

A three-population constrained discrimination procedure

David Patterson
University of Montana

February 14, 2014

Abstract

Classification rules with a reserve judgment option provide a way to satisfy constraints on the misclassification probabilities when there is a high degree of overlap among the populations. Constructing rules which maximize the probability of correct classification while satisfying such constraints is a difficult optimization problem. This paper uses a result of Anderson (1969) on the form of the optimal solution to develop a relatively simple and computationally fast method for three populations which has a nonparametric quality in controlling the misclassification probabilities. Simulations demonstrate that this procedure performs well.

Keywords: classification, constrained discrimination, reserve judgment option, reject option

1 Introduction

The usual classification rules in discriminant analysis are *forced* rules; they classify an observation into one of K populations ($K \geq 2$) even if there is doubt about which population an observation belongs to. If two or more of the populations overlap substantially then the probability of errors can be quite high for even the best forced rules. Adding a reserve judgment option (also called a “reject” or “in doubt” option) to the classification choices where no decision is made provides a way to address this problem; individuals for which the classification is not clear can be put in this category and are, by definition, neither correctly classified nor misclassified. Clearly, a correct classification is preferred to no decision, so the goal is to minimize the use of the reserve judgment option while satisfying one or more constraints on the misclassification probabilities. Rules with a reserve judgment option which attempt to satisfy such constraints will be referred to as *constrained* rules.

As an example, suppose it is possible to determine which of three types of a disease a patient has based on an expensive and invasive procedure such as a biopsy. As a substitute for the biopsy, it

is desired to classify patients based on easy-to-measure blood chemistry variables. However, the disease types may not be well separated on these variables and even an optimal discriminant rule may have a high misclassification rate. A possible strategy would be to classify those patients on the basis of blood chemistry for which the classification is clear and put the remainder in the reserve judgment category; those in the latter group are biopsied.

Constructing optimal or nearly optimal constrained rules is a difficult problem, particularly for three or more populations. While there has been considerable research for the two-population problem, there has been far less for $K \geq 3$. In this paper we propose and examine a robust procedure for the three-population problem based on a theoretical result about the form of the optimal solution from Anderson (1969). It stems from the work of Gallagher, Lee & Patterson (1997), but provides a simpler and more straightforward way to estimate a solution in the three-population case.

The organization of the paper is as follows. Section 2 provides the background for the problem and presents the form of the optimal solution for the K -population case. Section 3 discusses an approach to estimating the optimal solution in the three-population case which is nonparametric in controlling the misclassification probabilities. Sections 4 presents the design and Section 5 the results of a simulation to evaluate the performance of the proposed approach. Section 6 is the discussion and conclusions.

2 Background

Let H_1, \dots, H_K denote the K populations. The goal is to construct a rule to classify an object into one of these populations based on \mathbf{x} , a p -variate vector of measurements obtained from the object. The distribution of \mathbf{x} in population H_i is given by probability density f_i , assumed known for now, and the prior probability that an object comes from H_i by π_i (where $\sum_1^K \pi_i = 1$). A forced rule is characterized by a partition $\{R_1, \dots, R_K\}$ of \mathbb{R}^p where \mathbf{x} is classified as coming from H_i if and only if $\mathbf{x} \in R_i, i = 1, \dots, K$. A classification rule with a reserve judgment option can be characterized by a partition $\{R_0, R_1, \dots, R_K\}$ of \mathbb{R}^p where R_0 is the reserve judgment region. Let

$$q_{ij} = \int_{R_j} f_i(\mathbf{x}) d\mathbf{x}, \quad i = 1, \dots, K; j = 0, 1, \dots, K \quad (1)$$

be the probability that either type of rule classifies a random observation from H_i into H_j , letting $j = 0$ represent the reserve judgment category. This is a misclassification probability for $i \neq j$ and $j \neq 0$. The probability of correct classification of a random observation is

$$\sum_{i=1}^K \pi_i q_{ii}. \quad (2)$$

The forced rule which maximizes (2) classifies \mathbf{x} into the population H_k for which the posterior probability $p_k(\mathbf{x}) = \pi_k f_k(\mathbf{x}) / \sum_{i=1}^K \pi_i f_i(\mathbf{x})$ is a maximum.

For rules with a reserve judgment option, the goal is to maximize (2) subject to constraints on the misclassification probabilities or some function of them. One possible constraint is simply on the overall (unconditional) probability of misclassification,

$$\sum_{i=1}^K \pi_i \sum_{\substack{j=1 \\ j \neq i}}^K q_{ij} \leq \alpha \quad (3)$$

where $0 < \alpha < 1$. Chow (1957) showed that the rule which maximizes (2) subject to (3) is defined by

$$\begin{aligned} R_j &= \{\mathbf{x} : p_j(\mathbf{x}) = \max_{i=1, \dots, K} p_i(\mathbf{x}) \text{ and } p_j(\mathbf{x}) \geq \beta\}, \quad j = 1, \dots, K, \\ R_0 &= \{\mathbf{x} : \max_{i=1, \dots, K} p_i(\mathbf{x}) < \beta\} \end{aligned} \quad (4)$$

for some $0 < \beta < 1$. The constant β represents a threshold which the maximum posterior probability must meet in order for an observation to be classified. The determination of β is not easy except in very simple cases as it involves the evaluation of p -dimensional integrals like (1).

In this paper, we consider constraints on each of the $K(K-1)$ conditional probabilities of misclassification since one may wish to exercise greater control over some misclassifications than others. The constraints are

$$q_{ij} \leq \alpha_{ij}, \quad i, j = 1, \dots, K; \quad i \neq j \quad (5)$$

where the α_{ij} are constants between 0 and 1. Anderson (1969) showed that, under very general regularity conditions, the rule which maximizes the probability of correct classification (2) subject to (5) has the following form:

$$R_j = \{\mathbf{x} : L_j(\mathbf{x}) = \max_{i=0, 1, \dots, K} L_i(\mathbf{x})\}, \quad j = 0, 1, \dots, K \quad (6)$$

where

$$L_j(\mathbf{x}) = \pi_j f_j(\mathbf{x}) - \sum_{\substack{i=1 \\ i \neq j}}^K \lambda_{ij} \pi_i f_i(\mathbf{x}), \quad j = 1, \dots, K, \quad (7)$$

$$\text{and } L_0(\mathbf{x}) = 0,$$

and where the λ_{ij} are unique (except on a set of measure 0) non-negative constants. This formulation of the L_j 's is slightly different from, but equivalent to, that given by Anderson (he did not include π_i in the summation of the second term on the right hand side of L_j ; it was, in effect, folded into λ_{ij}). The formulation in (7), however, is more convenient for the analysis here.

The rule is unchanged if the $L_j(\mathbf{x})$'s are divided by $\sum_{i=1}^K \pi_i f_i(\mathbf{x})$. This yields the following definition of the L_j 's for the optimal rule as a function of the posterior probabilities:

$$L_j(\mathbf{x}) = p_j(\mathbf{x}) - \sum_{\substack{i=1 \\ i \neq j}}^K \lambda_{ij} p_i(\mathbf{x}), \quad j = 1, \dots, K, \quad (8)$$

and $L_0(\mathbf{x}) = 0$.

