by $\hat{\sigma}^2$ in (2.2) and (2.3). Further modification of (2.2) would be required to make it exactly unbiased (see e.g. Stein (1981)) but since the Stein estimators will not be applicable in the rest of this paper there is no need to address this matter in more detail. Also expressions such as (2.5) are no longer strictly valid. However, they serve only to establish approximate unbiasedness for small t and we may expect this still to be true especially if the degrees of freedom on which $\hat{\sigma}^2$ is based is large. Finally, since (2.3) (or its version with σ^2 estimated) estimates a risk which is always non-negative, a further refinement is obtained by truncating at 0; therefore, in practice we shall use $\max\{0, \hat{R}(\boldsymbol{T}, t)\}$ rather than (2.3) as it stands.

3. Application to linear model selection

Returning to the setup of Section 1, let $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$ be an orthonormal basis for M and let $\boldsymbol{A}$ be the $n \times m$ matrix with columns $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$. Then $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_m$ and $\boldsymbol{A}\boldsymbol{A}' = P_M$. Define $\boldsymbol{T} = \boldsymbol{A}'\boldsymbol{Y}$ and $\boldsymbol{\theta} = \boldsymbol{A}'\boldsymbol{\mu}$. Then $\boldsymbol{T}$ is distributed as in Section 2. We are interested in the estimator $\boldsymbol{\delta}(\boldsymbol{Y}) = P_{L(\boldsymbol{Y},p)}\boldsymbol{Y}$ of $\boldsymbol{\mu}$. The corresponding estimator of $\boldsymbol{\theta}$ is $\boldsymbol{A}'P_{L(\boldsymbol{Y},p)}\boldsymbol{Y}$ and we now express this in terms of $\boldsymbol{T}$.

$\boldsymbol{A}'$ yields a linear transform from M onto $\mathcal{R}^m$ by setting $\boldsymbol{t} = \boldsymbol{A}'\boldsymbol{y}$ for any $\boldsymbol{y} \in M$. For any $\boldsymbol{t} \in \mathcal{R}^m$, we have $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{t} \in M$ so that $\boldsymbol{A}$ yields the inverse linear transformation from $\mathcal{R}^m$ to M. Also, if $\boldsymbol{y}, \boldsymbol{z} \in M$ and $\boldsymbol{t} = \boldsymbol{A}'\boldsymbol{y}$, $\boldsymbol{s} = \boldsymbol{A}'\boldsymbol{z}$, then $\boldsymbol{t}'\boldsymbol{s} = \boldsymbol{y}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{z} = \boldsymbol{y}'P_M\boldsymbol{z} = \boldsymbol{y}'\boldsymbol{z}$ so that inner products (and norms) are preserved. Further, any linear subspace L of M has a unique image linear subspace in $\mathcal{R}^m$ given by

$$L^* = \boldsymbol{A}'L = \{\boldsymbol{t} \in \mathcal{R}^m : \boldsymbol{t} = \boldsymbol{A}'\boldsymbol{y} \text{ with } \boldsymbol{y} \in L\}$$

which is such that $L = \boldsymbol{A}L^*$, where

$$\boldsymbol{A}L^* = \{\boldsymbol{y} \in M : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{t} \text{ with } \boldsymbol{t} \in L^*\}.$$

Then L and L^* have the same dimension and it is readily seen that $P_{L^*} = \boldsymbol{A}'P_L\boldsymbol{A}$ and $P_L = \boldsymbol{A}P_{L^*}\boldsymbol{A}'$. Define $\mathcal{L}^* = \{L^* : L \in \mathcal{L}\}$ and let $L^*(\boldsymbol{T}, p)$ be the nearest subspace of dimension p to $\boldsymbol{T}$ among all $L^* \in \mathcal{L}^*$, i.e.

$$\|P_{L^*(\boldsymbol{T},p)}\boldsymbol{T}\| = \max\{\|P_{L^*}\boldsymbol{T}\| : L^* \in \mathcal{L}^* \text{ and } \dim(L^*) = p\}.$$

Then we have $L^*(\boldsymbol{T}, p) = \boldsymbol{A}'L(\boldsymbol{Y}, p)$ and $P_{L^*(\boldsymbol{T},p)}\boldsymbol{T} = \boldsymbol{A}'P_{L(\boldsymbol{Y},p)}\boldsymbol{A}\boldsymbol{T} = \boldsymbol{A}'P_{L(\boldsymbol{Y},p)}\boldsymbol{Y}$. Hence in terms of $\boldsymbol{T}$ the estimator $\boldsymbol{A}'P_{L(\boldsymbol{Y},p)}\boldsymbol{Y}$ of $\boldsymbol{\theta}$ becomes $P_{L^*(\boldsymbol{T},p)}\boldsymbol{T}$ and this may be taken as the estimator $\boldsymbol{h}(\boldsymbol{T})$ of $\boldsymbol{\theta}$ whose risk estimation is studied in Section 2. Clearly, it is homogeneous of degree 1. The risk in (1.5) may now be expressed as

$$(3.1)\quad R_p = E\|P_{L(\boldsymbol{Y},p)}\boldsymbol{Y} - \boldsymbol{\mu}\|^2 = E\|\boldsymbol{A}P_{L^*(\boldsymbol{T},p)}\boldsymbol{T} - \boldsymbol{A}\boldsymbol{\theta}\|^2 = E\|P_{L^*(\boldsymbol{T},p)}\boldsymbol{T} - \boldsymbol{\theta}\|^2$$

which is just $R(\boldsymbol{\theta})$ of Section 2 with the present choice of $\boldsymbol{h}(\boldsymbol{T})$. We first consider the case σ^2 known and apply the risk estimators (2.3). We now have

$$\boldsymbol{g}(\boldsymbol{T}) = \boldsymbol{T} - P_{L^*(\boldsymbol{T},p)}\boldsymbol{T} = P_{L^*(\boldsymbol{T},p)^\perp}\boldsymbol{T}$$

which may be expressed in terms of $\boldsymbol{Y}$ as

$$(3.2)\quad \boldsymbol{g}(\boldsymbol{T}) = \boldsymbol{A}'\boldsymbol{Y} - \boldsymbol{A}'P_{L(\boldsymbol{Y},p)}\boldsymbol{Y} = \boldsymbol{A}'P_M\boldsymbol{Y} - \boldsymbol{A}'P_{L(\boldsymbol{Y},p)}\boldsymbol{Y} = \boldsymbol{A}'P_{M|L(\boldsymbol{Y},p)}\boldsymbol{Y}$$

so that

$$\|\boldsymbol{g}(\boldsymbol{T})\|^2 = \|A'P_{M|L(\boldsymbol{Y},p)}\boldsymbol{Y}\|^2 = \|P_{M|L(\boldsymbol{Y},p)}\boldsymbol{Y}\|^2.$$

