STAGEWISE CLASSIFICATION ALGORITHMS FOR SELECTING A SUBSET OF GROUPS OF ALLOCATION VARIABLES

BU-871-M May, 1989

J. C. Evans
N S W Agriculture and Fisheries, Haymarket 2000, Australia

and

S. J. Schwager
Biometrics Unit, Cornell University, Ithaca, N.Y. 14853, U. S. A.

SUMMARY

Two stagewise classification algorithms are given, one with Type I error control and one without. They use a test of additional classification accuracy at each stage to decide which groups of variables to add or drop, if any. The standardized difference in estimated Bayes risk between two subsets of groups of allocation variables is the test statistic used. For a multinormal example, the algorithms are compared by the estimated Bayes risks of their ultimately selected subsets. Stepwise and simultaneous stepdown classifications do not perform as well as minimal-best-subset classification. To improve the optimality of a subset selected by a stagewise classification, it is necessary to append an extra test of accuracy of the selected subset versus the full set of groups.

Key Words and Phrases: Bayes classification accuracy; Estimated Bayes risk; Multivariate repeated measurements; Stepwise classification; Stepdown classification; Remote sensing; Stagewise subset selection.

Running Title: Stagewise Selection of Allocation Variables

1. Introduction

Evans and Schwager (1989) described two all-possible-subsets approaches for selecting subsets of groups of allocation variables. One approach used a discrimination criterion, Wilks's lambda, and the other used a classification accuracy criterion, the standardized estimate of Bayes risk. For each type, two algorithms were given. The first was the simultaneous procedure advocated by McKay and Campbell (1982a,b), which leads to "adequate" subsets. The second leads to a "minimal-best" subset. In the case of discrimination, Rao's test of additional discrimination (1973, p.556) is used to compare subsets. Adequate subsets are those that give essentially the same discrimination as by all groups. The minimal-best subset is the smallest of the best (=minimal lambda) subsets of each size that retains most of the discrimination from all groups. For classification, a test of additional allocation accuracy is used. Adequate subsets are those that give essentially the same accuracy as the overall best (=minimal standardized estimate of Bayes risk) subset rather than by all groups because the latter is not necessarily the best. The minimal-best-subset classification seeks to select the smallest of the best subsets of each size that retains most of the allocation accuracy.

Grouping of variables was used both to exploit the natural grouping that often occurs and to reduce computational burden through having to consider only combinations of groups instead of individual variables. Nevertheless, grouping includes the special case of one variable per group. Natural grouping occurs in remote sensing studies, such as the detailed example given in Section 5, where reflectance variables for several wavelengths are recorded simultaneously at each date. Once a subset of dates has been selected to give a sufficiently accurate and early classifi-

cation, the selection of a subset of wavelengths could be considered for those dates.

However, even with grouping, all-possible-subsets approaches may still become computationally prohibitive. Roughly speaking, the commonly accepted limit of about 20 variables in all-possible-subsets discrimination suggests an upper limit of 20 groups of variables for all-possible-subsets-of-groups discrimination. But this depends on the number of classes and observations as well as the type of computer and whether or not it is dedicated. When computing does become prohibitive, stagewise selection of groups should be considered, along with the inherently greater risk that the chosen subset will be suboptimal. Stagewise discrimination methods include the simultaneous stepdown algorithm of Mudholkar and Subbaiah (1980) and the extension of the stepwise algorithm of Jennrich (1977) to groups. These algorithms use Rao's test at each step to decide which group, if any, to add or drop. But the selected subset cannot be guaranteed to be the best subset of its size for discrimination, let alone for classification. Instead, classification analogues to stagewise discrimination methods should be used to ensure that the selected subset of groups is as near as possible to optimal for allocation. Such stagewise classification algorithms are simply constructed by replacing Rao's test by a test of additional allocation accuracy. The simultaneous stepdown classification method is equivalent to a backward stepwise classification, but with both a prespecified order of testing groups (by increasing importance or discriminatory ability) and overall probability of a Type I error. It requires a lower individual significance level at each step and so should select fewer redundant groups, but the chosen subset depends on the testing order. As illustrated later, the stepwise and stepdown

classifications give, respectively, about 50% and 80% reductions in CPU time compared with all-possible-subsets classification.

Again, as developed for all-possible-subsets Bayes classification by Evans and Schwager (*loc. cit.*), the most appropriate test statistic for stagewise classification is the standardized difference in estimated Bayes risks between two subsets of groups. They estimated the required Bayes risks and their (co)variances from test data in preference to other methods for the sake of theoretical and computational simplicity due to the independence of test and reference samples. Previous authors had only used the unstandardized difference in estimated Bayes risks, thus ignoring their (co)variances. Even then, only the forward stepwise algorithm of Habbema and Hermans (1977) was generally applicable to two or more multivariate normal or non-normal populations with equal or unequal misclassification costs and prior probabilities. Their stopping rule compared the decrease in estimated Bayes risk, due to adding a variable, with a subjective threshold value. Instead, a statistical test is needed to assess objectively the significance of a decrease.