Setting all the λ_{ij} 's equal to 0 gives the optimal forced rule. Anderson showed that if the optimal forced rule satisfies the constraints in (5), then it is also the optimal constrained rule. The optimal rule in (4) (which maximizes the probability of correct classification subject to a single constraint on the unconditional probability of misclassification) can also be expressed in the form of (6) by setting all the λ_{ij} equal to $\beta/(1 - \beta)$.

Anderson's result provides only the form of the optimal rule; it does not provide a way to determine the values of the optimal λ_{ij} 's, a very difficult optimization problem involving $K(K - 1)$ parameters.

Rules with a reserve judgment option, like forced rules, can be approached from a decision theoretic viewpoint in which the goal is to find a rule which minimizes the expected cost of classification. Let c_{ij} be the cost of classifying an observation from H_i into H_j , with a correct classification assumed to have zero cost (that is, $c_{ii} = 0$). Let c_{i0} represent the cost of not classifying an object from H_i , reflecting, perhaps, the additional cost of obtaining further information on the object in order to make a definite classification (e.g., the cost of the biopsy in the disease example). Commonly, the $c_{i0}, i = 1, \dots, K$, would all be equal (and would be less than the costs of misclassification). The expected cost of using the rule $\{R_0, R_1, \dots, R_K\}$ is

$$\sum_{i=1}^K \pi_i \sum_{j=0}^K c_{ij} q_{ij}. \quad (9)$$

The rule which minimizes (9) is

$$R_k = \left\{ \mathbf{x} : \sum_{i=1}^K c_{ik} p_i(\mathbf{x}) = \min_{j=0,1,\dots,K} \sum_{i=1}^K c_{ij} p_i(\mathbf{x}) \right\}, \quad k = 0, 1, \dots, K. \quad (10)$$

While the decision theoretic approach yields an explicit minimum expected cost rule with reserve judgment option, it may be the case that costs are difficult to specify and an approach based on constraints on the misclassification probabilities is preferred, just as in forced discrimination. That is the approach taken in this paper. However, there is a connection between the two approaches that is given in the following proposition, which is easily proved with a little algebra:

Proposition 1. *The minimum expected cost (Bayes) rule with reserve judgment option in (10) is equivalent to the optimal constrained rule with reserve judgment option in (6) and (8) when*

$c_{i0} = 1$, $c_{ii} = 0$ and $c_{ij} = 1 + \lambda_{ij}$, $i \neq j$.

Thus the optimal constrained rule is a Bayes rule. The quantity λ_{ij} can be viewed as representing the additional cost of misclassifying an observation from H_i into H_j above the cost of not classifying the observation.

In practice, one can only approximate the optimal forced or constrained classification rules since the f_i are not known. A training sample of objects whose classification is known is therefore necessary. Many classification methods rely on estimating the population densities and/or posterior probabilities from the training sample, either parametrically (e.g., linear discriminant analysis) or nonparametrically (e.g., neural networks, nonparametric density estimates). Ripley (1996) discusses many of these methods.

Estimating the parameters of the optimal constrained rule is difficult. The two-population case is relatively straightforward and has been extensively studied, both when the f_i are known and unknown. Habbema, Hermans & Van Der Brugt (1974) discuss parametric solutions. Broffitt, Randles & Hogg (1976) give a rank procedure which is distribution-free in controlling the misclassification probabilities. A large number of recent papers in the machine learning and statistics literature (e.g., Herbei & Wegkamp, 2006; Bartlett & Wegkamp, 2008; Bousair, Beasuseroy & Grall-Maës, 2008) have examined various properties of two-population classifiers with a reserve judgment option. Jeske, Liu, Bent & Borneman (2007) and Choi, Yeo, Kwon & Kim (2011) discuss applications in gene expression data.

Research on rules with a reserve judgment option for three or more populations is much sparser. Grall-Maës & Beasuseroy (2009) present a very general theoretical framework for such problems, but practical procedures are limited. Huberty (1994) discusses a general K -population procedure implemented in PROC DISCRIM in the SAS software (SAS Institute, Inc., 2008). It allows a user to specify a threshold value between $1/K$ and 1 such that an observation is classified only if its maximum estimated posterior probability $\hat{p}_{(K)}(\mathbf{x}) = \max(\hat{p}_1(\mathbf{x}), \dots, \hat{p}_K(\mathbf{x}))$ is greater than the threshold value and is placed in the in-doubt category, otherwise. This follows the form of the rule (4) which maximizes the probability of correct classification subject to a constraint on the overall probability of misclassification. Huberty suggests trying different threshold values and observing the resulting misclassification and no-classification rates in the training sample. We call this the p_{\max} procedure. Yu, Jeske, Ruegger & Borneman (2010) suggest a procedure based instead on the value of the difference between the maximum estimated posterior probability and the second highest, $\hat{p}_{(K)}(\mathbf{x}) - \hat{p}_{(K-1)}(\mathbf{x})$. The larger the difference, the more certain one is of the classification so, again, one chooses a threshold value and classifies an observation into the population with highest posterior probability provided that $\hat{p}_{(K)}(\mathbf{x}) - \hat{p}_{(K-1)}(\mathbf{x})$ exceeds the threshold. They choose the threshold to minimize the expected cost of a decision, but one could also follow the strategy of Huberty and base the threshold on observed correct classification and misclassification rates for the training sample. We call this the p_{diff} procedure when used in this latter fashion.

Gallagher et al. (1997) and Lee, Gallagher & Patterson (2003) are apparently the only authors who have attempted to estimate Anderson's optimal solution for $K \geq 3$ for a given set of constraints as in (5). Their approach has two steps. The first step is to use the training samples to compute estimates \hat{f}_i of the population densities f_i and to compute \hat{f}_i for all the points in the training sample. The second is to find the set of λ_{ij} 's which maximizes the proportion of training sample points correctly classified subject to constraints on the proportions of training sample points misclassified. That is, replace f_i by \hat{f}_i in equations (6) and (7) and find λ_{ij} 's to maximize

$$\sum_{i=1}^K \pi_i \left[\frac{1}{n_i} (\# \text{ of training sample observations from } H_i \text{ in } R_i) \right] \quad (11)$$

subject to

$$\frac{1}{n_i} (\# \text{ of training sample observations from } H_i \text{ in } R_j) \leq \alpha_{ij}, \quad i, j = 1, \dots, K; \quad i \neq j. \quad (12)$$

There are two advantages to using the empirical distribution of the training sample points in order to estimate the optimal λ_{ij} 's. First, it avoids potentially difficult numerical integrations in (1). Second, it has a nonparametric quality in controlling the misclassification rates. Even if the estimates of the population densities are based on incorrect assumptions about the population distributions (for example, using linear discriminant analysis for non-normal populations), this approach ensures that the constraints on the misclassification probabilities will be satisfied for the sample and, therefore, approximately satisfied for the population, at least for large samples. This procedure may not provide good estimates of the optimal λ_{ij} 's if the f_i are poorly estimated, but it will tend to control the misclassification rates regardless.