To express $\hat{\psi}(\boldsymbol{T}, t)$ in terms of $\boldsymbol{Y}$, let $\boldsymbol{W}$ be $N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$-distributed independently of $\boldsymbol{Y}$ and put $\boldsymbol{U} = \boldsymbol{A}'\boldsymbol{W}$ so that $\boldsymbol{U}$ is $N_m(\boldsymbol{0}, \sigma^2\boldsymbol{I}_m)$-distributed independently of $\boldsymbol{T}$ as in Section 2. Define

$$\tilde{\boldsymbol{T}} = \boldsymbol{T} + t\boldsymbol{U} = \boldsymbol{A}'\boldsymbol{Y} + t\boldsymbol{A}'\boldsymbol{W} = \boldsymbol{A}'(\boldsymbol{Y} + t\boldsymbol{W}) = \boldsymbol{A}'\tilde{\boldsymbol{Y}}$$

with $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + t\boldsymbol{W}$. Then using (2.4) and (3.2)

$$(3.3)\quad \begin{aligned}\hat{\psi}(\boldsymbol{T}, t) &= t^{-1}E[\boldsymbol{U}'\boldsymbol{g}(\tilde{\boldsymbol{T}}) \mid \boldsymbol{T}] \\ &= t^{-1}E[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{A}'P_{M|L(\tilde{\boldsymbol{Y}},p)}\tilde{\boldsymbol{Y}} \mid \boldsymbol{T}] \\ &= t^{-1}E[\boldsymbol{W}'P_{M|L(\tilde{\boldsymbol{Y}},p)}\tilde{\boldsymbol{Y}} \mid P_M\boldsymbol{Y}] \\ &= t^{-1}E[\boldsymbol{W}'P_{M|L(\tilde{\boldsymbol{Y}},p)}\tilde{\boldsymbol{Y}} \mid \boldsymbol{Y}].\end{aligned}$$

Here conditioning on $\boldsymbol{T}$ may be replaced by conditioning on $P_M\boldsymbol{Y}$ since $\boldsymbol{T}$ and $P_M\boldsymbol{Y}$ are 1-1 functions of each other. As noted in Section 1 $P_{M|L(\tilde{\boldsymbol{Y}},p)}\tilde{\boldsymbol{Y}}$ is actually only a function of $P_M\tilde{\boldsymbol{Y}} = P_M\boldsymbol{Y} + tP_M\boldsymbol{W}$ which is independent of $P_{M^\perp}\boldsymbol{Y}$; hence we may replace conditioning on $P_M\boldsymbol{Y}$ by conditioning on both $P_M\boldsymbol{Y}$ and $P_{M^\perp}\boldsymbol{Y}$

or, equivalently, by conditioning on $\boldsymbol{Y} = P_M \boldsymbol{Y} + P_{M^\perp} \boldsymbol{Y}$. We shall now denote (3.3) by $\hat{\psi}_p(\boldsymbol{Y}, t)$ so that the estimator of R_p in (3.1) becomes

$$\hat{R}_p(\boldsymbol{Y}, t) = m\sigma^2 + \|P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y}\|^2 - 2\hat{\psi}_p(\boldsymbol{Y}, t). \tag{3.4}$$

Notice that we may write

$$\begin{aligned} R_p &= E\|P_{L(\boldsymbol{Y},p)} \boldsymbol{Y} - \boldsymbol{\mu}\|^2 \\ &= E\|P_M(\boldsymbol{Y} - \boldsymbol{\mu}) - P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y}\|^2 \\ &= m\sigma^2 + E\|P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y}\|^2 - 2\psi_p(\boldsymbol{\mu}) \end{aligned} \tag{3.5}$$

where

$$\psi_p(\boldsymbol{\mu}) = E(P_M(\boldsymbol{Y} - \boldsymbol{\mu}))' P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y} = E(\boldsymbol{Y} - \boldsymbol{\mu})' P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y}. \tag{3.6}$$

Therefore each of the last two terms in (3.4) estimates the corresponding term in (3.5). It is readily seen that we also have

$$\hat{\psi}_p(\boldsymbol{Y}, t) = \psi_p\left(\frac{1}{t} P_M \boldsymbol{Y}\right) \tag{3.7}$$

as the parallel of (2.4). It is instructive to look at the case where $L(\boldsymbol{Y}, p) = L_p$ is independent of $\boldsymbol{Y}$ (e.g. in case there is only one subspace $L_p \in \mathcal{L}$ of dimension p). Then (3.6) yields

$$\begin{aligned} \psi_p(\boldsymbol{\mu}) &= E(P_M(\boldsymbol{Y} - \boldsymbol{\mu}))' P_{M|L_p} \boldsymbol{Y} \\ &= E(P_{L_p}(\boldsymbol{Y} - \boldsymbol{\mu}) + P_{M|L_p}(\boldsymbol{Y} - \boldsymbol{\mu}))' P_{M|L_p}(\boldsymbol{Y} - \boldsymbol{\mu}) \\ &= E\|P_{M|L_p}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2 = \sigma^2 \dim(M \mid L_p) = \sigma^2(m - p) \end{aligned}$$

and (3.7) shows that also $\hat{\psi}_p(\boldsymbol{Y}, t) = \sigma^2(m - p)$. The criterion (3.4) then reduces to $\|P_{M|L_p} \boldsymbol{Y}\|^2 + \sigma^2(2p - m)$ which is a version of the well-known C_p-criterion (equivalent to the $\alpha - 2$ version of FPE_α; see Section 1 or e.g. Venter and Steel (1992)). Use of the C_p-criterion on the nearest subspace $L(\boldsymbol{Y}, p)$ is tantamount to estimating the risk of the NS estimator by

$$\|P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y}\|^2 + \sigma^2(2p - m) \tag{3.8}$$

which amounts to estimating $\psi_p(\boldsymbol{\mu})$ by $\sigma^2(m \quad p)$ but this may be quite erroneous if $L(\boldsymbol{Y}, p)$ does depend on $\boldsymbol{Y}$ as will be illustrated in Section 4 below.

