The Use of Estimated Bayes Risk as a Criterion for Selecting Groups of Allocation Variables

BU-870-M May 1989

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

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

Estimated Bayes risk, standardized, is proposed as a criterion for examining all possible subsets of groups of allocation variables. The subset of a specified size that achieves minimal estimated Bayes risk is defined as the best subset of that size for classification purposes. The standardized difference in estimated Bayes risk between two subsets of groups of allocation variables is proposed as a test statistic for additional classification accuracy. This test is used in a "minimal-best-subset" algorithm that aims to select the smallest subset retaining most of the classification accuracy. A multivariate normal example demonstrates that all-possible-subsets and minimal-best discrimination procedures based on Wilks's lambda and Rao's test do not necessarily identify the best subsets.

Key Words and Phrases: Bayes classification; Allocation variable selection; All possible subsets; Multivariate repeated measurements; Remote sensing.

Running Title: Using Bayes Risk to Select Allocation Variables

1. INTRODUCTION

In classification studies, observations may be made on several types of variables at a number of times for each sampling unit. These multivariate repeated measurements, or sequential multivariate data, usually must be obtained first for each sampling unit in a reference (= design, calibration, index, or training) sample from each class, to estimate parameters for establishing an appropriate classification rule. This rule is then used to allocate each unidentified sampling unit into one of the several classes under consideration.

As the total number of variables can be very large with sequential multivariate data, many variables may not improve classification accuracy in the presence of the others and are thus redundant. It is therefore important to consider methods for selecting subsets of variables that maximize accuracy at each stage of selection and that ultimately select subsets providing essentially the same accuracy as the full (or the most accurate) set of variables. Selecting a subset not only reduces future cost and time for measurement and computing, but can actually decrease misclassification error rates (McKay and Campbell 1982b). When there are many variables, however, the number of combinations of variables to be considered and thus the number of selection steps required may become prohibitively expensive or even infeasible. To reduce the number of steps required with multivariate repeated measurements data, the variables observed at the same time can be regarded as a natural group, and these groups can be selected rather than individual variables. Indeed, this is an essential first step in remote sensing studies wherein reflectance

variables for several wavelengths are recorded simultaneously at each date. Once a subset of dates has been selected to give an adequately accurate and early classification, then the selection of a subset of wavelengths could be considered for the chosen dates separately or jointly. In this way, using grouping to reduce the number of steps of computation, subsets of, say, 20 groups of 4 variables each, giving a total of 80 variables, may be considered for selection rather than just 20 single variables. Then the commonly accepted limit of about 20 variables in all-possible-subsets discrimination (McCabe 1975) may be replaced by an upper limit of about 20 groups of variables for discrimination or classification, although the upper limit of 20 groups may need to be reduced in cases with large numbers of classes, variables per group, and observations. This grouping structure will be adopted throughout this paper; it includes the special case of individual variables when group size is one. Section 2 establishes a notation for statistics based on subsets of groups of variables, for use in later sections.

Most existing subset selection methods use a measure of discrimination among classes as the selection criterion. These discrimination methods select subsets that either maximize class separation at each stage or ultimately retain essentially the same separation as with all groups of variables, or both. But there is no guarantee that the selected subsets will give maximal allocation accuracy (Habbema and Hermans 1977). One such method is the empirical all-possible-subsets discrimination procedure of McCabe (1975); the best subset of each size is defined as the subset with minimal Wilks's lambda. Section 3 of this paper presents an application of McCabe's algorithm to the case of grouped variables (Evans 1984). Evans's minimal-best-subset discrimination method, which uses Rao's test of addi-

tional discrimination due to adding a subset, is also given to select the smallest subset of groups that retains most of the discrimination from all groups.

In their review of variable selection for allocation, McKay and Campbell (1982b) suggested that none of the currently available procedures can be recommended strongly over others. The best of these methods use estimated Bayes risk or a closely related statistic, possibly with equal misclassification costs or prior probabilities or both, as a criterion for the empirical comparison of subsets of (groups of) variables. When costs are equal, the Bayes risk is simply the probability of a misclassification. Bayes risk, as the selection criterion, is the appropriate link between the practical aim of allocating future sampling units and the Bayes classification rule that minimizes Bayes risk for the selected subset. It also provides a natural stopping rule: if the addition of extra groups decreases the risk by more than a given threshold, then that justifies their inclusion. McLachlan (1976, 1980), Costanza and Afifi (1979), and Schaafsma and van Vark (1979) considered such stopping rules for the case of two multivariate normal populations with common covariance matrices, costs, and priors; Henschke and Chen (1974) did the same for two or more populations and unequal priors; van Vark (1976) did likewise for two or more populations with unequal covariance matrices and priors. All except McLachlan first ordered the variables using a stepwise discrimination approach to form progressively larger subsets, then they applied a stopping rule using allocation accuracy to select the best subset. Costanza and Afifi selected the subset yielding the highest estimated overall probability of a correct allocation; Schaafsma and van Vark aimed to minimize the pooled or average probability of misclassification; Henschke and Chen aimed

to select a subset that minimized the estimated Bayes risk; and van Vark sought to select the subset that minimized the average of the estimated posterior probabilities of misclassification of test individuals. McLachlan proffered an allocation criterion that could potentially be used both for the ordering of variables and as a stopping rule. He developed an asymptotic method for obtaining an approximate confidence level corresponding to no increase in the mean of the two estimated misclassification probabilities due to deleting a subset. There was also a close correspondence between this confidence level and the significance probability associated with Rao's test, but that fact does not carry over to the case of more than two multivariate normal populations (Habbema and Hermans 1977). None of these methods is generally applicable because of the assumptions of multivariate normality and equal costs. Habbema and Hermans considered two or more populations, allowed unequal costs, and used nonparametric estimation of probability density functions (p.d.f.s) to avoid the restrictive distributional assumption of normality. They proposed a forward stepwise algorithm that uses estimated Bayes risk both as a selection criterion to be minimized at each step and in a stopping rule. This approach is an excellent one but there still remain some problems. First, the stopping rule empirically compares decrease in risk with a subjective threshold value; a statistical test is needed to assess objectively the significance of a decrease in estimated Bayes risk due to adding a subset. Second, to estimate Bayes risks they use estimates of misclassification probabilities (= error rates) but ignore their variances and covariances. These (co)variances should be used to estimate the variances and covariances of estimated Bayes risks for different subsets, which should then be used in conjunction with the estimated Bayes risks to enhance the comparability of