Gallagher et al. (1997) used mixed integer programming to calculate λ_{ij} 's using this framework, but the procedure was very slow and didn't always converge because of the numerical difficulties involved, even for three populations with small sample sizes. Lee et al. (2003) then used linear programming to estimate the optimal λ_{ij} 's which was much faster but involved approximating the problem and did not guarantee that the solution found satisfied the constraints for the training sample. We propose a simple alternative procedure for estimating the optimal λ_{ij} 's for the three-population problem that satisfies the constraints for the training sample. It depends on a geometric representation of the estimated posterior probabilities for the training sample.

3 Three-population problem

To examine the form of the optimal solution for $K = 3$, consider the L_i 's of (8) where they are written as functions of the posterior probabilities:

$$\begin{aligned} L_1(\mathbf{x}) &= p_1(\mathbf{x}) - \lambda_{21}p_2(\mathbf{x}) - \lambda_{31}p_3(\mathbf{x}) \\ L_2(\mathbf{x}) &= p_2(\mathbf{x}) - \lambda_{12}p_1(\mathbf{x}) - \lambda_{32}p_3(\mathbf{x}) \\ L_3(\mathbf{x}) &= p_3(\mathbf{x}) - \lambda_{13}p_1(\mathbf{x}) - \lambda_{23}p_2(\mathbf{x}) \\ L_0(\mathbf{x}) &= 0. \end{aligned} \tag{13}$$

The regions R_0, \dots, R_3 in (6) can be conveniently displayed with a ternary plot which displays any point $p(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), p_3(\mathbf{x}))$ as a point on the interior of an equilateral triangle using the Cartesian coordinates $(p_1/\sqrt{3} + 2p_3/\sqrt{3}, p_1)$. The vertices of the triangle represent the points $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ and are numbered 1, 2, and 3, respectively. For any point on the interior of the triangle, the value of $p_i(\cdot)$ is the distance from that point to the side opposite vertex i .

Consider any rule defined by (13) and (6) with non-negative λ_{ij} 's. The regions R_0, \dots, R_3 are defined by the line segments $L_i = 0, i = 1, 2, 3$, and the line segments $L_i = L_j, i \neq j$. $L_i = 0$ is a line segment from the $i - j$ side (at the point where $p_i = \lambda_{ji}/(1 + \lambda_{ji})$ and $p_j = 1 - p_i$) to the $i - k$ side (at the point where $p_i = \lambda_{ki}/(1 + \lambda_{ki})$ and $p_k = 1 - p_i$). The regions are 3-, 4-, or 5-sided polygons depending on which pairs of the line segments $L_i = 0, i = 1, 2, 3$, intersect each other. Figure 1 illustrates one possible rule in which there are two of the possible three intersections.

[Figure 1 here]

For comparison, ternary plots of the p_{\max} and p_{dif} procedures discussed earlier are illustrated in Figure 2. The p_{\max} procedure classifies an observation into the population with the highest posterior probability as long as that maximum probability exceeds a threshold. This procedure is equivalent to Anderson's optimal procedure with all λ_{ij} 's equal. The p_{dif} procedure classifies an observation into the population with the highest posterior probability as long as the difference between the highest probability and the second highest exceeds a threshold. This procedure does not have an equivalence to the optimal procedure.

[Figure 2 here]

While the ternary plot provides a convenient geometric representation of the form of the optimal solution, it doesn't necessarily make the optimization problem any easier. Expressing the optimal rule in terms of the posterior probabilities rather than the population densities changes the p -dimensional integrals in (1) to two-dimensional integrals but the joint distribution of the posterior probabilities $(p_1(\mathbf{x}), p_2(\mathbf{x}), p_3(\mathbf{x}))$ must be derived for each population. Therefore, following the approach of Gallagher et al. (1997) and Lee et al. (2003), we use the empirical

distribution of the training sample points on the ternary plot to estimate the optimal solution. That is, estimate the unknown population densities, either parametrically or nonparametrically, from the training samples, and use the estimated densities to estimate the posterior probability function $(\hat{p}_1(\mathbf{x}), \hat{p}_2(\mathbf{x}), \hat{p}_3(\mathbf{x})) = [1/\sum_1^3 \hat{f}_j(\mathbf{x})](\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \hat{f}_3(\mathbf{x}))$. Plot each training sample point on the ternary triangle based on its estimated posterior probabilities, using different plotting symbols to distinguish between the three populations. If the populations are well separated, then the training sample points from H_1 should be clustered near vertex 1, and those from H_2 and H_3 should be clustered near vertices 2 and 3, respectively. Second, use the empirical joint distribution of the posterior probabilities for the training sample to estimate the optimal λ_{ij} 's; that is, maximize the proportion of training sample points correctly classified (equation 11) subject to the constraints on the proportions of training sample points misclassified (equation 12).

Unfortunately, Gallagher et al. (1997) found this optimization problem extremely difficult, so we propose the following simplified procedure: find the lines $L_1 = 0$, $L_2 = 0$ and $L_3 = 0$ which would individually be optimal if they did not intersect each other. That is, find λ_{21} and λ_{31} to maximize

$$\frac{1}{n_1}(\# \text{ of points from } H_1 \text{ in the region } L_1 > 0) \quad (14)$$

subject to the constraints

$$\begin{aligned} \frac{1}{n_2}(\# \text{ of points from } H_2 \text{ in the region } L_1 > 0) &\leq \alpha_{21} \\ \text{and } \frac{1}{n_3}(\# \text{ of points from } H_3 \text{ in the region } L_1 > 0) &\leq \alpha_{31} \end{aligned} \quad (15)$$

Proceed analogously in positioning $L_2 = 0$ and $L_3 = 0$, thus obtaining a set of λ_{ij} 's. Consider the classification rule $\{R_0, R_1, R_2, R_3\}$ as determined by these λ_{ij} 's from (13). Since $R_i \subseteq \{\mathbf{x} : L_i(\mathbf{x}) > 0\}$, $i = 1, 2, 3$, the constraints are satisfied by this solution, although it may not maximize (11) among all constrained rules of this form.

The actual process of positioning the line $L_1 = 0$ to maximize (14) subject to the constraints in (15) can be easily accomplished using the ‘‘Pivot Algorithm’’ which is described in the Appendix. This algorithm has been implemented in R (R Development Core Team, 2013) and runs quickly even for large examples. Because the empirical distribution of the points on the ternary triangle is discrete, there are an infinite number of maximizing lines $L_1 = 0$. Hence, only lines which go through two training sample points or a training sample point and a vertex are considered. There may be several maximizing lines for each vertex; the combination of solutions across the three vertices which maximizes the total empirical correct classification rate is used. If there is more than one, we use the one that minimizes the area of the reserve judgment region on the ternary plot. This process will be referred to as the pivot procedure.