A case of special interest is the choice $t = \infty$ for which we get $\hat{\psi}_p(\boldsymbol{Y}, \infty) = \psi_p(\boldsymbol{0})$. By (3.6)

$$\begin{aligned} \psi_p(\boldsymbol{0}) &= E_{\boldsymbol{0}}(P_M \boldsymbol{Y})' P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y} \\ &= E_{\boldsymbol{0}}\|P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y}\|^2 \\ &= \sigma^2 E\|P_{M|L(\boldsymbol{Z},p)} \boldsymbol{Z}\|^2 \\ &= \sigma^2(m - E\|P_{L(\boldsymbol{Z},p)} \boldsymbol{Z}\|^2) \end{aligned}$$

where E_0 indicates that expectation is to be calculated taking $\boldsymbol{\mu} = \mathbf{0}$ in the expressions involving $\boldsymbol{Y}$ and where $\boldsymbol{Z}$ is $N_n(\mathbf{0}, \boldsymbol{I}_n)$-distributed. Hence $\hat{\psi}_p(\boldsymbol{Y}, \infty) = \sigma^2(m - E\|P_{L(\boldsymbol{Z},p)}\boldsymbol{Z}\|^2)$ is independent of $\boldsymbol{Y}$ and depends only on p. With $t = \infty$ the criterion (3.4) therefore has the form

$$\hat{R}_p(\boldsymbol{Y}, \infty) = \|P_{M|L(\boldsymbol{Y},p)}\boldsymbol{Y}\|^2 + \sigma^2(2E\|P_{L(\boldsymbol{Z},p)}\boldsymbol{Z}\|^2 - m) \tag{3.9}$$

which differs from version (3.8) of the C_p criterion only in the replacement of the term $2p$ by the more complicated expression $2E\|P_{L(\boldsymbol{Z},p)}\boldsymbol{Z}\|^2$.

As in Section 2 (3.7) is typically not useful for calculation purposes and the equivalent of (2.6) must be used. Further, when σ^2 is not known but an estimator $\hat{\sigma}^2$ is available (e.g. as given in Section 1) then we proceed as described at the end of Section 2, i.e. we replace σ^2 by $\hat{\sigma}^2$ in (3.4) and in the calculation of $\hat{\psi}_p(\boldsymbol{Y}, t)$ we take $\boldsymbol{W}$ to be $N_n(\mathbf{0}, \hat{\sigma}^2\boldsymbol{I}_n)$-distributed. In practice we shall also truncate the resulting criterion at 0.

Thus (3.4) (or its variant when σ^2 is unknown) is a new family of linear model selection criteria. We select $\hat{p}$ to minimize $\hat{R}_p(\boldsymbol{Y}, t)$ over the available dimensions p. Then the selected model is that corresponding to the linear subspace $L(\boldsymbol{Y}, \hat{p})$ in $\mathcal{L}$ and the final estimator of $\boldsymbol{\mu}$ is $P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y}$. Some considerations regarding the choice of t were given in Section 2. Those centered on the issue of risk estimation but the results are now used as selection criteria so that those considerations may no longer be relevant. The main concern must now be the effect that the choice of t has on the final estimator $P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y}$ of $\boldsymbol{\mu}$. We shall consider this issue in the context of a special case in the next section.

4. Selecting and estimating non-zero means

In this section we specialize to the following canonical **selection and estimation** problem:
We are given independent observations Y_i assumed to be $N(\mu_i, \sigma^2)$-distributed, $i = 1, \ldots, n$; for some given integer m we know that $\mu_i = 0$ for $i > m$ while for $i \leq m$ the μ_i's may or may not be zero. Our goal is to select (identify) the non-zero μ_i's and to estimate them; a secondary goal is to estimate σ^2 as well.

It is easily seen that the variable selection problem of multiple linear regression with an orthogonal design matrix can be transformed into this form (see Venter and Steel (1992)). If we should happen to know the value of σ^2 we may disregard $Y_{m+1}, \ldots, Y_n$ since they only contribute information on σ^2. This special case is discussed in Venter and Steel (1992, 1994) where several selection criteria are derived and compared. If σ^2 is unknown but $m < n$ we have the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-m}\sum_{j=m+1}^{n} Y_j^2 \tag{4.1}$$

and, especially if $n - m$ is large, this case can be treated much like the σ^2 known case, although the issue of improving on (4.1) when we should conclude that many of the μ_i's are zero for $i \leq m$, is also pertinent. Another important special case

is when σ^2 is unknown but $m = n$, i.e. we have no designated μ_i's which are definitely known to be zero a priori. Aspects of this problem have been discussed fairly extensively in the literature under the title of "identifying active contrasts". This literature goes back to the half-normal plot of Daniel (1959) but more recent references are Dong (1993), Box and Meyer (1986) and Venter and Steel (1996). This case also plays a pivotal role in the recent work of Donoho and Johnstone (1994) in whose work our Y_i's are the wavelet transforms of an initial series of observations. Thus it is clear that this canonical problem is of substantial interest.