subsets. As algorithms incorporating these needs will be computationally extensive, it is important to choose estimators of misclassification probabilities and their (co)variances that do not unduly increase the burden. Although jackknife, bootstrap, and other estimators may be preferred to resubstitution and error count (or holdout) estimators for a given subset of variables (Hand 1986), the need for additional intensive computation, such as in the leave-one-out approach, is magnified when many different subsets must be studied.

To overcome these problems, Section 4 proposes the all-possible-subsets and minimal-best-subset Bayes classification procedures of Evans (1984). The best subset of each size is defined to be the subset that achieves maximal accuracy in the classification of test (= follow-up or holdout) sampling units. The selection criterion used is estimated Bayes risk, standardized for comparability of subsets; minimal risk represents maximal classification accuracy. Then the selected subset of a given size, based on the available reference and test sampling units, is guaranteed to be the best subset of that size for Bayes classification purposes, unlike the subsets selected with a discrimination criterion. From among the best subsets of all sizes, the minimal-best classification algorithm aims to select the smallest subset that retains most of the classification accuracy of the overall best subset of groups. This algorithm uses a test of additional accuracy due to adding or replacing a subset, based on the standardized difference in estimated Bayes risk between two subsets. The error count estimator has been chosen in preference to the resampling and other methods in Hand (1986) for the sake of simplicity in derivation and computation of error rates and their (co)variances due to the independence of the test sample from the reference sample. Offset against this benefit

of independence is a reduction in the number of sampling units available for designing the classifier. If there is a limited number of extra units available for testing then a suitably smaller percentage of the total number should be assigned to the test sample than the reference sample. Resubstitution was not used because of the underestimation of error rates when the reference sample is used first to build a classification rule and then to assess it. Section 4 includes a description of the usual Bayes classification rule and the estimation, standardization, and use of Bayes risks and their differences as criteria for selecting subsets of groups of allocation variables. Section 5 gives a remote sensing example illustrating that the all-possible-subsets and minimal-best-subset discrimination procedures do not necessarily select the best subsets for Bayes classification purposes.

2. NOTATION FOR GROUPED VARIABLES

Let Y_{gd} denote variable d $(1 \le d \le D)$ of group g $(1 \le g \le G)$; $\underline{Y}_g = (Y_{g1},\cdots,Y_{gD})^t$ denotes the vector of D variables in group g; $\underline{Y} = (\underline{Y}_1^t,\cdots,\underline{Y}_G^t)^t$ denotes the vector of all GD variables. (The superscript t indicates transposition.) Let $\underline{Y}_{(u)}$ denote the subvector of $\underline{Y}$ corresponding to an arbitrary set of $u \le G$ groups $g_1,\cdots,g_u$, so $\underline{Y}_{(u)}$ is the concatenation of $\underline{Y}_{g1},\cdots,\underline{Y}_{gu}$. The notation $\underline{Y}_{(u)}$, which indicates only the size of a set of groups and not the u specific groups in it, will be used for the sake of simplicity wherever possible.

This paper gives details for the common case of multivariate repeated measurements just described, with G groups of D variables. The variables $Y_{1d},\cdots,Y_{Gd}$ are G repeated measurements on variable d, one of the D different variable types. However, all methods presented here are directly

applicable to the more general case in which group g has D_g variables, g=1,$\cdots$,G. This includes the special cases of (i) unequal numbers of repeated measurements on the D different types of variables and (ii) GD variables grouped by a means other than repeated measurements, such as by variable types. In the repeated measurements cases, as all variables are observed simultaneously, or essentially so, on a sampling unit at each of the G times, the GD variables fall naturally into G groups of variables. Then selecting a subset of the times can substantially reduce future measurement costs, and if the subset does not contain the later time(s), earlier classifications can be made.

The classification methods of later sections are applicable to distributions other than the multivariate normal, but for simplicity the latter is assumed, with different mean vectors and common or distinct covariance matrices for all classes. In the classification problem, $\underline{Y}$ is observed on a sampling unit from an unknown class $k\varepsilon\{1,\cdots,K\}$; on the basis of the observation $\underline{y}$ the unit is classified as being from class $j\varepsilon\{1,\cdots,K\}$. Before any unidentified sampling units can be classified, the parameters of a classification rule usually must be estimated from observations on $N \equiv \sum_{k=1}^{K} r_k$ independent calibration sampling units, r_k being from class k, and then the rule should be tested on $M \equiv \sum_{k=1}^{K} m_k$ independent test sampling units, m_k being from class k.