Section 2 summarizes the test of additional reduction in estimated Bayes risk proposed by Evans and Schwager (*loc. cit.*). Sections 3 and 4 give the stepwise and simultaneous stepdown classification algorithms that use this test statistic both as a selection criterion and in a stopping rule. These approaches satisfy the major requirement of McKay and Campbell (1982b) that "Stopping rules based on probabilistic arguments would be particularly valuable; it is surely desirable to be able to claim that one subset provides an allocation error rate that is significantly lower than that provided by another," but not their other need that there be a "check on the performance of a subset against the performance of the full set of

variables." An extra test of additional accuracy could be appended to the stagewise algorithms to compare the selected subset with the full set of groups. Of course, the full set of groups may not be the subset that gives the greatest accuracy, but the latter is unknown throughout the stagewise process and thus cannot be used. If the selected subset did give a significantly lower accuracy than the full set, then the next larger subset would be selected and the check repeated, and so on. When there are insufficient reference sampling units to develop a classification rule based on all groups of variables, the stagewise selected subset could be checked against the best subset of the same (or a given) size. Section 5 presents a remote sensing example in which the stagewise methods of Sections 3 and 4 are implemented and compared by the standardized estimated Bayes risk of their terminal subsets.

2. A Test of Additional Classification Accuracy

To construct a Bayes classification rule, parameters are estimated from r_k reference sampling units of each class $k \in \{1,\cdots,K\}$, with all $N = \sum_{k=1}^{K} r_k$ sampling units being sampled independently. Such a rule could be based on all G available groups of variables in $\underline{Y} \equiv (\underline{Y}_1^t,\cdots,\underline{Y}_G^t)^t$ or on a subset $\underline{Y}_{(v)}$ of $v \leq G$ groups, where $\underline{Y}_{(v)}$ is the concatenation of $\underline{Y}_{g_1},\cdots,\underline{Y}_{g_v}$. In the simplest case of multivariate repeated measurements, to be considered here, each group corresponds to one date or time and is taken to be of size D, although all methods given here are directly applicable to the general case. To assess a Bayes rule based on an arbitrary subset of groups, the estimation and standardization of Bayes risk are based on the classification of $M \equiv \sum_{k=1}^{K} m_k$ independent test (or holdout) sampling units. The test of additional reduction in Bayes risk due to adding a second subset of u groups to a first subset of v groups is now given.

First, using the first subset of v groups, perform a Bayes classification of the M test sampling units. That is, following the simplest and most usual rule (e.g., Mardia, Kent, and Bibby 1979, p.308), classify each observation $\underset{\sim}{y}_{(v)}$ as coming from the class $k\varepsilon\{1,\cdots,K\}$ that minimizes $\sum_{j=1}^{K} \pi_j C_{kj} \hat{f}(\underset{\sim}{y}_{(v)}|j)$, where π_j is the prior probability that the observation came from class j, C_{kj} is the cost of misclassification of a sampling unit from class j into class k, $C_{kk} = 0$ for every k, and $\hat{f}(\underset{\sim}{y}_{(v)}|j)$ is the estimated probability density function (p.d.f.) for $\underset{\sim}{y}_{(v)}$ from class j. Obtain the estimate of Bayes risk, $R_v \equiv R(g_1,\cdots,g_v)$, where $g_1,\cdots,g_v\varepsilon\{1,\cdots G\}$ identify the v groups involved, using

$$\hat{R}_v = \sum_{j=1}^{K} \pi_j \sum_{k=1}^{K} C_{kj} p_v(k|j) \ , \qquad (1)$$

where $p_v(k|j)$ is the proportion of the m_j class j test observations misclassified into class $k \neq j$. Using the K independent test samples, from which corresponding estimated misclassification probabilities necessarily have zero covariance, the variance of $\hat{R}_v$ is

$$V(\hat{R}_v) = \sum_{j=1}^{K} \pi_j^2 \left\{ \sum_{k=1}^{K} C_{kj}^2 \sigma^2_{kj(v)}/m_j + 2 \sum_{1\leq k<\ell\leq K} C_{kj} C_{\ell j} \sigma_{kj(v),\ell j(v)}/m_j \right\} , \qquad (2)$$

where $\sigma^2_{kj(v)}/m_j$ is the variance of $p_v(k|j) \equiv \bar{q}_{kj(v)}$, the average for class j test observations of values of a variable $q_{k(v)}$ that takes either the value 1 if an observation is classified into class k or the value 0 otherwise; similarly, $\sigma_{kj(v),\ell j(v)}/m_j$ is the covariance of $p_v(k|j)$ and $p_v(\ell|j)$. To obtain an estimate $\hat{V}(\hat{R}_v)$ of $V(\hat{R}_v)$, replace $\sigma^2_{kj(v)}$ and $\sigma_{kj(v),\ell j(v)}$, respectively, by $s^2_{kj(v)}$ and $s_{kj(v),\ell j(v)}$, the usual unbiased sample variance of $q_{k(v)}$ and sample covariance of $q_{k(v)}$ with $q_{\ell(v)}$ taken over class j sampling units. [For all-possible-subsets classification, $z_v \equiv \hat{R}_v/[\hat{V}(\hat{R}_v)]^{\frac{1}{2}}$ was used as an accuracy criterion for empirically comparing a subset of v groups with other subsets: the smaller the z_v, the higher the accuracy.]