As before we start with the case σ^2 known and we evaluate the criterion (3.4) for the present problem. To put it in the form of Section 1, let $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ be the standard orthonormal basis of $\mathcal{R}^n$ (i.e. $\boldsymbol{a}_i$ has i-th component equal to 1 and all other components equal to 0). Then

$$\boldsymbol{Y} = \sum_{i=1}^{n} Y_i \boldsymbol{a}_i \quad \text{and} \quad \boldsymbol{\mu} = \sum_{i=1}^{n} \mu_i \boldsymbol{a}_i \tag{4.2}$$

so that M and $M^\perp$ are spanned by the orthonormal sets $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$ and $\{\boldsymbol{a}_{m+1}, \boldsymbol{a}_{m+2}, \ldots, \boldsymbol{a}_n\}$ respectively. A typical subspace L of M consists of $\boldsymbol{\mu}$'s of the form (4.2) having some of its components 0. If $\mu_i = 0$ for $i \notin J \subseteq \{1, \ldots, m\}$ then $L = \text{span}\{\boldsymbol{a}_j, j \in J\}$. We let $\mathcal{L}$ be the family of 2^m such subspaces obtained by varying J over all subsets of $\{1, \ldots, m\}$. For $J = \emptyset$ we take $L = \{\boldsymbol{0}\}$. In general we have

$$\|P_{L^\perp} \boldsymbol{Y}\|^2 = \|\boldsymbol{Y}\|^2 - \|P_L \boldsymbol{Y}\|^2 = \|\boldsymbol{Y}\|^2 - \sum_{j \in J} Y_j^2. \tag{4.3}$$

For J such that it has a fixed number p of elements (4.3) is minimized by taking J the set of indices of the p largest $|Y_j|$'s among $|Y_1|, |Y_2|, \ldots, |Y_m|$, i.e. if $|Y_{l_1}| < |Y_{l_2}| < \cdots < |Y_{l_m}|$ are the order statistics and $J(\boldsymbol{Y}, p) = \{l_{m-p+1}, l_{m-p+2}, \ldots, l_m\}$, then $L(\boldsymbol{Y}, p) = \text{span}\{\boldsymbol{a}_j, j \in J(\boldsymbol{Y}, p)\}$. Consequently the NS estimators are given by

$$P_{L(\boldsymbol{Y},p)} \boldsymbol{Y} = \sum_{i=m-p+1}^{m} Y_{l_i} \boldsymbol{a}_{l_i} = \sum_{j=1}^{m} Y_j I(|Y_j| > |Y_{l_{m-p}}|) \boldsymbol{a}_j \tag{4.4}$$

and its j-th component estimates μ_j by Y_j if $|Y_j|$ is among the p largest $|Y_i|$'s and by 0 otherwise. Here $I(A)$ is the indicator function of the event A. Now

$$P_{M|L(\boldsymbol{Y},p)} \boldsymbol{Y} = P_M \boldsymbol{Y} - P_{L(\boldsymbol{Y},p)} \boldsymbol{Y} = \sum_{j=1}^{m} Y_j I(|Y_j| \le |Y_{l_{m-p}}|) \boldsymbol{a}_j$$

so that from (3.3) we get

$$\hat{\psi}_p(\boldsymbol{Y}, t) = t^{-1} E\left[\sum_{j=1}^{m} W_j \tilde{Y}_j \tilde{I}_j \mid \boldsymbol{Y}\right] \tag{4.5}$$

where $\tilde{Y}_j = Y_j + tW_j$ and $\tilde{I}_j = I(|\tilde{Y}_j| \leq |\tilde{Y}_{\tilde{l}_{m-p}}|)$ for $j = 1, \ldots, m$ with $|\tilde{Y}_{\tilde{l}_1}| < |\tilde{Y}_{\tilde{l}_2}| < \cdots < |\tilde{Y}_{\tilde{l}_m}|$ the order statistics of $|\tilde{Y}_1|, |\tilde{Y}_2|, \ldots, |\tilde{Y}_m|$ and where $W_1, W_2, \ldots, W_m$ are independent $N(0, \sigma^2)$-distributed given $\boldsymbol{Y}$. Using the antithetic device mentioned in Section 2, if $\tilde{Y}_j^* = Y_j - tW_j$ and $\tilde{I}_j^*$ is defined as $\tilde{I}_j$ but in terms of the $\tilde{Y}_j^*$'s, then

$$\hat{\psi}_p(\boldsymbol{Y}, t) = (2t)^{-1} E\left[\sum_{j=1}^{m} W_j(\tilde{Y}_j \tilde{I}_j - \tilde{Y}_j^* \tilde{I}_j^*) \mid \boldsymbol{Y}\right]. \tag{4.6}$$

Simulation evaluation of $\hat{\psi}_p$ using this expression is more efficient than using (4.5). The selection criterion (3.4) now becomes

$$\begin{aligned} \hat{R}_p(\boldsymbol{Y}, t) &= m\sigma^2 + \sum_{j=1}^{m} Y_j^2 I(|Y_j| \leq |Y_{l_{m-p}}|) - 2\hat{\psi}_p(\boldsymbol{Y}, t) \\ &= m\sigma^2 + \sum_{j=1}^{m-p} Y_{l_j}^2 - 2\hat{\psi}_p(\boldsymbol{Y}, t). \end{aligned} \tag{4.7}$$

By (3.9), for the choice $t = \infty$, this criterion reduces to

$$\hat{R}_p(\boldsymbol{Y}, \infty) = \sum_{j=1}^{m-p} Y_{l_j}^2 + \sigma^2(2h(m,p) - m) \tag{4.8}$$

where, using (4.4) with the Y_j's replaced by the Z_j's,

$$\begin{aligned} h(m,p) &= E\|P_{L(\boldsymbol{Z},p)}\boldsymbol{Z}\|^2 \\ &= E\sum_{j=1}^{m} Z_j^2 I(|Z_j| > |Z_{l_{m-p}}|) \\ &\quad - E\sum_{i=m-p+1}^{m} Z_{l_i}^2 \\ &= \sum_{i=m-p+1}^{m} \frac{2^m m!}{(i-1)!(m-i)!} \int_0^\infty x^2 \phi(x) \left[\Phi(x) - \frac{1}{2}\right]^{i-1} [1 - \Phi(x)]^{m-i} dx \end{aligned}$$

where ϕ and Φ are the $N(0,1)$-density and distribution functions respectively. The values of $h(m,p)$ can be calculated by numerical integration. Table 1 compares the values of $h(m,p)$ with p and it is evident that $h(m,p)$ rises much faster to m than p does. Keeping in mind that (4.8) is especially appropriate if $\boldsymbol{\mu}$ is close to $\mathbf{0}$ we see that the C_p criterion (3.8) viewed as a risk estimator gives values that may be much too low at least when $\boldsymbol{\mu}$ is close to $\mathbf{0}$. Motivation for taking the choice $t = \infty$ seriously will be given below.

In the case σ^2 unknown but $m < n$ we use the estimator (4.1) in the place of σ^2 in (4.7) or (4.8) and in the distribution of $W_1, \ldots, W_m$. The case σ^2 unknown but $m = n$ falls outside of the scope of this paper.