To illustrate the approach, consider Data Set 36 on ‘‘Chemical and Overt Diabetes’’ in Andrews & Herzberg (1985). The data set consists of several variables measured on 145 patients who had

been clinically classified into one of three groups: 1: overt diabetic ($n_1 = 33$), 2: chemical diabetic ($n_2 = 36$), 3: normal ($n_3 = 76$). Only two explanatory variables are used, insulin resistance (IR) and relative weight (RW), since use of all the variables gives almost perfect discrimination among the groups. A scatterplot (Figure 3(a)) shows considerable overlap among the three groups and forced classification by linear discriminant analysis (with prior probabilities proportional to the sample sizes) shows high misclassification rates (Table 1a). A constrained rule based on the posterior probabilities from LDA was estimated by the pivoting algorithm with all constraints (α_{ij} 's) equal to 0.1. The classification matrix (Table 1b) shows that the constraints are satisfied for the training sample (as they must be by construction) and that the number of correctly classified observations drops by 20 from the LDA forced rule, but the number of misclassified observations drops by 34. Both the forced and constrained rule are displayed on a ternary plot in Figure 3(b). The misclassification rates of both the forced and constrained rules when applied to the training sample will tend, of course, to be optimistically biased as estimates of the misclassification rates for new cases.

[Figure 3 here]

[Table 1 here]

4 Monte Carlo Simulation

A Monte Carlo study was carried out to examine the performance of the pivot procedure. Two types of population distributions were used: bivariate normal and 10% contaminated bivariate normal. Eight different sets of mean positions were used, five with equal and three with unequal covariance matrices. In the equal covariance case, the common covariance matrix was I_2 , the 2×2 identity matrix. In the unequal covariance case, the first group's covariance matrix was diagonal (1, 0.25) and the second and third's were diagonal (0.25, 1). The configurations are reported in Tables 2 and 3 along with the distance between each pair of means. For the equal covariance case, the distance is the Mahalanobis distance, $[(\mu_i - \mu_j)' \Sigma^{-1} (\mu_i - \mu_j)]^{1/2}$ where μ_i and μ_j are the group means and Σ the common covariance matrix (I_2 in this case). For the unequal covariance case, Σ was replaced by the average of the two population covariance matrices in the Mahalanobis distance formula. The population configurations are illustrated in Figure 4. In the 10% contaminated normal distributions, 10% of the population came from a bivariate normal with the same mean but a covariance matrix 100 times as large. Two training sample sizes were used: 40 from each population (120 total) and 100 (300 total). In each case, the simulation was repeated 2000 times. In each of the simulations, a training sample was generated, linear discriminant analysis (with equal priors) was used to estimate the posterior probabilities for the training sample, and the pivot procedure was used to estimate a set of λ_{ij} 's for a constrained solution with two different sets of constraints: all the α_{ij} 's equal to 0.05 and all equal to 0.10.

The p_{\max} and p_{dif} procedures were also computed for comparison. The threshold for each rule was the smallest value such that the constraints on the misclassification probabilities were satisfied for the training sample. All the resulting rules were used to classify 15000 new random observations from each population in each simulation. The linear discriminant analysis forced classification was also computed for each new observation. All simulations were programmed in R (R Development Core Team, 2013); function `lda` in package MASS (Venables & Ripley, 2002) was used for linear discriminant analysis computations.

Tables 2 and 3 here]

[Figure 4 here]

Linear discriminant analysis was used to estimate the posterior probabilities in all simulations even for the unequal covariance matrix and contaminated normal situations (for which it is not optimal) in order to assess the performance of the procedures in these situations.

In configurations E1 and E2, it was possible to calculate Anderson’s optimal constrained rule based on the true population densities. This is because these configurations are symmetric and, since the optimal solution is unique, it follows that the optimal λ_{ij} ’s are all equal. Monte Carlo integration was used to calculate the optimal common λ and the optimal correct classification rate based on two million randomly generated observations from each population. The optimal classification probabilities are included in the results for E1 and E2 in the next section.

5 Simulation Results

The considerations in assessing the performance of a constrained rule are threefold: first, the rule should achieve the desired constraints, at least on average. Second, it should not be too conservative; the average misclassification rates shouldn’t all be significantly less than their bounds. Third, the correct classification rate should seem reasonable when compared to the forced rule (and compared to the optimal solution for E1 and E2).

Only the results for training sample size 100 from each population and constraints all equal to 0.1 are reported here because the results from sample size 40 and from constraints all equal to 0.05 show the same overall patterns. Figure 5 shows the results for the normal population case. The p_{\max} and p_{dif} procedures had almost the same mean classification rates for all configurations so only the results for p_{\max} are displayed. The similarity is likely due to the fact that with three populations, the maximum estimated posterior probability and the difference between the maximum and second highest posterior probabilities are highly correlated. The pivot procedure consistently achieved the desired misclassification rates on average without being too conservative; the only exceptions were average misclassification rates of about 0.11 in a few cases. The p_{\max} and p_{dif} procedures, on the other hand, tended to be more conservative with misclassification

rates more often below the target of 0.1 than the pivot method. The standard deviations of the estimated misclassification rates were similar for the pivot, p_{\max} and p_{dif} procedures and didn't exceed 0.04 except for position U3, where they were as high as 0.053. Most of the variation is due to the variation in the true misclassification rates of the 2000 estimated rules as the variation due to estimation based on 15000 new observations is small (standard deviation $\sqrt{.1 \times .9/15000} = 0.0025$). The constraints on the misclassification probabilities are goals that can never be guaranteed to be achieved (except by a solution which puts all observations in R_0).

[Figure 5 here]

The mean correct classification rates for the pivot method were always higher than for p_{\max} and p_{dif} , sometimes substantially so, particularly in the unequal covariance cases. They closely matched the forced LDA rates for populations which were well separated from the others and closely matched the optimal rates for E1 and E2. An interesting phenomenon in configuration E3 is that the p_{\max} and p_{dif} procedures rarely classified observations into population 1. This is because population 1 is between and close to populations 2 and 3. Therefore, the true posterior probability $p_1(\mathbf{x})$ is never very high for any \mathbf{x} ; this will also generally be true of the estimated posterior probability $\hat{p}_1(\mathbf{x})$ which makes it difficult to meet the threshold for classification.

In the contaminated normal case, there were substantial differences between the p_{\max} and p_{dif} procedures in some of the configurations so both are included in the plots of results (Figure 6). All three constrained procedures consistently achieved the target misclassification rate, on average, but the pivot procedure had consistently higher correct classification rates than the p_{\max} and p_{dif} procedures. The pivot procedure's average correct classification rates were significantly below the optimal rates in E1 and E2; this is not surprising since LDA was used to compute the estimated posterior probabilities in this non-normal situation. The p_{dif} procedure had noticeably higher correct classification rates than the p_{\max} procedure in several cases, but it's not clear why. The standard deviations of the misclassification rates for the constrained procedures were similar to those in the normal case with a maximum of 0.055.

[Figure 6 here]

For training sample size 40, the mean classification rates were almost identical to those for sample size 100, but the standard deviations were about 50% larger, on average. The general pattern and conclusions for misclassification bound of 0.05 were very similar to those for 0.10.

6 Discussion and Conclusions

The pivot procedure provides a straightforward algorithm for computing a constrained solution for three populations which satisfied the desired constraints on average across different population configurations and distributions without being too conservative. It also matched the optimal

solution where it could be computed. These results suggest that optimizing each population separately, as the pivot method does, does not make the procedure too conservative nor to incur much of a penalty relative to the optimal solution, at least in the symmetric configurations. The p_{\max} and p_{dif} procedures also achieved the desired constraints on average, but had lower correct classification rates.