Second, using the second subset of u groups together with the first v groups, reclassify the test data and obtain $\hat{R}_{u+v}$ and $\hat{V}(\hat{R}_{u+v})$.

Third, calculate the decrease in estimated Bayes risk

$$\hat{R}_{u\cdot v} \equiv \hat{R}_v - \hat{R}_{u+v} \tag{3}$$

attributable to adding the u groups. $\hat{R}_{u\cdot v}$ is expected to be positive under the alternative hypothesis $H_1: R_v > R_{u+v}$ but zero (or negative) under the null hypothesis $H_0: R_v = R_{u+v}$. As $\hat{R}_{u\cdot v}$ is a simple difference between two asymptotically normal quantities (as $m_k \to \infty$, $k=1,\cdots,K$), it is itself distributed asymptotically normally with mean $E(\hat{R}_{u\cdot v}) = R_v - R_{u+v} \equiv R_{u\cdot v}$ and variance

$$V(\hat{R}_{u\cdot v}) = V(\hat{R}_v) + V(\hat{R}_{u+v}) - 2\,\mathrm{Cov}(\hat{R}_v, \hat{R}_{u+v}) \ . \tag{4}$$

To obtain an estimate $\hat{V}(\hat{R}_{u\cdot v})$ of $V(\hat{R}_{u\cdot v})$, we first need to define $\mathrm{Cov}(\hat{R}_v, \hat{R}_{u+v})$ and estimate it. Then, if it is assumed for small m_k $(k=1,\cdots,K)$ that $\hat{R}_{u\cdot v}$ is approximately normal, the distribution of $\hat{R}_{u\cdot v}$ is fully specified by $E(\hat{R}_{u\cdot v})$ and $V(\hat{R}_{u\cdot v})$, thus facilitating the definition of an appropriate statistic for testing the additional accuracy due to the u groups,

$$z_{u\cdot v} \equiv \hat{R}_{u\cdot v} / [\hat{V}(\hat{R}_{u\cdot v})]^{\frac{1}{2}} \ . \tag{5}$$

Evans and Schwager (*loc.cit.*) gave an example, based on 100 bootstrap samples, that supported this assumption of normality for the case of small m_k $(2 \le m_k \le 19)$.

The covariance of $\hat{R}_v$ and $\hat{R}_{u+v}$ is

$$\mathrm{Cov}(\hat{R}_v, \hat{R}_{u+v}) = \sum_{j=1}^{K} \pi_j^2 \sum_{k=1}^{K} \sum_{\ell=1}^{K} C_{kj} C_{\ell j} \sigma_{kj(v),\ell j(u+v)} / m_j \tag{6}$$

where $\sigma_{kj(v),\ell j(u+v)}/m_j$ is the covariance between $p_v(k|j)$ and $p_{u+v}(\ell|j)$.

An unbiased estimator, $\hat{\mathrm{Cov}}(\hat{R}_v, \hat{R}_{u+v})$, is given by replacing each $\sigma_{kj(v),\ell j(u+v)}$ by the usual unbiased sample covariance, $s_{kj(v),\ell j(u+v)}$, of $q_{k(v)}$ and $q_{\ell(u+v)}$ over the m_j test observations in class j.

Under the null hypothesis H_0: $R_{u+v} = R_v$, the standardized difference between $\hat{R}_v$ and $\hat{R}_{u+v}$, namely $z_{u\cdot v}$, has an asymptotically N(0,1) distribution. Invoking the assumption for small m_k of approximate normality of $\hat{R}_{u\cdot v}$, and thus an assumption that $z_{u\cdot v}$ is distributed approximately as N(0,1) under H_0, a one-sided test of H_0: $R_{u+v} = R_v$ versus the alternative H_1: $R_{u+v} < R_v$ can be performed by comparing $z_{u\cdot v}$ with the upper $100(1-\alpha)\%$ point Z_α of the N(0,1) distribution. If $z_{u\cdot v} > Z_\alpha$ then reject H_0 at significance level α and state that the u groups have increased accuracy; otherwise accept H_0. If all m_k and thus $M = \sum_{k=1}^{K} m_k$ are very small, it may be preferable to compare $z_{u\cdot v}$ with the upper $100(1-\alpha)\%$ point of the t-distribution based on M-K or M-1 d.f.