Table 1. Values of $h(m,p)$ for $m = 10$

p	0	1	2	3	4	5	6	7	8	9	10
$h(m,p)$	0	3.80	5.88	7.40	8.37	9.03	9.47	9.74	9.90	9.98	10.00

Suppose now that we have chosen $\hat{p}$ to minimize (4.7) or its version with σ^2 estimated, in either case truncated at 0. To simplify notation, let $\hat{\boldsymbol{\mu}}$ denote the resulting final NS estimator given by (4.4) with p replaced by $\hat{p}$. Then $\hat{\boldsymbol{\mu}}$ still depends on t and our next aim is to consider the effect of t. We shall use the relative risk of $\hat{\boldsymbol{\mu}}$ given by

$$\rho(\boldsymbol{\mu}/\sigma, t) - E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2/m\sigma^2 \tag{4.9}$$

as a measure in terms of which to judge the effect of t. It is readily seen that this depends only on $\boldsymbol{\mu}/\sigma$ and so it involves no loss of generality to take $\sigma = 1$ until further notice. One way towards choosing t is provided by the minimax approach which chooses t to minimize

$$\sup\{\rho(\boldsymbol{\mu}, t) : \boldsymbol{\mu} \in M\}. \tag{4.10}$$

In a related context Venter and Steel (1994) conjectured that this maximum is reached on the equi-angular line segment $\mu_i = d$ for $i = 1, \ldots, m$ for some $d > 0$. We believe that this is also true here, i.e. that (4.10) is equal to

$$\sup\{\rho(\boldsymbol{\mu}, t) : \mu_i = d \text{ for } i = 1, \ldots, m \text{ with } d > 0\}. \tag{4.11}$$

In view of the complexity of $\hat{\boldsymbol{\mu}}$ it hardly seems possible to prove this; of course (4.11) is at most (4.10). Whether or not this conjecture is in fact true, a possible strategy towards choosing t is to minimize (4.11) and we now describe the results of a numerical study to this effect.

We varied t over the values 0.001 (which serves as a good proxy for 0), 0.1(0.1)3.0 as well as 4, 5, 6 and ∞ and we used $B = 1000$ in the internal simulation calculation of $\hat{\psi}_p(\boldsymbol{Y}, t)$ by (4.6). We further varied the "spacing" d of the line segment in (4.11) over the values 0, 0.5(0.25)2.5 and 3, 4, 5. At each of these d-values we estimated the relative risk (4.9) by the average of 1000 simulation repetitions and then fitted a spline to estimate the functional dependence of the relative risk on d; from this we then determined the position and value of the maximum (4.11). These steps were repeated 5 times and the averages and standard errors were calculated. Double precision IMSL routines were used for these calculations. Figure 1 (Configuration 5) gives an illustration of typical results for the cases $t = 1$, $\sqrt{2}$ and ∞ when $m = 10$ and σ^2 is known. In Table 2 the column headed "Maximal relative risk" provides an abstract of the results of this calculation for a number of t-values for the same case. The "maximal relative risk" decreases as t increases up to about $t = 1$ after which it essentially remains

Table 2. Maximal and minimal relative risks of $\hat{R}_p(\boldsymbol{Y}, t)$ ($m = 10$, σ^2 known).

t	Maximal rel. risk	Standard error	Position d of max.	Standard error	Minimal rel. risk	Standard error
0.001	1.648	0.005	1.880	0.008	0.368	0.008
0.1	1.595	0.005	1.836	0.018	0.360	0.010
0.2	1.491	0.004	1.612	0.022	0.306	0.008
0.3	1.399	0.002	1.470	0.007	0.277	0.005
0.4	1.344	0.003	1.388	0.007	0.207	0.003
0.6	1.296	0.003	1.340	0.006	0.169	0.006
0.8	1.275	0.006	1.309	0.006	0.145	0.004
1.0	1.259	0.006	1.299	0.005	0.122	0.004
$\sqrt{2}$	1.253	0.004	1.288	0.004	0.115	0.005
2.0	1.255	0.003	1.274	0.005	0.104	0.008
3.0	1.257	0.006	1.279	0.003	0.102	0.007
5.0	1.247	0.005	1.271	0.008	0.112	0.004
∞	1.269	0.000	1.253	0.001	0.109	0.001

constant. This indicates that for any choice $t \geq 1$ an estimator close to minimax on the configuration in (4.11) is obtained. This remains true for other choices of m and also for the case where σ^2 is unknown as is evident from Table 3 which covers some of these cases.

The behavior of the relative risk (4.9) in configurations other than the least favorable used in the above calculation are also of interest. Of course $\boldsymbol{\mu} \in L(\boldsymbol{\mu}) = \text{span}\{\boldsymbol{a}_j : \mu_j \neq 0, 1 \leq j \leq m\}$ and if we knew which components of $\boldsymbol{\mu}$ were actually non-zero we would also know $L(\boldsymbol{\mu})$ and then we would estimate $\boldsymbol{\mu}$ by $P_{L(\boldsymbol{\mu})}\boldsymbol{Y}$ (i.e. a non-zero μ_i would be estimated by Y_i and a zero μ_i would be estimated by 0). The relative risk of this "estimator" is

$$E\|P_{L(\boldsymbol{\mu})}\boldsymbol{Y} - \boldsymbol{\mu}\|^2/m = E\|P_{L(\boldsymbol{\mu})}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2/m = p(\boldsymbol{\mu})/m \tag{4.12}$$

where $p(\boldsymbol{\mu}) = \dim(L(\boldsymbol{\mu}))$ is the number of non-zero components of $\boldsymbol{\mu}$. Hopefully the estimator $\hat{\boldsymbol{\mu}}$ will be adaptive in the sense that its relative risk is only slightly larger than $p(\boldsymbol{\mu})/m$. In the configuration of (4.11) we have $L(\boldsymbol{\mu}) = M$ and $p(\boldsymbol{\mu})/m = 1$ and the aim of the above choice of t was to limit as much as possible the amount by which the relative risk of $\hat{\mu}$ exceeds 1 in this case. At the other extreme (4.12) is minimal when $\boldsymbol{\mu} = \boldsymbol{0}$ where $p(\boldsymbol{\mu})/m = 0$ and here the relative risk of $\hat{\mu}$ should be particularly small. We conjecture that this is the most favorable case where $\hat{\boldsymbol{\mu}}$ achieves its minimal risk but again it seems unlikely that this could be proved formally. All our numerical experience to date confirms this conjecture. Tables 2 and 3 also shows values of this "minimal" relative risk as a function of t for various cases. Again it decreases as t increases up to about 1 after which it essentially remains constant. Thus choosing $t \geq 1$ is not only advantageous from a minimax