Let $\underline{y}_{ki}$ denote the length GD observation vector (G groups of D variables) for a sampling unit $i\varepsilon\{1,\cdots,r_k,r_k+1,\cdots,r_k+m_k\}$ randomly selected from class $k\varepsilon\{1,\cdots,K\}$. In the case of multivariate normality, the class k mean vector $\underline{\mu}_k$ and covariance matrix $\underline{\Sigma}_k$ (or common $\underline{\Sigma}$) are usually unknown and estimated by the reference sample mean vector $\bar{\underline{y}}_{k\cdot} \equiv \sum_{i=1}^{r_k} \underline{y}_{ki}/r_k$ and within-class mean squares and products (MSP) matrix $\underline{S}_k \equiv \underline{W}_k/(r_k-1)$, where $\underline{W}_k \equiv \sum_{i=1}^{r_k} (\underline{y}_{ki}-\bar{\underline{y}}_{k\cdot})(\underline{y}_{ki}-\bar{\underline{y}}_{k\cdot})^t$ is the within-class-k sums of squares and

products (SSP) matrix (or by $\underline{S} \equiv \underline{W}/(N-K)$, where $\underline{W} \equiv \sum_{k=1}^{K} \underline{W}_k$ is the pooled within-classes SSP matrix).

In the later sections, sampling units also must be classified by an arbitrary subvector $\underline{y}_{ki(u)}$ of $\underline{y}_{ki}$. Using the earlier notation of this section, such classification rules are established and tested using the reference and test observations $\underline{y}_{ki(u)}$, the length uD vector of observations of variables in groups g_1 to g_u, D variables per group, on sampling unit $i\varepsilon\{1,\cdots,r_k+m_k\}$ from class $k\varepsilon\{1,\cdots,K\}$. In the multivariate normal case, the mean vectors and covariance matrices based on u groups are simply uD × 1 subvectors and uD × uD submatrices, respectively, of those based on all groups. Subvectors of $\underline{\mu}_k$ and $\bar{\underline{y}}_{k\cdot}$ are denoted by $\underline{\mu}_{k(u)}$ and $\bar{\underline{y}}_{k\cdot(u)}$. Submatrices of $\underline{\Sigma}_k$, $\underline{\Sigma}$, $\underline{S}_k$, $\underline{S}$, $\underline{W}_k$, and $\underline{W}$ are denoted by $\underline{\Sigma}_{k(u)}$, $\underline{\Sigma}_{(u)}$, $\underline{S}_{k(u)}$, $\underline{S}_{(u)}$, $\underline{W}_{k(u)}$, and $\underline{W}_{(u)}$.

3. ALL-POSSIBLE-SUBSETS AND MINIMAL-BEST-SUBSET DISCRIMINATION

The usual selection criterion in discrimination methods is Wilks's lambda, as in the all-possible-subsets procedure proposed by McCabe (1975) for selecting subsets of single variables. His procedure is extended here to the case of grouped variables.

Wilks's lambda is first calculated for every subset of groups of size u=1,···,G. Based on u groups $g_1,\cdots g_u$, Wilks's lambda is denoted by Λ_u or $\Lambda(g_1,\cdots,g_u)$ and defined as the determinantal ratio

$$\Lambda_u \equiv |\underline{W}_{(u)}|/|\underline{W}_{(u)} + \underline{B}_{(u)}| \qquad (1)$$

where $\bar{\underline{y}}_{\cdot\cdot} \equiv \sum_{k=1}^{K}\sum_{i=1}^{r_k} \underline{y}_{ki}/N$ is the sample grand mean and $\underline{B}_{(u)}$ is a uD × uD submatrix of $\underline{B} \equiv \sum_{k=1}^{K} r_k(\bar{\underline{y}}_{k\cdot} - \bar{\underline{y}}_{\cdot\cdot})(\bar{\underline{y}}_{k\cdot} - \bar{\underline{y}}_{\cdot\cdot})^t$, the among-classes SSP matrix based on K-1 d.f. The smaller the Λ_u, the greater the discrimination among classes.

Next, identify the best (= smallest lambda) subset of size 1, of size 2, and so on. That is, find the subset of groups $g_1,\cdots,g_u$ that achieves

$$\underset{g_1,\cdots,g_u \in \{1,\cdots,G\}}{\text{minimum}} \Lambda(g_1,\cdots,g_u) \tag{2}$$

for each u=1,$\cdots$,G-1. McKay and Campbell (1982a) advocated a probabilistic alternative to this empirical comparison of all possible subsets. Their method "employs significance testing with protection of the simultaneous significance level for all (Rao's) tests of additional information carried out" in isolating a number of adequate subsets that give essentially the same discrimination as the original set of variables. The choice of which particular "best" or "adequate" subset to use for discrimination or future allocation purposes is then a compromise between ease or cheapness of measurement and accuracy and earliness of classification.

Evans (1984) proposed a probabilistic "minimal-best-subset" method for selecting a single subset from among the best subsets. The groups not included in the best subsets of size G-1,G-2,$\cdots$,1, and of size 0, are tested successively for their discrimination additional to that of the included groups (using Rao's test of each best subset versus the full set); or until rejection of one of the null hypotheses of no additional discrimination. If and when such a rejection occurs, select the groups included in the best subset at the previous step; otherwise select no groups. The final subset from this non-simultaneous test procedure is taken as the minimal-best subset. If GD>N-K, these algorithms cannot be applied, as comparisons then cannot be made against the full set of groups. Indeed, in the all-possible-subsets discrimination, the most groups that can be considered is the largest u for which uD $\leq$ N-K, thus giving a nonsingular $\underset{-u}{W}$.

A serious deficiency of these and other discrimination methods that use Wilks's lambda as the selection criterion is that they assume a common $\underset{\sim}{\Sigma}_{(u)}$ for all classes. Only upon the ultimate use of a selected subset in a Bayes classification rule can unequal $\underset{\sim}{\Sigma}_{k(u)}$ be utilized, an apparent inconsistency with the selection process. Discrimination methods also ignore possibly unequal costs of misclassification and prior probabilities of an unclassified sampling unit arising from each of the classes. But in the all-possible-subsets and minimal-best-subset classification methods of Section 4, misclassification costs, prior probabilities, and possibly unequal $\underset{\sim}{\Sigma}_{k(u)}$ are incorporated at every stage.