The fact that the p_{\max} and p_{dif} procedures were generally inferior to the pivot procedure is not too surprising given that they each depend on a single parameter while the pivot procedure has six. The p_{\max} and p_{dif} procedures can also be easily applied to any number of populations, but would likely become increasingly inferior to an optimal solution. Anderson's optimal solution, on the other hand, has $K(K - 1)$ parameters for K populations and presents increasing difficulties as K increases. The pivot method does not generalize easily to even four populations; the analog to the ternary plot would be a tetrahedron for $K = 4$. The strategy of optimizing each population separately could still be employed and would involve estimating $K - 1$ parameters for each population, which could feasibly be accomplished with a numerical optimization routine, at least for moderate values of K .

It is also likely that for more than three populations, a user would not be interested in controlling all $K(K - 1)$ individual misclassification probabilities. Chow's (1957) rule in (4) maximizes the probability of correct classification subject to a single constraint on the unconditional overall probability of misclassification. The p_{\max} procedure can be easily adapted to estimate this rule by letting the threshold be the smallest value such that the sum of the proportions of misclassified points (weighted by the prior probabilities) is less than or equal to the constraint.

These scenarios represent two extremes, however: all $K(K - 1)$ constraints or a single constraint. Anderson (1969) provides related results which cover a wide variety of intermediate cases. For example, he shows that the optimal rule subject to a subset of the constraints in (5) has the same form as in (7) but with the λ_{ij} 's corresponding to any missing constraints set equal to 0. He also provided the form of the optimal solution with constraints on sums of the misclassification probabilities (e.g., $q_{21} + q_{31} \leq \alpha$); essentially, there is a single λ for each constraint. This would simplify estimation of the optimal solution.

Finally, the connection between Anderson's (1969) optimal solution and the Bayes minimum expected cost rule established in Proposition 1 makes a hybrid approach possible. Recall that λ_{ij} in (8) can be viewed as representing the additional cost of misclassifying an observation from H_i into H_j above the cost of not classifying an observation. One might be willing to specify the relative size of these additional costs (for example, by specifying that λ_{21} is twice λ_{12}) without specifying the actual values. In addition, one could specify a constraint on the maximum of the misclassification probabilities, that is $q_{ij} \leq \alpha$ for all $i \neq j$. This might be appealing to a user who is willing to specify the relative costs but wants to ensure that none of the misclassification probabilities are too big. The nonparametric approach could be easily adapted to find the

maximum probability of correct classification subject to such a specification.

References

- Anderson, J. A. (1969). Constrained discrimination between k populations. *J.R. Stat. Soc. Ser. B*, *31*, 123–139.
- Andrews, D. F. & Herzberg, A. M. (1985). *Data*. Springer-Verlag.
- Bartlett, P. L. & Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, *9*, 1823–1840.
- Bounsair, A., Beausero, P., & Grall-Maës, E. (2008). General solution and learning method for binary classification with performance constraints. *Pattern Recogn. Lett.*, *29*, 1455–1465.
- Broffitt, J. D., Randles, R. H., & Hogg, R. V. (1976). Distribution-free partial discriminant analysis. *J. Amer. Statist. Assoc.*, *71*, 934–939.
- Choi, H., Yeo, D., Kwon, S., & Kim, Y. (2011). Gene selection and prediction for cancer classification using support vector machines with a reject option. *Comput. Statist. Data Anal.*, *55*, 1897–1908.
- Chow, C. K. (1957). An optimum character recognition system using decision functions. *IRE Trans. Elect. Comput.*, *6*, 247–254.
- Gallagher, R. J., Lee, E. K., & Patterson, D. A. (1997). Constrained discriminant analysis via 0/1 mixed integer programming. *Ann. Oper. Res.*, *74*, 65–88.
- Grall-Maës, E. & Beausero, P. (2009). Optimal decision rule with class-selective rejection and performance constraints. *IEEE Trans. Pattern Anal. Mach. Intell.*, *31*, 2073–2082.
- Habbema, J. D. F., Hermans, J., & Van Der Brug, A. T. (1974). Cases of doubt in allocation problems. *Biometrika*, *61*, 313–322.
- Herbei, R. & Wegkamp, M. H. (2006). Classification with reject option. *Canad. J. Statist.*, *34*, 709–721.
- Huberty, C. J. (1994). *Applied Discriminant Analysis*. Wiley.
- Jeske, D. R., Liu, Z., Bent, E., & Borneman, J. (2007). Classification rules that include neutral zones and their application to microbial community profiling. *Comm. Statist. Theory Methods*, *36*, 1965–1980.
- Lee, E. K., Gallagher, R. J., & Patterson, D. A. (2003). A linear programming approach to discriminant analysis with a reserved-judgment region. *INFORMS J. Comput.*, *15*, 23–41.

- R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge.
- SAS Institute, Inc. (2008). *SAS/STAT 9.2 User's Guide*. Cary, NC.
- Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.
- Yu, H., Jeske, D., Ruegger, P., & Borneman, J. (2010). Neutral zone classifiers using a decision-theoretic approach with application to DNA array analyses. *J. Agric. Biol. Environ. Stat.*, 15, 474–490.

Appendix: The Pivot Algorithm

The pivot algorithm finds the line segments across a corner of the ternary plot triangle which maximize the number of points from the target population while satisfying constraints on the number of points from the other two populations. Figure 7 illustrates the situation. The goal is to find the line segment across corner 1 (that is, from a point on the 1-2 side to a point on the 1-3 side) which maximizes the number of points from H_1 (denoted by numerals) in the region R_1 while satisfying two constraints: the number of points from H_2 (lowercase letters) cannot exceed $l_2 = [\alpha_{21}n_2]$ and the number of points from H_3 (uppercase letters) cannot exceed $l_3 = [\alpha_{31}n_3]$ ($[\]$ denotes the greatest integer function). Because of the discreteness of the problem, there are an infinite number of possible solutions so we restrict the candidates to line segments which go through a data point and a vertex or which go through two data points. A solution is considered feasible only if the number of points on the interior of R_1 (not including points on the boundary) from H_2 is less than or equal to $l_2 - 1$ and the number of points from H_3 is less than or equal to $l_3 - 1$ (the -1 accounts for the fact that there may be points from H_2 and/or H_3 on the boundary).

The algorithm starts by considering a line segment with one endpoint at vertex 3 and the other endpoint at vertex 1. Move the latter endpoint along the 1-2 side toward vertex 2. As it moves (thus pivoting on vertex 3) the line segment will hit points from the training samples. Keep track of the number of points in R_1 from each of the three training samples. Eventually, the segment will hit the l_2^{th} point from H_2 or the l_3^{th} point from H_3 (step 1 in Figure 7). Pivot on the last point hit; that is, keep it fixed and continue moving the endpoint on the 1-2 side toward vertex 2. The endpoint on the 1-3 side will simultaneously start moving toward vertex 1. Points may be lost or gained as the line pivots. If a point is hit where further pivoting would cause a violation of the constraints, then stop and pivot on this new point. It's also possible that a point hit from H_2 or H_3 will be lost from R_1 on further pivoting. If that will cause the number of points from H_2

and H_3 in R_1 to both be strictly less than their limits, then stop and pivot on this new point. Continue until the endpoint on the 1-2 side reaches vertex 2. During this whole process, keep track of which candidate lines maximize the number of points from H_1 in R_1 . Figure 7 shows the sequence of pivots; the resulting candidate line segments are defined by the following pairs of points: i-1, i-F, i-J, and J-4. Each of these includes 5 points from population 1 in R_1 .