Table 3. Maximal and minimal relative risks.

m	$n-m$	Criterion	Maximal relative risk	Minimal relative risk
10	∞	$\hat{R}_p(\boldsymbol{Y}, t-0.001)$	1.65	0.366
		$\hat{R}_p(\boldsymbol{Y}, t=1)$	1.26	0.122
		$\hat{R}_p(\boldsymbol{Y}, t=\sqrt{2})$	1.25	0.115
		$\hat{R}_p(\boldsymbol{Y}, t=\infty)$	1.27	0.109
		C_p, GCV, AKA	1.65	0.572
20	∞	$\hat{R}_p(\boldsymbol{Y}, t=0.001)$	1.65	0.264
		$\hat{R}_p(\boldsymbol{Y}, t=1)$	1.18	0.057
		$\hat{R}_p(\boldsymbol{Y}, t=\sqrt{2})$	1.18	0.037
		$\hat{R}_p(\boldsymbol{Y}, t=\infty)$	1.19	0.034
		C_p, GCV, AKA	1.65	0.572
50	∞	$\hat{R}_p(\boldsymbol{Y}, t=0.001)$	1.65	0.146
		$\hat{R}_p(\boldsymbol{Y}, t=1)$	1.11	0.011
		$\hat{R}_p(\boldsymbol{Y}, t=\sqrt{2})$	1.11	0.008
		$\hat{R}_p(\boldsymbol{Y}, t=\infty)$	1.11	0.005
		C_p, GCV, AKA	1.65	0.572
10	20	$\hat{R}_p(\boldsymbol{Y}, t=0.001)$	1.65	0.381
		$\hat{R}_p(\boldsymbol{Y}, t=1)$	1.28	0.194
		$\hat{R}_p(\boldsymbol{Y}, t=\sqrt{2})$	1.27	0.204
		$\hat{R}_p(\boldsymbol{Y}, t=\infty)$	1.27	0.183
		C_p	1.65	0.583
		GCV	1.66	0.634
		AKA	1.49	0.664
20	20	$\hat{R}_p(\boldsymbol{Y}, t=0.001)$	1.64	0.314
		$\hat{R}_p(\boldsymbol{Y}, t=1)$	1.20	0.133
		$\hat{R}_p(\boldsymbol{Y}, t=\sqrt{2})$	1.19	0.110
		$\hat{R}_p(\boldsymbol{Y}, t=\infty)$	1.20	0.115
		C_p	1.65	0.583
		GCV	1.65	0.682
		AKA	1.40	0.727
50	20	$\hat{R}_p(\boldsymbol{Y}, t-0.001)$	1.64	0.239
		$\hat{R}_p(\boldsymbol{Y}, t=1)$	1.14	0.069
		$\hat{R}_p(\boldsymbol{Y}, t=\sqrt{2})$	1.14	0.071
		$\hat{R}_p(\boldsymbol{Y}, t-\infty)$	1.14	0.068
		C_p	1.65	0.583
		GCV	1.62	0.765
		AKA	1.28	0.829