4. ESTIMATED BAYES RISKS FOR ALL POSSIBLE SUBSETS

A classification rule ϕ based on u arbitrary groups of variables is defined by specifying $\phi(k|\underset{\sim}{y}_{(u)})$, the probability under rule ϕ of classifying a sampling unit into class k after observing $\underset{\sim}{y}_{(u)}$ on it. For the most usual Bayes rule with respect to prior probabilities $\underset{\sim}{\pi} \equiv (\pi_1,\cdots,\pi_K)^t$ of a sampling unit being from classes k=1,$\cdots$,K,

$$q_{k(u)} \equiv \phi(k|\underset{\sim}{y}_{(u)}) = \begin{cases} 1 & \text{iff } \sum_{j=1}^{K} \pi_j C_{kj} f(\underset{\sim}{y}_{(u)}|j) = \underset{\ell=1,\cdots,K}{\text{minimum}} \sum_{j=1}^{K} \pi_j C_{\ell j} f(\underset{\sim}{y}_{(u)}|j) \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where C_{kj} is the cost of misclassification of a sampling unit from class j as being from class k, $C_{kk} = 0$ for every k, $\Sigma_{k=1}^{K}\pi_k = 1$, $\pi_k > 0$ (k=1,$\cdots$,K), and $f(\underset{\sim}{y}_{(u)}|j)$ is the p.d.f. for $\underset{\sim}{y}_{(u)}$ from class j. The Bayes risk $R_u(\phi,\underset{\sim}{\pi})$ of misclassification of a sampling unit can be estimated by

$$\hat{R}_u(\phi,\underset{\sim}{\pi}) \equiv \sum_{j=1}^{K} \pi_j \sum_{k=1}^{K} C_{kj} p_u^{\phi}(k|j) \ , \tag{4}$$

where $p_u^{\phi}(k|j)$ is the proportion of the class j test sampling units misclassified into class $k \neq j$ based on the u groups of variables used in ϕ. See Appendix 1 for a derivation of (4).

The classification rule ϕ will be the simple form (3) of the Bayes rule and $\underline{\pi}$ will be regarded as fixed, so ϕ and $\underline{\pi}$ will be omitted from the notations $R_u(\phi,\underline{\pi})$ and $p_u^{\phi}(k|j)$, giving R_u and $p_u(k|j)$. The new notation $R(g_1,\cdots,g_u)$ can then be used to identify the u groups involved in R_u.

As $\hat{R}_u$ is a linear function of (correlated) means that are distributed asymptotically normally (as $m_k \to \infty$, $k=1,\cdots,K$), given the reference data, it is itself asymptotically normal with mean $E(\hat{R}_u) = R_u$ and variance $V(\hat{R}_u)$. As illustrated by the example in Section 5, it can be reasonable to assume that $\hat{R}_u$ is approximately normal for small m_k ($k=1,\cdots,K$). Making this assumption here, although it is not necessary, facilitates the definition of a simple empirical criterion to quantify the Bayes classification accuracy attributable to an arbitrary subset of u groups:

$$z_u \equiv \hat{R}_u / [\hat{V}(\hat{R}_u)]^{\frac{1}{2}} \quad , \tag{6}$$

where $\hat{V}(\hat{R}_u)$ is the estimated variance of $\hat{R}_u$.

Now $V(\hat{R}_u)$ must be defined to specify fully the approximately normal distribution of $\hat{R}_u$. Using the K independent test samples from which corresponding estimated misclassification probabilities necessarily have zero covariance, Evans (1984) found this variance to be

$$V(\hat{R}_u) = \sum_{j=1}^{K} \pi_j^2 \left\{ \sum_{k=1}^{K} C_{kj}^2 \sigma_{kj(u)}^2 / m_j + 2 \sum_{1 \leq k < \ell \leq K} C_{kj} C_{\ell j} \sigma_{kj(u),\ell j(u)} / m_j \right\} \tag{7}$$

where $\sigma_{kj(u)}^2/m_j$ is the variance of $p_u(k|j) = \bar{q}_{kj(u)}$, the average of the $q_{k(u)}$ values from class j test observations, $\underline{y}_{ji(u)}$, $i=1,\cdots,m_j$, and $\sigma_{kj(u),\ell j(u)}/m_j$ is the covariance of $p_u(k|j)$ and $p_u(\ell|j)$. An unbiased esti-

mator of $V(\hat{R}_u)$ is obtained by replacing $\sigma^2_{kj(u)}$ and $\sigma_{kj(u),\ell j(u)}$ by their unbiased estimators $s^2_{kj(u)}$ and $s_{kj(u),\ell j(u)}$; let it be denoted by $\hat{V}(\hat{R}_u)$. (Note that $s^2_{kj(u)}$ [$s_{kj(u),\ell j(u)}$] is the usual unbiased sample variance [covariance] of $q_{k(u)}$ [$q_{k(u)}$ and $q_{\ell(u)}$], which may take the value 0 or 1, over all class j test sampling units.) Then standardized estimated Bayes risk is defined by z_u in (6), which can be used as the selection criterion in all-possible-subsets classification; the smaller the z_u, the greater the classification accuracy.