[Figure 7 here]

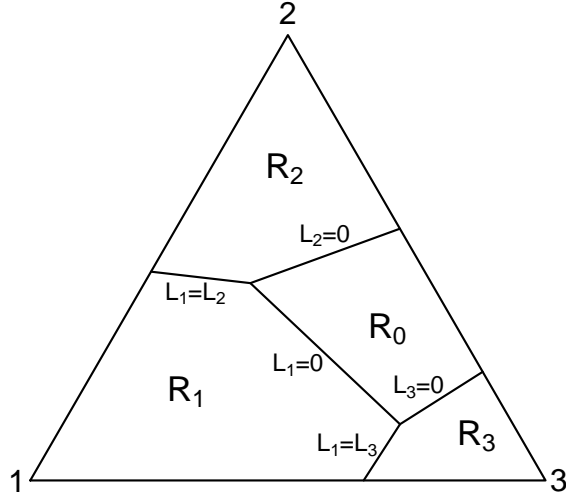


Figure 1: Anderson's solution with $\lambda_{21} = 0.7, \lambda_{31} = 0.2, \lambda_{12} = 0.5, \lambda_{32} = 1.3, \lambda_{13} = 1.2, \lambda_{23} = 3.1$

Table 1: Classification matrices for diabetes data: (a) forced discrimination by LDA; (b) pivot procedure constrained solution with all $\alpha_{ij} = 0.1$

(a)				(b)				
True Population	Classified Into			True Population	Classified Into			
	1	2	3		1	2	3	0
1	5	3	25	1	22	3	2	6
2	1	19	16	2	1	19	3	13
3	3	6	67	3	7	4	30	35

Table 2: Population configurations and Mahalanobis distances between means for equal covariance case.

Configuration	Means			Distance		
	Pop. 1	Pop. 2	Pop. 3	1-2	1-3	2-3
E1	(0, 0)	(-0.5, 0.866)	(0.5, 0.866)	1	1	1
E2	(0, 0)	(-1, 1.732)	(1, 1.732)	2	2	2
E3	(0, 0)	(-1, 0)	(1, 0)	1	1	2
E4	(0, 0)	(-0.5, 2.958)	(0.5, 2.958)	3	3	1
E5	(0, 0)	(0, 2)	(2.905, -0.75)	2	3	4

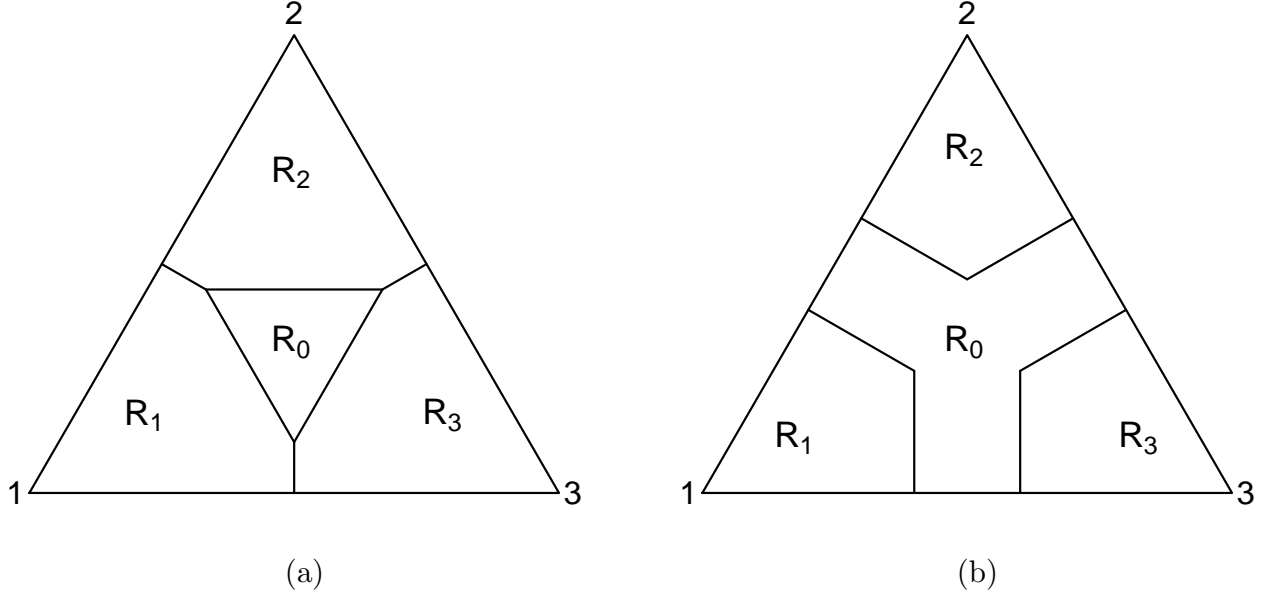


Figure 2: Example ternary plots for (a) p_{\max} procedure with threshold of 0.444 and (b) p_{dif} procedure with threshold of 0.2

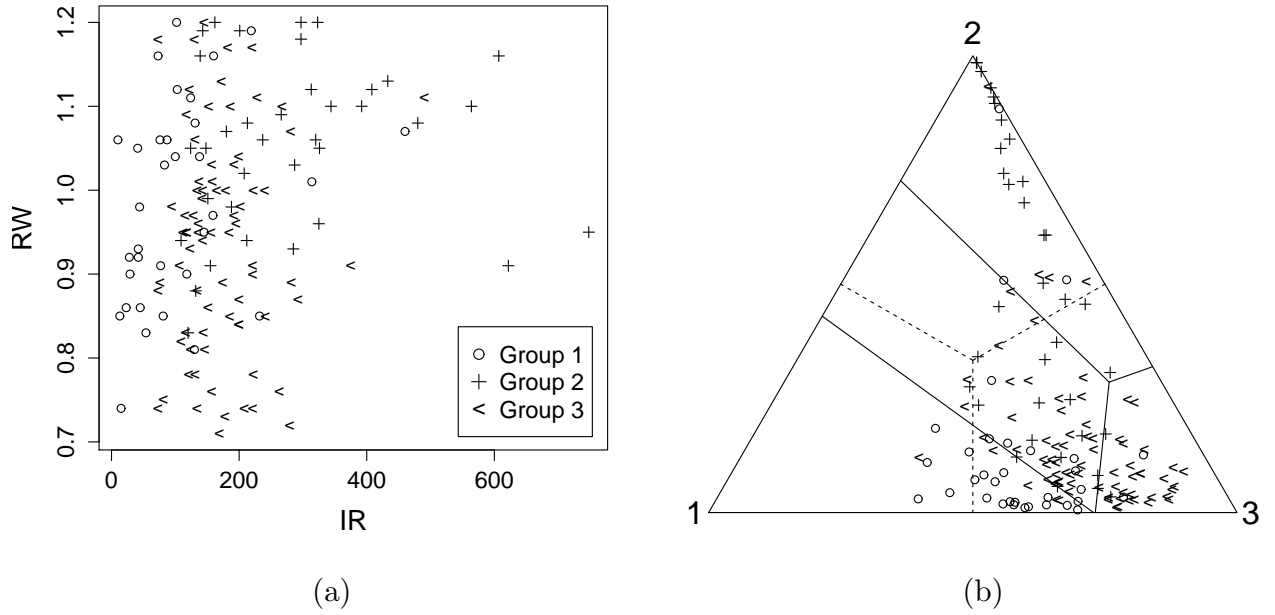


Figure 3: Diabetes data: (a) scatterplot of raw data; (b) ternary plot of posterior probabilities from LDA with proportional priors with boundaries for forced rule (dashed lines) and for optimal constrained rule estimated by pivot procedure (solid lines).