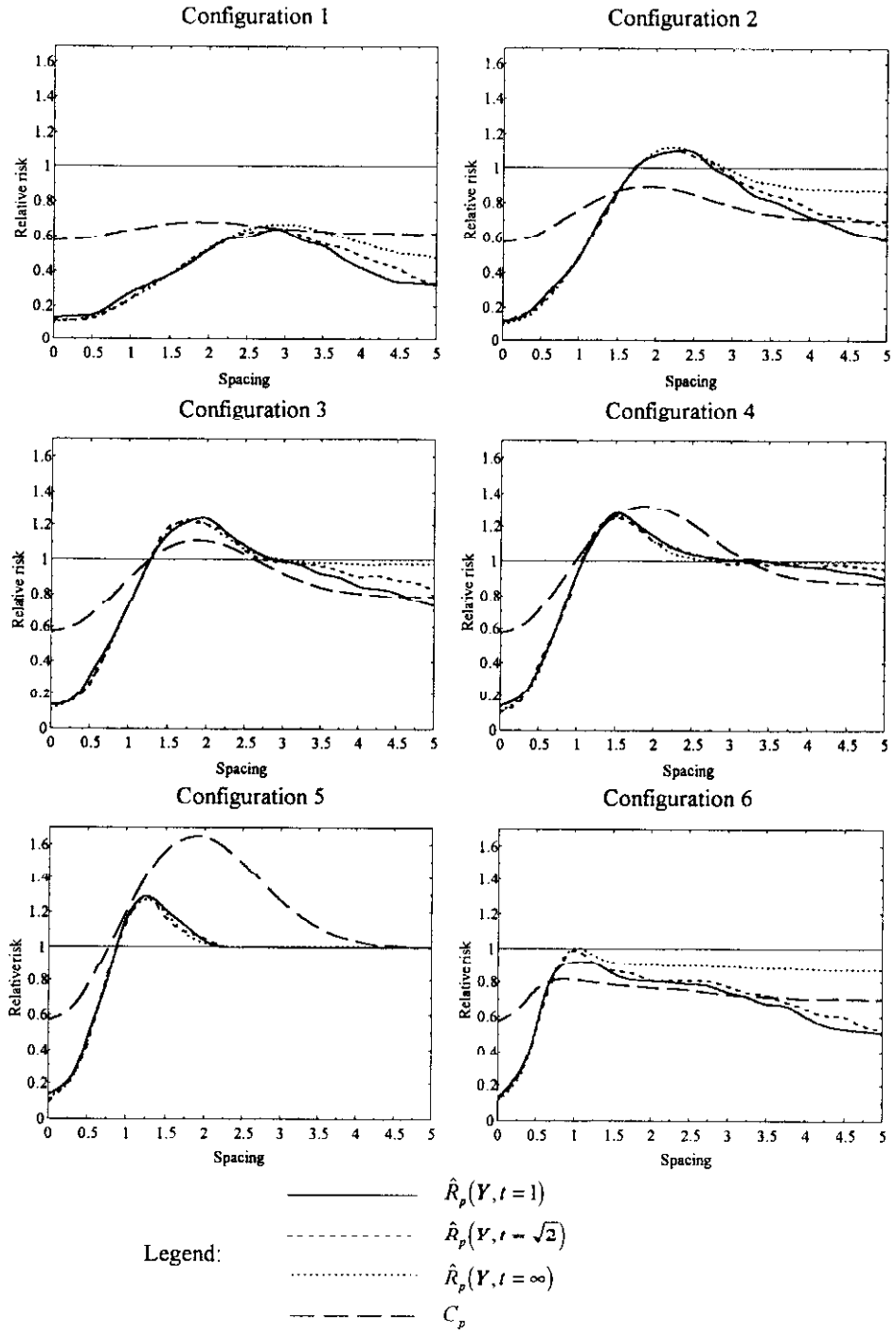

Fig. 1. Relative risks along six configurations.

point of view but it also enhances the performance of $\hat{\boldsymbol{\mu}}$ in its most favorable configuration.

There are a host of further configurations that may be considered. We briefly report on the following six which show some trends the truth of which we have observed also in other cases. Here we have $m = 10$ and we assume $\sigma^2 = 1$ and known. Then we take

Configuration 1: $\mu_1 = d$ and $\mu_i = 0$ for $i = 2, \ldots, 10$;
Configuration 2: $\mu_i = d$ for $i = 1, 2, 3$ and $\mu_i = 0$ for $i = 4, \ldots, 10$;
Configuration 3: $\mu_i = d$ for $i = 1, \ldots, 5$ and $\mu_i = 0$ for $i = 6, \ldots, 10$;
Configuration 4: $\mu_i = d$ for $i = 1, \ldots, 7$ and $\mu_i = 0$ for $i = 8, \ldots, 10$;
Configuration 5: $\mu_i = d$ for $i = 1, \ldots, 10$;
Configuration 6: $\mu_1 = d$, $\mu_2 = 2d$, $\mu_3 = 3d$ and $\mu_i = 0$ for $i = 4, \ldots, 10$.

For each of these configurations we varied d on the grid 0(0.25)5 and estimated the relative risk (4.9) by 1000 simulation repetitions and fitted a smoothing spline to the results. Figure 1 shows the results for the three choices $t = 1$, $\sqrt{2}$ and ∞. When $d = 0$ all these configurations revert to the most favorable case $\boldsymbol{\mu} = \mathbf{0}$ and it is evident that the relative risks are low there. Along configuration 1 the three choices do about equally well as long as the spacing d is small but for larger spacing the parametric bootstrap choice $t = 1$ seems the better option. This tendency is also visible for configurations 2, 3 and 4. Configuration 5 is the least favorable configuration where there is little difference between the choices. Configuration 6 is comparable to 2 in that they have the same number of non-zero μ_i's and again the choice $t = \infty$ shows up as the worst. The choice $t = \infty$ relates to the prior expectation that $\boldsymbol{\mu} = \mathbf{0}$ and it is therefore not surprising that it should turn out poorer when this is not true; what is a bit surprising is that it should be almost as good as the others in the least favorable configuration. We may summarize the findings of this numerical work by saying that, within the class of criteria (3.4) or (4.7), the parametric bootstrap choice $t = 1$ yields a model selection and estimation procedure for the problem of this section which seems overall quite satisfactory in terms of its final relative risk performance. We will refer to (3.4) with $t = 1$ as the "partial bootstrap" criterion for reasons explained in Section 5.

To put this finding into perspective, we also computed the final relative risks associated with $C_p (\equiv FPE_2)$, GCV and AKA as given by (1.1)–(1.3). As is argued in the appendix both GCV and AKA become equivalent to C_p when σ^2 is known (effectively n very large). Table 3 shows the relevant "maximal" and "minimal" relative risks. It is evident that the partial bootstrap criterion performs substantially better in both respects. Indeed, Configuration 5 of Fig. 1 shows that the partial bootstrap completely dominates C_p in very high dimensional configurations and this is also true for very low dimensional configurations as is illustrated by Configuration 1. Between these extremes, however, there are limited regions where C_p performs somewhat better than the partial bootstrap as is illustrated by Configurations 2, 3 and 6. These findings have also been checked for the case σ^2 unknown. On balance, where the partial bootstrap does not overshadow C_p, GCV and AKA, the differences between them are not substantial, which makes the partial bootstrap criterion an appealing option.

5. Concluding remarks

1. Breiman (1992) called (2.4) the "little bootstrap" estimator, "little" presumably refers to the small value of t required for approximate unbiasedness. It is now evident that there is good reason to prefer the value $t = 1$ instead which makes the "little" part of the name seem inappropriate. The idea is to apply bootstrap considerations to one term in (2.1) only, namely the term $E(\boldsymbol{T} - \boldsymbol{\theta})'\boldsymbol{g}(\boldsymbol{T})$ (which is the term that is also prominent in Stein estimation) rather than to all three terms. With $t = 1$ this term is estimated by the ordinary parametric bootstrap. For this reason (2.3) with $t = 1$ and its versions in special problems ((3.4) and (4.7)) are referred to as "partial bootstrap" risk estimators in the sense that only a part of the risk is involved in the bootstrapping.

2. We have carried out simulation studies also in the variable selection problem (Snyman (1994)) and while our findings by and large agree with the extensive results reported by Breiman (1992) on variable selection, they also show that the remarks above are valid in that context as well.

3. Although not required for selection purposes, the user may want to estimate the risk $E\|P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y} - \boldsymbol{\mu}\|^2$ of the final estimator $P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y}$. Since $\hat{p} = \hat{p}(\boldsymbol{Y})$ and $L(\boldsymbol{Y}, \hat{p}(\boldsymbol{Y}))$ are both homogeneous of degree 0, $\boldsymbol{\delta}(\boldsymbol{Y}) = P_{L(\boldsymbol{Y},\hat{p})}\boldsymbol{Y}$ is homogeneous of degree 1 so that the basic method of Section 2 may again be applied to get such an estimator. A simulation within a simulation calculation is required making it computationally demanding.