When it is necessary to identify the groups involved, $z_{u\cdot v}$ can be replaced by

$$z_{u\cdot v} = z(g_{v+1},\cdots,g_{v+u}|g_1,\cdots,g_v) \ . \tag{7}$$

3. Stepwise Classification

Step 0. Calculate $\hat{R}(g_1)$ for each group $g_1=1,\cdots,G$. Find the group g_1 with minimum $\hat{R}(g_1)$ and calculate

$$z(g_1|0) \equiv \hat{R}(g_1|0)/\{\hat{V}[\hat{R}(g_1|0)]\}^{\frac{1}{2}}, \tag{8}$$

where zero denotes random allocation due to the null group g_0 (see Appendix). Enter group g_1 if $z(g_1|0) > Z_\alpha$; otherwise stop.

Step 1. After selecting v groups ($1 \le v \le G-1$), calculate for each of the remaining $u = G-v$ groups g_k, $k\varepsilon\{v+1,\cdots,G\}$,

$$z(g_k|g_1,\cdots,g_v) \equiv \hat{R}(g_k|g_1,\cdots,g_v)/\{\hat{V}[\hat{R}(g_k|g_1,\cdots,g_v)]\}^{\frac{1}{2}} . \qquad (9)$$

Find the group g_k with maximum $z(g_k|g_1,\cdots,g_v)$ and enter it as the $(v+1)^{\underline{st}}$ group if its $z(g_k|g_1,\cdots,g_v) > Z_\alpha$; otherwise stop.

Step 2. Before considering the addition of a $(v+2)^{\underline{nd}}$ group ($3 \le v+2 \le G$), calculate for each of the currently entered v+1 groups g_k, $k\varepsilon\{1,\cdots,v+1\}$, the quantity $z(g_k|g_1,\cdots,g_v)$ using Equation (9), where $g_1,\cdots,g_v$ identify the other v of the v+1 currently entered groups. Find the group g_k with minimum $z(g_k|g_1,\cdots,g_v)$ and remove it if its $z(g_k|g_1,\cdots,g_v) < Z_\alpha$. Now return to Step 1.

This procedure consists of alternations of the forward Step 1 and backward Step 2, and terminates when no further group can be entered or deleted, or possibly earlier if certain conditions are not met. For example, in the case of multivariate normality, the Bayes classification rule based on v groups requires that $Dv \le \underset{j\varepsilon\{1,\cdots,K\}}{\text{minimum}} (r_j-1)$ to ensure non-singularity of each reference mean squares and products matrix $\underline{S}_{j(v)}$ (or $Dv \le N-K$ for a pooled $\underline{S}_{(v)}$) and thus the existence of its inverse. This rule assigns a sampling unit to the class $k\varepsilon\{1,\cdots,K\}$ that minimizes

$$\sum_{j=1}^{K} \pi_j C_{kj} |\underline{S}_{j(v)}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\underline{y}_{(v)} - \bar{\underline{y}}_{j\cdot(v)})^t \underline{S}_{j(v)}^{-1} (\underline{y}_{(v)} - \bar{\underline{y}}_{j\cdot(v)})\right\} \qquad (10)$$

where π_j and C_{kj} are, respectively, the prior probabilities and misclassification costs defined in Section 2 and $\bar{\underline{y}}_{j\cdot(v)}$ is the class j reference-sample mean vector. Thus, in cases when $DG > \min_j(r_j-1)$ [or $DG > N-K$], the stepwise algorithm is made to terminate at the completion of the last step in which the inverse of $\underline{S}_{j(v)}$ (or $\underline{S}_{(v)}$) still exists, or when no further group can be added or dropped, whichever occurs first.

4. Simultaneous Stepdown Classification

Consider the expansion of

$$\hat{R}(g_1,\cdots,g_G|g_0) = \hat{R}(g_0) - \hat{R}(g_1,\cdots,g_G) \ , \tag{11}$$

where g_0 denotes random allocation with no groups (see Appendix), as a sum of G components:

$$\begin{aligned}\hat{R}(g_1,\cdots,g_G|g_0) &= [\hat{R}(g_0) - \hat{R}(g_1)] + [\hat{R}(g_1) - \hat{R}(g_1,g_2)] \\ &\quad + \cdots + [\hat{R}(g_1,\cdots,g_{G-1}) - \hat{R}(g_1,\cdots,g_G)] \\ &= \hat{R}(g_1|g_0) + \hat{R}(g_2|g_1) + \cdots + \hat{R}(g_G|g_1,\cdots,g_{G-1}). \end{aligned} \tag{12}$$