Under multivariate normality of $\underline{Y}$, as assumed later for the example in Section 5, the p.d.f. for class j is

$$f(\underline{y}|j) = (2\pi)^{-GD/2}|\underline{\Sigma}_j|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\underline{y}-\underline{\mu}_j)^t\underline{\Sigma}_j^{-1}(\underline{y}-\underline{\mu}_j)\right\}, \qquad (5)$$

where $\underline{\Sigma}_j$ (or a common $\underline{\Sigma}$) is required to be nonsingular so that its inverse exists. When $\underline{\Sigma}_j$ (or $\underline{\Sigma}$) and $\underline{\mu}_j$ are unknown, as is usual, $f(\underline{y}|j)$ is estimated by replacing $\underline{\Sigma}_j$ (or $\underline{\Sigma}$) and $\underline{\mu}_j$ by the reference MSP matrices $\underline{S}_j$ (or $\underline{S}$) and mean vectors $\bar{\underline{y}}_{j\cdot}$, respectively, where $\underline{S}_j$ (or $\underline{S}$) must be nonsingular, i.e., $GD \le r_j-1$ for $j=1,\cdots,K$ (or $GD \le N-K$). For u groups of variables rather than all of $\underline{y}$, uD replaces GD and the subscript (u) is added throughout where appropriate. Then the Bayes rule is to allocate a sampling unit to the class $k\varepsilon\{1,\cdots,K\}$ that minimizes

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

For all-possible-subsets classification, first perform a Bayes classification of the M test sampling units using each subset of u groups for $u=1,\cdots,G$, or until the last size for which $uD \le \underset{k\varepsilon\{1,\cdots,K\}}{\text{minimum}}(r_k-1)$ to

ensure nonsingularity of all $\underline{S}_{k(u)}$ (or uD ≤ N-K if using $\underline{S}_{(u)}$). Then calculate the standardized estimated Bayes risk $z_u = z(g_1,\cdots,g_u)$ for each subset of u groups. Next, identify the best (= smallest z_u) subset of size 1, of size 2, and so on. That is, find the subset of groups $g_1,\cdots,g_u$ that achieves

$$\underset{g_1,\cdots,g_u \in \{1,\cdots,G\}}{\text{minimum}} \quad z(g_1,\cdots,g_u) \tag{8}$$

for each u=1,$\cdots$,G-1. The choice of which "best" subset to use for future allocations is then a compromise between ease or cheapness of measurement and earliness and desired accuracy of classification.

The purpose now is to develop a test of additional reduction in Bayes risk due to adding a second subset of u groups to a first subset of v groups, i.e., to enable a probabilistic statement to be made about the significance of the difference in estimated Bayes risk between two subsets. Although one of these subsets is nested within the other, the methods to be given are valid for the comparison of any two subsets.

First, using the first subset of v groups, find $\hat{R}_v=\hat{R}(g_1,\cdots,g_v)$ and $\hat{V}(\hat{R}_v)$. 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

$$\hat{R}_{u\cdot v} \equiv \hat{R}_v - \hat{R}_{u+v} \tag{9}$$

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 $H_o: R_v = R_{u+v}$. 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

$$V(\hat{R}_{u\cdot v}) = V(\hat{R}_v) + V(\hat{R}_{u+v}) - 2\,\mathrm{Cov}(\hat{R}_v,\hat{R}_{u+v}) \ . \tag{10}$$

To obtain an estimate $\hat{V}(\hat{R}_{u\cdot v})$ of $V(\hat{R}_{u\cdot v})$, we first need to define $Cov(\hat{R}_v,\hat{R}_{u+v})$ 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:

$$z_{u\cdot v} \equiv \hat{R}_{u\cdot v}/[\hat{V}(\hat{R}_{u\cdot v})]^{\frac{1}{2}} . \tag{11}$$

The approximate normality of $\hat{R}_{u\cdot v}$ statistics for small m_k will be illustrated by the example in Section 5.

Evans (1984) found the covariance of $\hat{R}_v$ and $\hat{R}_{u+v}$ to be

$$Cov(\hat{R}_v,\hat{R}_{u+v}) = \sum_{j=1}^{K} \pi_j^2 \sum_{k=1}^{K} \sum_{\ell=1}^{K} C_{kj} C_{\ell j} \sigma_{kj(v),\ell j(u+v)}/m_j \tag{12}$$

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, $\hat{Cov}(\hat{R}_v,\hat{R}_{u+v})$, 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.

Although hold-out estimators of misclassification probabilities and their (co)variances have been used for the sake of simplicity, theoretically they could be replaced by other estimators and their (co)variances in the derivation of z_u and $z_{u \cdot v}$. But hold-out estimators will continue to be used here.

McKay and Campbell (1982b) suggested a probabilistic alternative to the empirical comparison of all possible subsets. They stated that "The ideal procedure would isolate a set of subsets whose corresponding estimated error rates are not significantly different among themselves but which are significantly lower than for all other subsets; suitable controls on significance levels would be needed." In principle, such a method can now be constructed analogously to that advocated by McKay and Campbell (1982a) for discrimination, by replacing Rao's test with the test of additional reduction in Bayes risk to isolate a number of adequate subsets that essentially retain the same classification accuracy as for all groups. But, in practice, the full set of groups does not necessarily give the smallest Bayes risk and thus cannot be used as a benchmark for such tests, although the overall-best (=smallest z) subset could be used in its place, provided that it is better than random allocation (Appendix 2). Then, from among the adequate subsets, a subjective choice must be made as to which subset should be used for allocation. An optimal subset to choose would be one that retains most of the allocation accuracy of the overall best subset — which every adequate subset does, by definition — and contains as few