Table 3: Population configurations and distances between means for unequal covariance case.

Configuration	Means			Distance		
	Pop. 1	Pop. 2	Pop. 3	1-2	1-3	2-3
U1	(0, 0)	(-0.25, 0.75)	(0.25, 0.75)	1	1	1
U2	(0, 0)	(0, 0.791)	(1.99, 0.395)	1	2.57	4
U3	(0, 0)	(0, 0)	(2, 0)	0	2.53	4

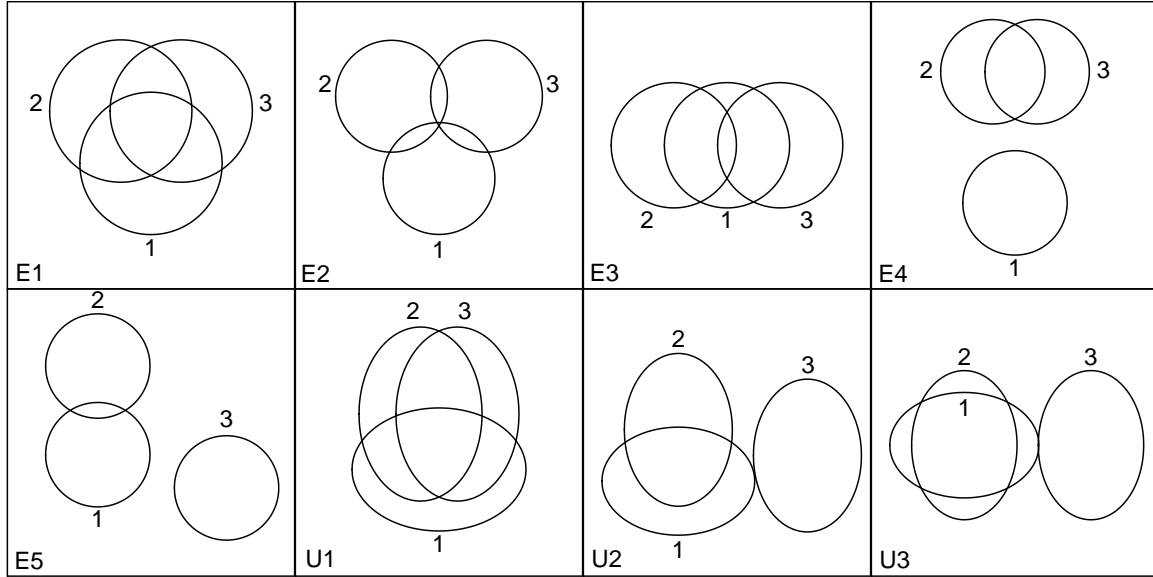


Figure 4: Mean configurations with 50% contours of normal densities

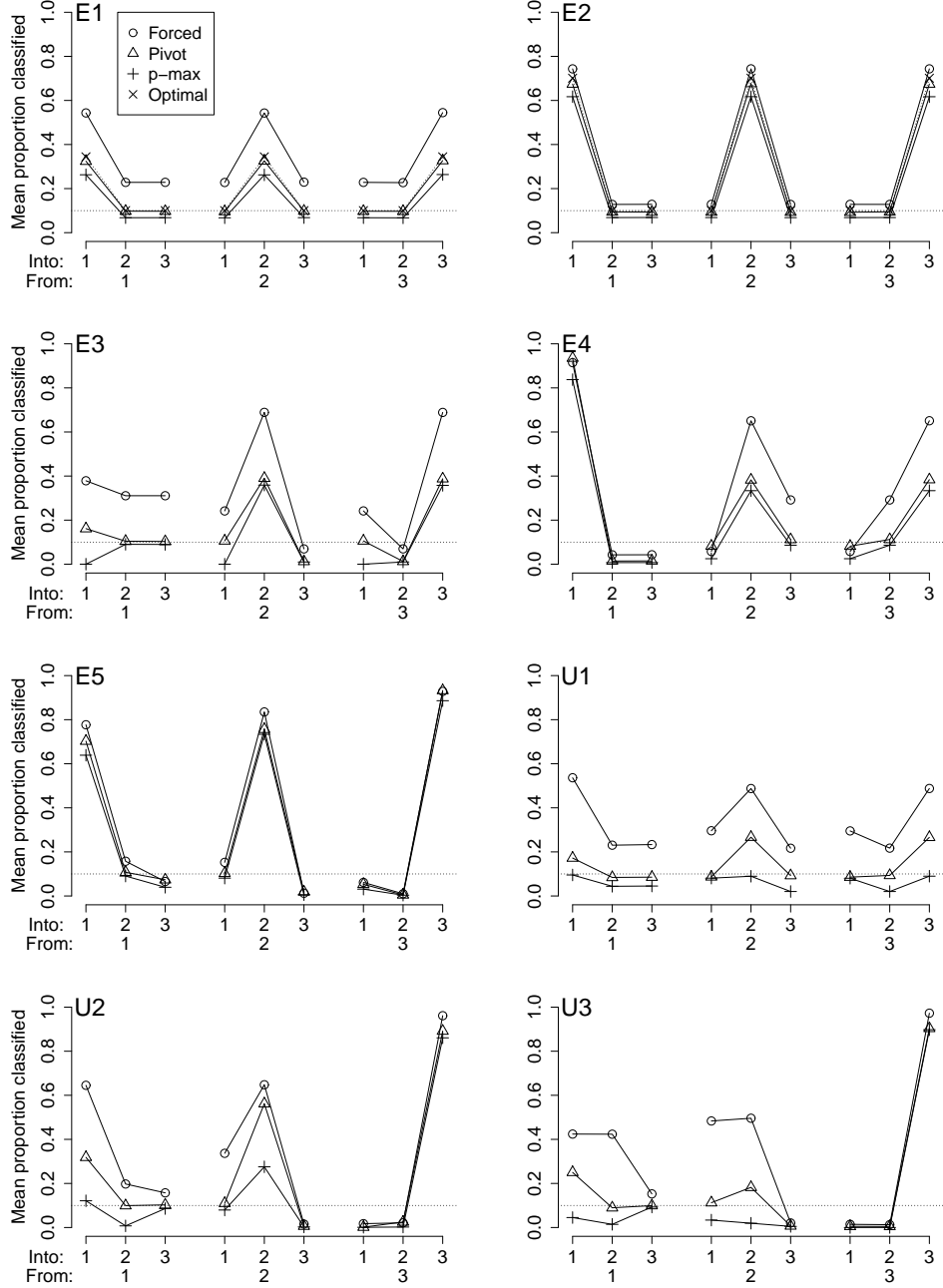


Figure 5: Simulation results for normal populations displaying mean classification rates across 2000 simulations. Rules are: LDA forced rule (\circ), pivot algorithm constrained rule (\triangle), p_{\max} constrained rule ($+$), optimal constrained rule (\times) for E1 and E2. Target misclassification rate is 0.1 (horizontal dotted line).