Now, instead of using $\hat{R}(g_1,\cdots,g_G|g_0)$ to test the null hypothesis H_0 of no reduction in Bayes risk due to groups $g_1,\cdots,g_G$, its components can be tested separately in a stepdown approach to identify which groups, if any, lead to a rejection of H_0. That is, the expansion of $\hat{R}(g_1,\cdots,g_G|g_0)$ corresponds to a decomposition of H_0 into G component hypotheses, $H_0 = \cap_{k=1}^{G} H_{0g_k}$, where H_{0g_k} is the null hypothesis of no reduction in Bayes risk due to group g_k. Under H_0, as $\hat{R}(g_k|g_0,g_1,\cdots,g_{k-1})$ for k=1,$\cdots$,G are distributed independently and approximately normally for small m_j (j=1,$\cdots$,K), the corresponding

$$z(g_k|g_0,g_1,\cdots,g_{k-1}) = \hat{R}(g_k|g_0,g_1,\cdots,g_{k-1})/\{\hat{V}[\hat{R}(g_k|g_0,g_1,\cdots,g_{k-1})]\}^{\frac{1}{2}} \tag{13}$$

are approximately independent. In the stepdown approach, the component hypotheses are tested in the order $H_{0g_G},\cdots,H_{0g_1}$ corresponding to the order of increasing importance from g_G to g_1. For each k=G,$\cdots$,1, $z(g_k|g_0,g_1,\cdots,g_{k-1})$ is compared to Z_{α_k}, where $\alpha_k \tilde{=} P[z(g_k|g_0,g_1,\cdots,g_{k-1}) > Z_{\alpha_k}|H_{0g_k}]$ is a prespecified Type I error probability and Z_{α_k} is the upper

$100(1-\alpha_k)\%$ point of the N(0,1) distribution. Following Calinski and Kaczmarek (1977), an overall Type I error probability of approximately $\alpha \equiv 1 - \prod_{k=1}^{G}(1-\alpha_k)$ is achieved by setting

$$\alpha_k \equiv 1 - (1 - \alpha)^{\left(\sum_{j=1}^{k} \frac{1}{j} \right)^{-1}} \quad \text{for } k=1,\cdots,G \ . \tag{14}$$

They prefer these $\alpha_1 > \cdots > \alpha_G$ to constant α_k because they give higher overall power.

The details of stepdown classification are now given.

Step 1. If $z(g_G|g_1,\cdots,g_{G-1}) > Z_{\alpha_G}$ then reject H_{0g_G} at level α_G and H_0 at level α and stop procedure by selecting all groups. Otherwise drop group g_G and go to Step 2.

Step 2. If $z(g_{G-1}|g_1,\cdots,g_{G-2}) > Z_{\alpha_{G-1}}$ then reject $H_{0g_{G-1}}$ at level α_{G-1} and H_0 at level α and stop procedure by selecting groups $g_{G-1},\cdots,g_1$. Otherwise drop group g_{G-1} and go to Step 3.

And so on.

Step G. If $z(g_1|g_0) > Z_{\alpha_1}$ then reject H_{0g_1} at level α_1 and H_0 at level α and stop procedure by selecting group g_1. Otherwise select no groups and stop.

This algorithm at most requires only the calculation of G test statistics based on $\hat{R}(g_1,\cdots,g_G)$, $\hat{R}(g_1,\cdots,g_{G-1}),\cdots,$ $\hat{R}(g_1)$, and $\hat{R}(g_0)$, and their estimated (co)variances.

If $DG > \min_j(r_j-1)$ [or $DG > N-K$] in the case of multivariate normality, then this stepdown algorithm can only be partially implemented. This could be done by forcing the non-selection of the groups $g_G, g_{G-1}, \cdots$ for which $DG, D(G-1), \cdots > \min_j(r_j-1)$ and testing onwards from the step k at which $Dk \leq \min_j(r_j-1)$ is first satisfied.

5. A Remote Sensing Example

Stepwise and simultaneous stepdown classifications were applied to Landsat data from an agricultural survey in the Hillston area of New South Wales, Australia, in the wheat growing season of 1983 (Dawbin and Evans 1988). Strictly forward and strictly backward selections of dates were used to exemplify the range of possible stepwise classifications. For the simultaneous stepdown selection of dates, two different orders of testing were tried: chronological $(1,\cdots,G)$ and reverse chronological $(G,\cdots,1)$, corresponding to, respectively, a possibly increasing classification accuracy through a cropping season and an increasing earliness of classification.