groups as possible. One obvious choice would be the smallest of the best subsets of different sizes, that are also adequate; this subset is guaranteed to be optimal. An alternative to McKay and Campbell's approach is the minimal-best-subset classification of Evans (1984) for objectively selecting a single subset from among the best subsets that is as small as possible while retaining as much accuracy as possible. This method involves a series of dependent tests of additional reduction in Bayes risk and is not a simultaneous test procedure. Comparisons are made of all G groups versus the best subset of size G-1, the latter versus the best subset of size G-2,$\cdots$, and the best subset of size 1 versus no groups, i.e., random allocation corresponding to the null group g_0 (Appendix 2). Although these comparisons are of subsets that are not necessarily nested, the test of accuracy still applies after replacing $\hat{R}_{u\cdot v}=\hat{R}_v-\hat{R}_{u+v}$ by $\hat{R}_{s2\cdot s1}=\hat{R}_{s1}-\hat{R}_{s2}$ and making other contingent notational changes to allow for arbitrary subsets of size s1 and s2. If and when this test leads to the rejection of one of the null hypotheses of no additional accuracy by the larger best-subset then the latter is selected; otherwise select no groups. This chosen subset is taken as the minimal-best subset. As a final check that this chosen subset is adequate, i.e., retains most of the classification accuracy, an extra test could be done to compare the chosen subset with the overall best set of groups. If the chosen subset significantly decreases accuracy then the next larger subset should be selected and the check repeated, and so on, until an optimal subset is obtained. If $GD > \min_j(r_j-1)$ in the case of multivariate normality then comparisons can only begin with the best subset of the largest size v for which all $\underset{\sim}{S}_{j(v)}$ are still nonsingular versus the best subset of next largest size,

thus ignoring all larger subsets. In that situation, the best subset from the sizes considered would have to be used in place of the overall best subset.

5. A REMOTE SENSING EXAMPLE

All-possible-subsets and minimal-best-subset discrimination and classification methods 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). Fields of fallow and woodlands and uncommon crop classes have been excluded from this study but additional fields of the K=13 common crop classes in the Hillston area have been included. These 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 $\bar{\underline{y}}_{j\cdot}$ 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 normally distributed, Landgrebe (1980) has demonstrated that they are approximately so. Accordingly, multivariate normality will be assumed here.

The statistics $\bar{\underline{y}}_k$. and $\underline{S}_k$, k=1,···,13, were used to establish the sample Bayes rules for classifying the M test observations on the basis of each subset of groups of size u=1,···,5. Also, a totally random allocation (i.e., with no groups) of the M test observations was performed as a check on how much better are the Bayes rules built from reference data. For each Bayes rule, based on the anticipated relative proportions of classes in the study district, the prior probabilities π_k, k=1,···,13, were (as percentages) 2, 8, 13, 24, 2, 2, 4, 2, 2, 6, 2, 9, and 24. Relative misclassification costs obtained from the district's agronomist were as follows: C_{kj}=1 for k=1,···6, j=7,···,11; 1 for k=7,8, j=1,···,6,9,10,11; 1 for k=9,10,11, j=1,···,8; 2 for k=12,13, j=1,···,11; 2 for k=1,···,11, j=12,13; and 0 otherwise.

From all-possible-subsets classification, the best (=minimal z_u) 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 (worse than the best date alone) and z_0=11.7. The minimal-best classification method (α=.05) chose May plus October. As this subset is clearly optimal, it is unnecessary to do a simultaneous classification to find adequate subsets. Dawbin and Evans (1988) considered only five subsets, not including May + October, and chose all dates except October 7. Their subset is the best here of size 4 and has the advantage that classifications could have been done earlier on

September 21 but the disadvantage that four dates were needed to achieve the same accuracy as the minimal-best-subset of May + October. 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.

To test the normality of distribution of estimated Bayes risks and their differences, 100 separate bootstrap samples were taken from the test fields, i.e., m_j fields randomly sampled with replacement from the m_j fields in class j, for all j=1,···,13. All samples were classified by selected subsets of dates to give 100 estimated Bayes risks for each subset. Subsets considered were: null; April; April + May; April + May + August; April + May + August + September; and April + May + August + September + October. A Shapiro-Wilk test of normality (SAS Institute 1985b) was applied to the risks for each subset and for each successive difference. Normality was not rejected for any of the differences (p-values = .13, .13, .65, .63, .42) but was rejected for two of the subsets (p-values = .26, .52, .002, .15, .02, .17). Even though the m_j and r_j were small, as in many practical situations, the assumption in Section 4 of approximate normality has been partially justified here.

To give all-possible-subsets discrimination a fair chance of selecting classification-best subsets, all M+N fields were used to calculate Wilks's lambda for each subset. Also, to ensure a proper comparison of classification with the discrimination that assumes a common covariance matrix and prior probability for all crop types, the classification rule was modified to incorporate the same assumptions. Table 1 summarizes the resulting all-possible-subsets classification and discrimination of the original data. Only one of the discrimination-best (= minimal Λ_u) subsets, namely

that for size 1, was also best for classification. The Λ-best subsets of sizes 2, 3, and 4 had z-values that were 22, 24, and 41% higher, respectively, than the corresponding z-best subsets. The minimal-best discrimination subset, and the only adequate subset by the method of McKay and Campbell (1982a), consisted of all dates (α=.05). This retention of all dates (z_5=3.5) conflicts with the choice of only two dates, August and September (z_2=4.5), by the minimal-best classification method. From inspection of the progressive reductions in z_u from the z-best subset of size 1 to the best of size 4 (Table 1), it appears that the optimal subset (i.e., the smallest subset that retains most of the accuracy) is April + May + August (z_3=3.4). Thus, the minimal-best discrimination approach retains too many dates and the classification approach apparently too few – but the test of additional accuracy indicated that the best three dates were not significantly better than the best pair.

[Table 1 goes about here]

To enable a full comparison of discrimination and classification, each of the 100 bootstrap samples was used in place of the original sample of test fields and submitted with the original reference fields for analysis by the all-possible-subsets algorithms. As the optimality of a minimal-best subset is unknown, whereas at least one of the adequate subsets from simultaneous classification is guaranteed to be optimal, only the performance of minimal-best classification is studied here. Methods of analysis and interpretation were as used for the original data in the previous paragraph and the results are now summarized. Of the Λ-best subsets of one to four dates, only 36, 36, 20, and 13% of them, respectively, were also z-best; for those that weren't z-best, their average increase in estimated

Bayes risk over the z-best subset of the same size was 20, 23, 37, and 41%, respectively. For each of the 100 new data sets, the minimal-best, and the only adequate, subset of dates for discrimination was the full set of dates.