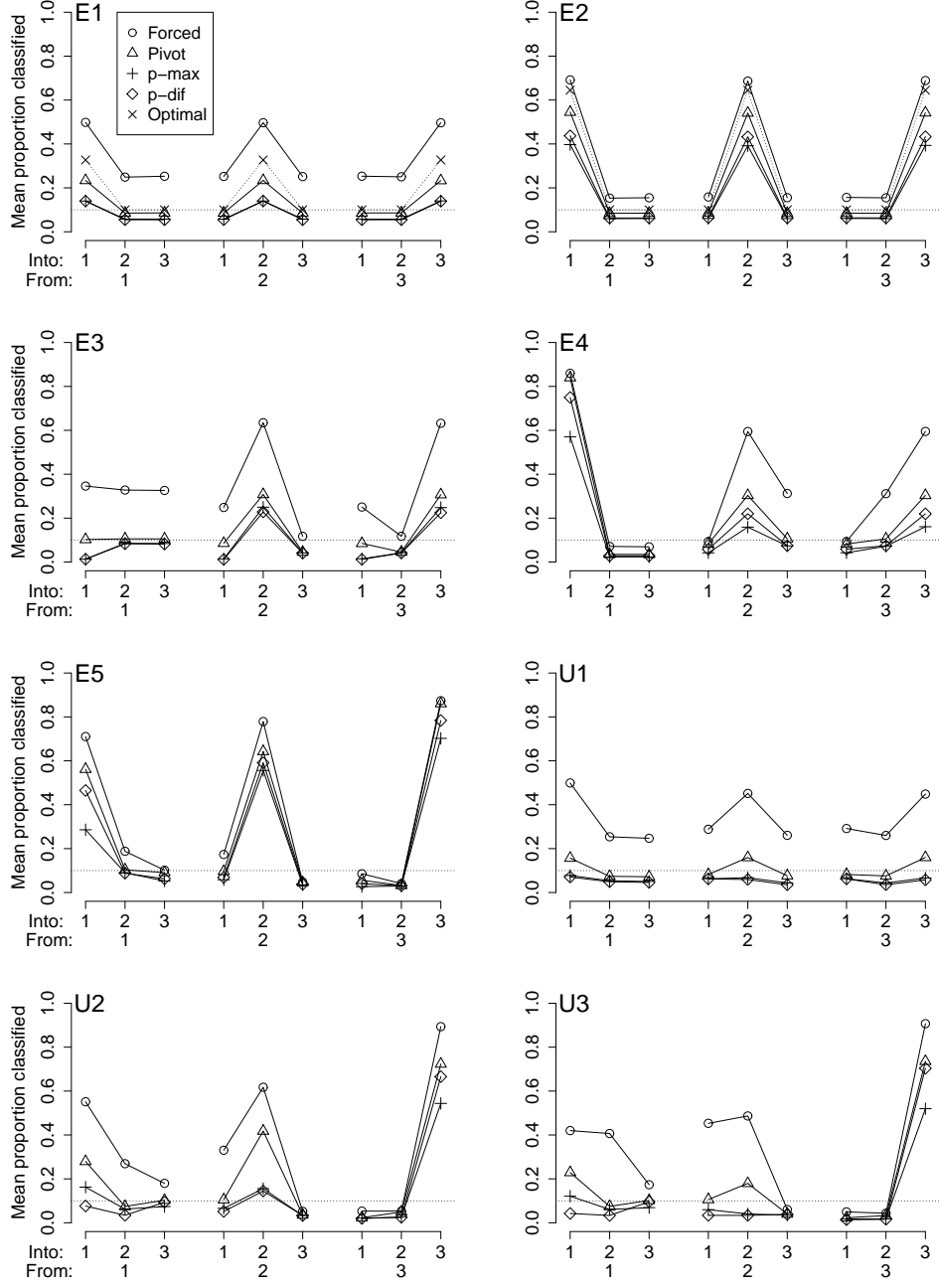


Figure 6: Simulation results for 10% contaminated normal populations displaying mean classification rates across 2000 simulations. Rules are: LDA forced rule (\circ), pivot algorithm constrained rule (\triangle), p_{\max} constrained rule (+), p_{dif} constrained rule (\diamond), optimal constrained rule (\times) for E1 and E2. Target misclassification rate is 0.1 (horizontal dotted line).

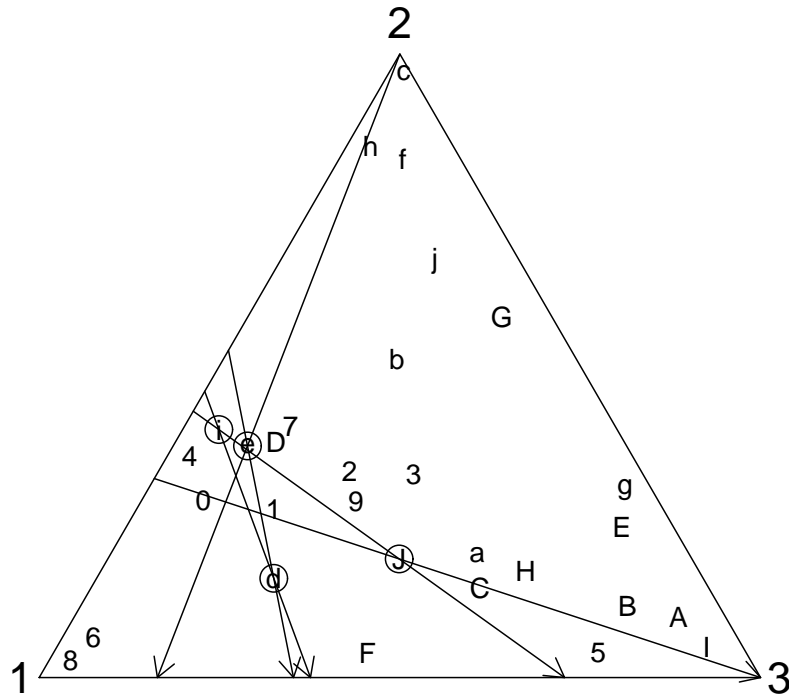


Figure 7: Sequence of steps for the Pivot Algorithm for finding candidates for $L_1 = 0$. Numerals, lowercase letters and uppercase letters represent points from populations 1, 2, and 3, respectively. Sample size is 10 for each population and the misclassification bounds are 0.2. The sequence of pivot points is e, d, i, J.