Fields of fallow and woodlands and uncommon crop classes have been excluded from this study but additional fields of the common crop classes have been included. The K=13 common crop classes consist of 11 combinations of density and sowing date for cereal crops (3 oats, 2 barley, 6 wheat) and two pasture types (native and improved). At G=5 dates (April 14, May 16, August 4, September 21 and October 7), mean observations of D=4 reflectance variables were made on each of the 6, 18, 20, 26, 4, 6, 36, 18, 8, 36, 12, 12, and 38 fields, respectively, that were randomly sampled from the six wheat classes, two barley classes, three oats classes and two

pasture classes. Half of the fields were randomly chosen for reference and the other half used for testing within each class $j=1,\cdots,13$ (giving $N=M=120$). For each class, reference field means were averaged to obtain the class mean vector $\underline{\bar{y}}_j$. and then used to obtain the among-fields covariance matrix $\underline{S}_j$. The $\underline{S}_j$ differed significantly (using the generalized likelihood ratio test in Morrison 1978) among wheat, barley, oats, and pasture crop types ($\alpha=.01$) but were similar within each. Accordingly, a pooled covariance matrix was obtained for each crop type and used in place of the individual matrices.

Although the original Landsat data can take only nonnegative integer values from 0 to 255 and therefore cannot be strictly multivariate normal in distribution, Landgrebe (1980) has demonstrated that they are approximately so. Accordingly, multivariate normality will be assumed here. Unequal prior probabilities and unequal misclassification costs were used in each application of a sample Bayes rule and in the estimation of Bayes risks and their (co)variances. Anticipated relative proportions of classes in the study district were used as the prior probabilities π_k, $k=1,\cdots,13$, and were (as percentages) 2, 8, 13, 24, 2, 2, 4, 2, 2, 6, 2, 9, and 24. Relative misclassification costs were obtained from the district's agronomist and were: $C_{kj}=1$ for $k=1,\cdots,6$ and $j=7,\cdots,11$; 1 for $k=7,8$ and $j=1,\cdots,6,9,10,11$; 1 for $k=9,10,11$ and $j=1,\cdots,8$; 2 for $k=12,13$ and $j=1,\cdots,11$; 2 for $k=1,\cdots,11$ and $j=12,13$; and 0 otherwise.

To enable an assessment of the optimality of subsets of dates selected by stepwise and simultaneous stepdown classifications, the standardized estimate of Bayes risk was found for every subset. The best (=minimal z_v) subset of dates of size 1 was August ($z_1=5.1$); of size 2 was May and

October (z_2=3.0); of size 3 was May, August and September (z_3=2.9); and of size 4 was April, May, August and September (z_4=2.8). For all groups and no groups, respectively, z_5=6.1 and z_0=11.7. Comparison of these $z_0, z_1, \cdots, z_5$ values suggests that the "optimal" subset, i.e., the smallest subset that retains most of the classification accuracy, is May plus October. Both the backward and forward stepwise classifications (with α=0.05 at each step) chose the subset of May plus August (z=3.3) which is near to optimal.

Simultaneous stepdown classification (with an overall α=0.05; α_5=.0045, α_4=.0056, α_3=.0075, α_2=.0112, α_1=.0222) selected only October (z=6.0) under chronological ordering and only April (z=5.4) under reverse chronological ordering; each of these dates is suboptimal and neither is the best single date. Dawbin and Evans (1988) used a different testing order based both on agronomic grounds and on earliness and likely accuracy of classification: October, August, April, September, May. They retained all but October 7, thus selecting the subset with the overall smallest z-value and enabling an earlier classification on September 21. In contrast, the stepwise selected subset of May 16 plus August 4 gave an even earlier classification at the expense of a slight decrease in accuracy. As the agronomist (Dawbin) was mainly concerned with choosing dates to give an early and accurate classification, the subsequent selection of a subset of reflectance variables has not been considered.

A fuller study of the performance of the stepwise and simultaneous stepdown classifications was conducted by applying them to 100 separate bootstrap samples of the test fields, i.e., m_j fields randomly sampled with replacement from the m_j test fields in class j, for all j=1,$\cdots$,13. Table 1 gives for each method the percentages of the 100 selected subsets having

one to five groups; results for the minimal-best-subset method of Evans and Schwager (1989) are also given. The majority of subsets selected by all methods had three groups. Stepdown classification for reverse chronological order selected all five groups in almost one-third of cases but no subsets of four groups, whereas minimal-best-subset classification did not choose all five in any case. The stepdown method for chronological order chose no less than three groups. Table 2 gives for each method the percentage of selected subsets of each size that were not z-best and their corresponding average increase in z-value; results from the all-possible-subsets discrimination of Evans and Schwager are also given, i.e., the percentage of lambda-best subsets that were not z-best and their average inflation of z. None of these methods gives satisfactory results, with the discrimination method and backward stepwise classification giving the worst inflation in z-values for subsets of 3 and 4 groups.