Minimal-best classification selected subsets of size 1, 2, 3, and 4, respectively, in 5, 22, 61, and 12% of cases. In 91% of these cases, an optimal or near-optimal subset was selected, with the near-optimal subsets tending to include too few groups. In the near-optimal cases, the simple differences $z_{s_1} - z_{s_1+s_2}$ suggested a decrease in risk due to adding $s_2 \geq 1$ extra groups, whereas the $z_{1.s_1}$ statistics (which also involve the covariance between $\hat{R}_{s_1}$ and $\hat{R}_{s_1+1}$) showed no statistically significant improvements. For the clearly suboptimal cases, consisting of all 5 single-group subsets selected and 4 out of the 22 two-group subsets, an extra test of the selected subset versus the overall best set of groups could have indicated the need to add more groups and thus get closer to optimality. Overall, the minimal-best-subset classification method has performed well in its search for an optimal subset.

The Interactive Matrix Language (IML) procedure of the SAS package (SAS Institute Inc. 1985a) was used for all aspects of this example, including estimation, bootstrapping, and implementation of the all-possible-subsets discrimination and classification methods. Three different computers have been used during the development and the running of these programs: an IBM 3081 mainframe (OS VS2/MVS), a Prime 6350 minicomputer (Primos), and an NEC Powermate 2 microcomputer (MS-DOS). CPU and IO times were recorded for two sizes of data sets. Evans (1984) considered 50-100 observations over only 5 classes, for the same groups of variables as here. All-possible-subsets and minimal-best discrimination used only 2 secs CPU

and <1 sec IO, whereas classification used 2 mins CPU and 3 secs IO, on the IBM. For the larger data set here, involving 240 observations over 13 classes, discrimination took 17 secs CPU and 6 secs IO on the Prime. In contrast, classification using equal (unequal) covariance matrices took 1.75 (3.75) hours CPU and 0.5 (1.5) mins IO. All computing times were approximately trebled on the NEC. Although discrimination is much easier to program and faster to run, it often selects sub-optimal subsets for classification. Thus, to ensure that an optimal subset is selected, it is worthwhile to meet the once-up additional expense of classification, i.e., once the subset is selected, the classification rule is ready to apply to any number of unidentified sampling units. In any case, computing time is generally not a serious limitation with the ready availability of dedicated microcomputers.

6. SUMMARY AND CONCLUSIONS

All-possible-subsets discrimination methods can be inadequate for selection of a subset of allocation variables. There is consequently a need for subset selection methods based on a classification criterion. The estimated Bayes risk given in this paper is such a criterion. It was seen in the example of Section 5 that estimated Bayes risk is approximately normally distributed for small sample sizes. Thus the standardized form of estimated Bayes risk was justified as an appropriate criterion for comparing different subsets.

Grouping of variables according to time of measurement, type of variable, or other means has been proposed as a way of decreasing computing cost by selecting groups rather than individual variables. Nevertheless, when there are many groups and classes, computing expense may still become

prohibitive. In that case, it may be preferable only to consider all possible subsets of up to a certain number of groups. However, Wilks's lambda could also be used to compare the larger subsets to screen out likely sub-optimal subsets.

Although the all-possible-subsets and minimal-best-subset classification methods of this paper were only given in detail for the simple case of an equal number of variables per group, they are also readily applicable to groups of unequal sizes.

ACKNOWLEDGEMENTS

The authors are indebted to Catherine Plunkett and Vicki Harris for their help with computing and summarization of results.

REFERENCES

COSTANZA, M. C. and AFIFI, A. A. (1979), "Comparison of stopping rules in forward stepwise discriminant analysis," *Journal of the American Statistical Association*, 74, 777-785.

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. (1984), *Stagewise Selection and Classification of Multivariate Repeated Measurements*, Ph.D. Dissertation, 1983, Ann Arbor, Michigan: University Microfilms International.

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

HAND, D. J. (1986), "Recent advances in error rate estimation," *Pattern Recognition Letters*, 4, 335-347.

HENSCHKE, C. I. and CHEN, M. M. (1974), "Variable selection technique for classification problems," *Educational and Psychological Measurement*, 34, 11-18.

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.

MCCABE, G. P. (1975), "Computations for variable selection in multiple discriminant analysis," *Technometrics*, 17, 103-109.

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.

MCLACHLAN, G. J. (1976), "A criterion for selecting variables for the linear discriminant function," *Biometrics*, 32, 529-534.

MCLACHLAN, G. J. (1980), "On the relationship between the F test and the overall error rate for variable selection in two-group discriminant analysis," *Biometrics*, 36, 501-510.

MORRISON, D. F. (1978), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill.

RUBINSTEIN, R. Y. (1981), *Simulation and the Monte Carlo Method*, New York: Wiley Interscience.

SAS INSTITUTE INC. (1982), *SAS User's Guide: Statistics* (1982 ed.), Cary, North Carolina.

SAS INSTITUTE INC. (1985a), *SAS IML User's Guide for Personal Computers*, Version 6 Edition, Cary, North Carolina.

SAS INSTITUTE INC. (1985b), *SAS Procedures User's Guide for Personal Computers*, Version 6 Edition, Cary, North Carolina.

SCHAAFSMA, W. and VAN VARK, G. N. (1979), "Classification and discrimination problems with applications, Part IIa," *Statistica Neerlandica*, 33, 91-126.

VAN VARK, G. N. (1976), "A critical evaluation of the application of multivariate statistical methods to the study of human populations from their skeletal remains," *HOMO*, 27, 94-114.