4. In Sections 1, 3 and 4 the discussion mainly revolved around least squares estimators of the form $P_L\boldsymbol{Y}$. Here $L = L(\boldsymbol{Y})$ may depend on $\boldsymbol{Y}$ with $L(\boldsymbol{Y})$ homogeneous of degree 0 in order to ensure that $P_L\boldsymbol{Y}$ is homogeneous of degree 1. The risk estimation method of Section 2 then applies. More generally the method of Section 2 will still apply to families of estimators of the form $\boldsymbol{\delta}(\boldsymbol{Y}) = H_{L,a}\boldsymbol{Y}$ where a indexes the family $H_{L,a}$ which need not be projection matrices. Here $L = L(\boldsymbol{Y})$ and/or $a = a(\boldsymbol{Y})$ may depend on $\boldsymbol{Y}$ as long as they are both homogeneous of degree 0. An example is the ridge-family for which $H_{L,a} = \boldsymbol{X}_L(\boldsymbol{X}_L'\boldsymbol{X}_L + a\boldsymbol{I})^{-1}\boldsymbol{X}_L'$ where $\boldsymbol{X}_L$ is the matrix of columns corresponding to L of the given design matrix $\boldsymbol{X}$ of a regression variable selection problem. Further research is required to evaluate such generalizations.

5. The partial bootstrap selection criterion resulting from this work is an attractive alternative to criteria such as C_p and it is to be hoped that it will soon find its way into statistical packages.

Acknowledgements

This research was supported by grants from the FRD of South Africa.

Appendix

Motivation for the criterion (1.3)

If $\boldsymbol{Y}^* = \boldsymbol{\mu} + \boldsymbol{e}^*$ is a future observation on $\boldsymbol{Y}$, independent of $\boldsymbol{Y}$, and $P_L\boldsymbol{Y}$ is used as a predictor of $\boldsymbol{Y}^*$, then its prediction risk is

$$E\|P_L\boldsymbol{Y} - \boldsymbol{Y}^*\|^2 = n\sigma^2 + E\|P_L\boldsymbol{Y} - \boldsymbol{\mu}\|^2 = \sigma^2(n + \dim(L)) + \|P_{M|L}\boldsymbol{\mu}\|^2.$$

We also have

$$E(AKA) = E\frac{n+\dim(L)}{n-\dim(L)}\|P_{L^\perp}\boldsymbol{Y}\|^2 = \sigma^2(n+\dim(L)) + \frac{n+\dim(L)}{n-\dim(L)}\|P_{M|L}\boldsymbol{\mu}\|^2$$
$$= E\|P_L\boldsymbol{Y} - \boldsymbol{Y}^*\|^2 + \frac{2\dim(L)}{n-\dim(L)}\|P_{M|L}\boldsymbol{\mu}\|^2.$$

Hence, if $\boldsymbol{\mu} \in L$ (i.e. L is true) then $\|P_{M|L}\boldsymbol{\mu}\| = 0$ and AKA estimates $E\|P_L\boldsymbol{Y} - \boldsymbol{Y}^*\|^2$ unbiasedly. If $\boldsymbol{\mu} \notin L$ (so that L is not true), then $\|P_{M|L}\boldsymbol{\mu}\| > 0$ and AKA is an upwardly biased estimator of $E\|P_L\boldsymbol{Y} - \boldsymbol{Y}^*\|^2$. Hence, if we select L to minimize AKA we tend to avoid selection of an untrue L.

Equivalence of C_p, GCV and AKA when σ^2 is known

We argue that both GCV and AKA becomes equivalent to C_p if $n \to \infty$ while m remains fixed in the context of Section 4. We have $n^2GCV = \|P_{L^\perp}\boldsymbol{Y}\|^2(1 + 2n^{-1}\dim(L) + o(n^{-1}))$. With $L = \text{span}\{\boldsymbol{a}_j, j \in J\}$ we have $n^{-1}\|P_{L^\perp}\boldsymbol{Y}\|^2 = n^{-1}\sum_{j\notin J} Y_j^2 + n^{-1}\sum_{j>m} Y_j^2$. The first term here is $\le n^{-1}\sum_{j=1}^m Y_j^2 \to 0$ a.s. and the second term tends to σ^2 as $n \to \infty$. Consequently n^2GCV becomes $\|P_{L^\perp}\boldsymbol{Y}\|^2 + 2\dim(L)\sigma^2$ as n becomes large which is FPE_2 or C_p. The argument for AKA is similar.

REFERENCES

Akaike, L. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, **22**, 203–217.

Akaike, L. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), 267–281, Akademia Kiado, Budapest.

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

Box, G. E. P. and Meyer, R. D. (1986). An analysis of unreplicated fractional factorials, *Technometrics*, **28**, 11–18.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: $\boldsymbol{X}$-fixed prediction error, *J. Amer. Statist. Assoc.*, **87**, 738–754.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.*, **31**, 377–403.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two level experiments, *Technometrics*, **1**, 311–341.

Dong, F. (1993). On the identification of active contrasts in unreplicated fractional factorials, *Statistica Sinica*, **3**, 209–217.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81**, 425–455.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Snyman, J. L. J. (1994). Model selection and estimation in multiple linear regression, Ph.D. Thesis, Department of Statistics, Potchefstroom University.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution, *Ann. Statist.*, **9**, 1135–1151.

Venter, J. H. and Snyman, J. L. J. (1995). A note on the generalised cross-validation criterion in linear model selection, *Biometrika*, **82**, 215–219.

Venter, J. H. and Steel, S. J. (1990). Estimating risk reduction in Stein estimation, *Canad. J. Statist.*, **18**, 221–232.

Venter, J. H. and Steel, S. J. (1992). Some contributions to selection and estimation in the normal linear model, *Ann. Inst. Statist. Math.*, **44**, 281–297.

Venter, J. H. and Steel, S. J. (1994). Pre-test type estimators for selection of simple normal models, *J. Statist. Comput. Simulation*, **51**, 31–48.

Venter, J. H. and Steel, S. J. (1996). A hypothesis testing approach towards identifying active contrasts, *Technometrics*, **38**, 161–169.