However, whether or not a selected subset is the best of its size, it is more pertinently assessed by its degree of optimality, i.e., by the closeness of its z-value to that of the optimal subset, relative to the differences among $z_0, z_1, \cdots, z_5$ from which the optimal subset is determined. Minimal best classification gave 91% (near-) optimal and 9% suboptimal subsets, with the near- and suboptimal subsets tending to include too few dates. The suboptimal subsets consisted of all single dates and 18% of pairs of dates; in each case, an extra test of the final subset versus the optimal subset would have shown the need to add more dates. Table 3 gives suboptimality results for each stepwise and stepdown classification method. Overall, only the chronological stepdown method gave results comparable to the minimal-best classification, but unlike the other methods, a majority of suboptimal subsets consisted of all dates. As the optimal subset is

unknown in practice when using the stagewise methods, the full set of dates is the only logical subset against which to check the selected subset. Thus, when all dates are selected, as in some of the stepdown classifications, no check of optimality is possible. Apart from subsets containing all dates, other suboptimal subsets tended to include too few groups; however, for each method and in most cases, an extra test of accuracy versus the full set of dates would have suggested the need for more dates and led to the selection of a subset nearer to optimal. Without the extra test, the forward and backward stepwise algorithms and the reverse chronological stepdown method are not acceptable. However, with the appended test, these methods are comparable to the chronological stepdown classification. Therefore, in practice, if resources are unavailable for doing all-possible-subsets classification and instead a stagewise classification is done, then an extra test must be appended to check the adequacy of the selected subset. Even when resources are available and a minimal-best classification is done, an extra test must be done versus the overall-best subset to ensure that enough groups have been retained.

The Interactive Matrix Language (IML) procedure of the SAS package (SAS Institute Inc. 1985) was used for all aspects of this example, including estimation, bootstrapping and implementation of the classification methods. Three different computers were used during the development and running of these programs: an IBM 3081 mainframe (OS VS2/MVS), a Prime 6350 minicomputer (PRIMOS), and a NEC Powermate 2 microcomputer (MSDOS). CPU and IO times were recorded for two sizes of data sets. For 50-100 observations over only 5 classes, for the same groups of variables, on the IBM it took less than 2 mins CPU and 3 secs IO to do all of the minimal-best and other stagewise classifications. For the full data set here,

involving 240 observations over 13 classes, the all-possible-subsets and minimal-best classifications using equal (unequal) covariance matrices took 1.75 (3.75) hours CPU and 0.5 (1.5) mins IO on the Prime. In contrast, stepwise classification took one (two) hours CPU and 0.3 (0.8) mins IO, and stepdown classification used 20 (45) mins CPU and 0.1 (0.3) mins IO. Times on the NEC were about three times those on the Prime.

6. **Conclusions**

The standardized difference in estimated Bayes risk has been proposed for testing additional classification accuracy due to an added subset of allocation variables. Some stagewise algorithms that use this test at each stage have been given here. Based on the example of Section 5, the stepwise and simultaneous stepdown classifications do not perform as well as the minimal-best-subset classification of Evans and Schwager (1989). This was expected because, first, the minimal-best algorithm guarantees that the selected subset will be the best of its size and, second, the algorithm was designed to seek objectively the optimal subset, i.e., the smallest subset that retains most of the accuracy. Still, the minimal-best subset was not always optimal and an extra test versus the overall-best subset was needed to ensure near-optimality. Similarly, to improve the optimality of subsets selected by stepwise and stepdown methods, an extra test is needed to compare the chosen subset with the full set of groups. Nevertheless, these methods should only be used when computational cost saving is crucial, as the selected subset is less likely to be optimal than the minimal-best subset.

As the stepdown classification can lead to different subsets for different test orders, it should be used only when a single order (chronological, reverse, or other) is relevant to the study at hand. However, even when there is a single prespecified order of testing, this simultaneous method may not be applicable. For example, it may be required to perform successive tests of additional classification accuracy as the data for each date become available; then the nonsimultaneous stepwise algorithm, with only one candidate group for entry at each forward step, would instead be appropriate. In those cases where no single order is dictated, a stepwise classification should be used in preference to simultaneous stepdown classification.

REFERENCES

Calinski, T. and Kaczmarek, Z. (1977). A stepdown procedure of eliminating variables in multivariate analysis of variance. *Biometrical Journal* 19, 449-453.

Dawbin, K. W. and Evans, J. C. (1988). Large area crop classification in New South Wales, Australia, using Landsat data. *International Journal of Remote Sensing* 9(2), 295-301.

Evans, J. C. and Schwager, S. J. (1989). The use of estimated Bayes risk as a criterion for selecting groups of allocation variables. Report BU-870-M, Biometrics Unit, Cornell University, Ithaca, N.Y. Submitted for publication.

Habbema, J. D. F. and Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics* 19, 487-493.

Jennrich, R. I. (1977). Stepwise discriminant analysis. In *Statistical Methods for Digital Computers*, Vol. 3 of *Mathematical Methods for Digital Computers,* K. Enslein, A. Ralston, and H. S. Wilf (eds.), 76-95. New York: Wiley.

Landgrebe, D. A. (1980). The development of a spectral-spatial classifier for earth observational data. *Pattern Recognition* 12, 165-175.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* New York: Academic Press.