APPENDIX 1: DERIVATION OF ESTIMATED BAYES RISK

The Bayes risk of misclassification of a sampling unit by a classification rule ϕ, based on an observation vector $\underline{Y}$, and with respect to prior probabilities $(\pi_1,\cdots,\pi_K)^t \equiv \underline{\pi}$ of a sampling unit being from classes $k=1,\cdots,K$, is

$$R(\phi,\underline{\pi}) \equiv \int_{\underline{y}} \sum_{k=1}^{K} \phi(k|\underline{y}) \sum_{j=1}^{K} \pi_j C_{kj} f(\underline{y}|j) d\underline{y} \ , \qquad \text{(A1.1)}$$

where $\phi(k|\underline{y})$ is the probability assigned by rule ϕ to classifying a sampling unit into class k after observing $\underline{y}$ on it, C_{kj} is the cost of misclassification of a sampling unit from class j as being from class k, $C_{kk}=0$ for every k, $\sum_{k=1}^{K}\pi_k = 1$, $\pi_k > 0$ $(k=1,\cdots,K)$, and $f(\underline{y}|j)$ is the p.d.f. for $\underline{y}$ from class j.

A Bayes rule with respect to $\underline{\pi}$ is any rule ϕ that minimizes this Bayes risk. The simplest and most usual Bayes rule (Mardia, Kent, and Bibby 1979, p.308) is constructed essentially as follows. For each $k=1,\cdots,K$, define the variable $Q_k \equiv \phi(k|\underline{Y})$ whose possible values for an observation $\underline{y}$ are given by

$$q_k \equiv \begin{cases} 1 \text{ iff } \sum_{j=1}^{K} \pi_j C_{kj} f(\underline{y}|j) = \underset{\ell\in\{1,\cdots,K\}}{\text{minimum}} \sum_{j=1}^{K} \pi_j C_{\ell j} f(\underline{y}|j) \\ 0 \text{ otherwise,} \end{cases} \qquad \text{(A1.2)}$$

but with the restriction in the case of a non-unique minimum that only one of $q_1,\cdots,q_K$ is unity. That is, for each sampling unit, whose observation is $\underline{y}$, classify it as coming from the class $k\in\{1,\cdots,K\}$ that minimizes

$\sum_{j=1}^{K} \pi_j C_{kj} f(\underline{y}|j)$. The p.d.f. $f(\underline{y}|j)$ can take any parametric form. When the parameters are unknown they are estimated and plugged in to give $\hat{f}(\underline{y}|j)$ in place of $f(\underline{y}|j)$. If the parametric form is unknown then the p.d.f. can be estimated nonparametrically to give $\hat{f}(\underline{y}|j)$. Under this simple rule, the Bayes risk formula reduces to

$$R(\phi,\underline{\pi}) = \sum_{j=1}^{K} \pi_j \sum_{k=1}^{K} C_{kj} P^{\phi}(k|j) \ , \tag{A1.3}$$

where $P^{\phi}(k|j) \equiv E_j(Q_k)$; that is, the expectation for class j of Q_k is the probability that rule ϕ will misclassify a sampling unit from class j, on the basis of its observed value of $\underline{Y}$, as being from class k.

Although $P^{\phi}(k|j)$ is usually intractable to evaluate numerically, it can be readily estimated by $p^{\phi}(k|j) \equiv m^{\phi}(k|j)/m_j$, where $m^{\phi}(k|j)$ and $p^{\phi}(k|j)$ are the number and proportion, respectively, of the m_j class j test sampling units misclassified into class $k \neq j$. Equivalently, $p^{\phi}(k|j)$ is obtained as $\bar{q}_{kj}$, the average of the observed $q_k \equiv \phi(k|\underline{y}_{ji})$ values of Q_k on test sampling units $i=1,\cdots,m_j$ from class j. Then Bayes risk can be estimated by

$$\hat{R}(\phi,\underline{\pi}) = \sum_{i=1}^{K} \pi_j \sum_{k=1}^{K} C_{kj} p^{\phi}(k|j) \ . \tag{A1.4}$$

This entire analysis can be based on $\underline{y}_{(u)}$ instead of $\underline{y}$, giving (4).

APPENDIX 2. RANDOM ALLOCATION

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 (4) to obtain $\hat{R}_0$; and 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 (7). Equation (6) is used to obtain z_0. The 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

$$\hat{R}(g_{v+1},\cdots,g_{v+u}|g_1,\cdots,g_v) = \hat{R}(g_1,\cdots,g_v) - \hat{R}(g_1,\cdots,g_{v+u}) \quad , \quad (A2.1)$$

where $g_1,\cdots,g_{u+v} \in \{1,\cdots,G\}$. 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 (11).

TABLE 1. The Λ-best and z-best dates and their Λ_u and z_u values for each subset of size u=1,2,3,4.*

	u=1		u=2		u=3		u=4	
	Λ-best	z-best	Λ-best	z-best	Λ-best	z-best	Λ-best	z-best
DATES	3	3	2,3	3,4	2,3,4	1,2,3	1,2,3,4	2,3,4,5
Λ×100	17.2	17.2	4.5	4.7	1.5	1.6	0.6	0.9
Λ-RANK	1	1	1	2	1	2	1	3
z	7.0	7.0	5.5	4.5	4.2	3.4	4.5	3.2
z-RANK	1	1	3	1	4	1	4	1
	z(0)=13.1		z(1,2,3,4,5)=3.5		Λ(1,2,3,4,5)=0.3			

* Dates 1-5 correspond to chronological order. Λ-rank and z-rank values are ranks from 1 (lowest=best) to 5 or 10 (highest=worse) of Λ and z, respectively, among the 5 or 10 subsets of the same size u=1 or 4 or u=2 or 3. z(0) denotes the z_0 value for random classification.