McKay, R. J. and Campbell, N. A. (1982a). Variable selection techniques in discriminant analysis, I. Description. *British Journal of Mathematical and Statistical Psychology* 35, 1-29.

McKay, R. J. and Campbell, N. A. (1982b). Variable selection techniques in discriminant analysis, II. Allocation. *British Journal of Mathematical and Statistical Psychology* 35, 30-41.

Morrison, D. F. (1978). *Multivariate Statistical Methods*, Second Edition. Singapore: McGraw-Hill.

Mudholkar, G. S. and Subbaiah, P. (1980). A review of step-down procedures for multivariate analysis. In *Multivariate Statistical Analysis,* R. P. Gupta (ed.), 161-178. Amsterdam: North Holland Publishing Company.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications,* 2nd ed. New York: Wiley.

SAS Institute Inc. (1985). *SAS IML User's Guide for Personal Computers, Version 6 Edition.* Cary, North Carolina: SAS Institute Inc.

Appendix: Random Allocation

In the stagewise algorithms of Sections 3 and 4, it is necessary at either the first or last stage to decide whether to enter or retain any groups at all by a comparison of the classification accuracy obtained with that from a purely random allocation of test observations, i.e., based on no groups. Let g_0 denote a null group consisting of no groups. One way to do a random allocation is as follows. Partition the interval (0,1) into K subintervals k=1,$\cdots$,K of lengths equal to the prior probabilities π_k of a sampling unit arising from classes k=1,$\cdots$,K. Randomly generate an observation from the uniform(0,1) distribution. If the observation falls in subinterval k then classify the sampling unit as being from class k. In this way, allocate m_j test observations corresponding to each class j=1,$\cdots$,K and calculate the proportions $p_0(k|j)$, analogous to the earlier $p_v(k|j)$, of these observations misclassified into class k=1,$\cdots$,K. Then these proportions can be substituted into Equation (1) to obtain $\hat{R}_0$; the estimated variance, $\hat{V}(\hat{R}_0)$, of $\hat{R}_0$ can be obtained in the same way as $\hat{V}(\hat{R}_v)$ via Equation (2). For comparability with subsets of v groups, the accuracy due to zero groups is found similarly to z_v, namely $z_0 = \hat{R}_0/[\hat{V}(\hat{R}_0)]^{\frac{1}{2}}$. Decrease in estimated Bayes risk due to u groups over no groups is given by $\hat{R}_{u\cdot 0} \equiv \hat{R}_0 - \hat{R}_u$, an analogue of $\hat{R}_{u\cdot v} = \hat{R}_v - \hat{R}_{u+v}$ but with $\hat{R}_{u+0} \equiv \hat{R}_u$. When it is necessary to identify the groups involved, $\hat{R}_{u\cdot v}$ can be replaced by

$$\hat{R}(g_{v+1},\cdots,g_{v+u}|g_1,\cdots,g_v) = \hat{R}(g_1,\cdots,g_v) - \hat{R}(g_1,\cdots,g_{v+u}) \qquad (A1)$$

where $g_1,\cdots,g_{u+v} \varepsilon \{1,\cdots,G\}$. Similarly when the null group is involved, $\hat{R}_{u\cdot 0}$ can be replaced by $\hat{R}(g_1,\cdots,g_u|g_0)$. To find $z_{u\cdot 0}$, all of the necessary calculations are the same as for $z_{u\cdot v}$ in Equation (5).

Table 1. Percentages of selected subsets with different numbers of groups, for each classification method.

Number of Groups	Forward Stepwise	Reverse Chronological Stepdown	Chronological Stepdown	Backward Stepwise	Minimal Best
1	4	6	0	0	5
2	13	20	0	15	22
3	58	43	64	50	61
4	16	0	19	24	12
5	9	31	17	11	0

Table 2. Percentage of selected subsets of a given size that were not z-best of that size (and their average percentage increase in z-value), for each classification method and for minimal-lambda discrimination.

Number of Groups	Forward Stepwise	Reverse Chronological Stepdown	Chronological Stepdown	Backward Stepwise	Lambda Best
1	0(--)	67(22)	--(--)	--(--)	64(20)
2	62(21)	95(21)	--(--)	73(24)	64(23)
3	79(24)	28(17)	87(22)	72(30)	80(37)
4	56(25)	--(--)	21(17)	71(38)	87(41)

Table 3. Percentage of selected subsets of a given size that were suboptimal (and their average percentage increase in z-value over that for the best subset of the same size), for each classification method.

Number of Groups	Forward Stepwise	Reverse Chronological Stepdown	Chronological Stepdown	Backward Stepwise
1	100(0)	100(15)	--(--)	--(--)
2	54(22)	90(22)	--(--)	67(22)
3	19(36)	5(32)	8(56)	20(53)
4	12(59)	--(--)	0(--)	42(51)
5	0(--)	6(0)	41(0)	0(--)
Overall	24(28)	28(21)	12(24)	30(43)