Virginia Commonwealth University
VCU Scholars Compass

Statistical Sciences and Operations Research Publications

Dept. of Statistical Sciences and Operations Research

2010

Analysis of the Consistency of a Mixed Integer Programming-based Multi-Category Constrained Discriminant Model

J. Paul Brooks
Virginia Commonwealth University, jpbrooks@vcu.edu

Eva K. Lee
Georgia Institute of Technology

Follow this and additional works at: http://scholarscompass.vcu.edu/ssor_pubs

 Part of the Computer Sciences Commons, and the Mathematics Commons

Copyright © Springer

Downloaded from
http://scholarscompass.vcu.edu/ssor_pubs/2

This Article is brought to you for free and open access by the Dept. of Statistical Sciences and Operations Research at VCU Scholars Compass. It has been accepted for inclusion in Statistical Sciences and Operations Research Publications by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

Analysis of the Consistency of a Mixed Integer Programming-based Multi-Category Constrained Discriminant Model

Paul Brooks[1] and Eva Lee[2*]

Department of Statistical Sciences and Operations Research, Virginia Commonwealth University[1]

School of Industrial and Systems Engineering, Georgia Institute of Technology[2]

evakylee@isye.gatech.edu, Corresponding Author*

May 21, 2008

1 Abstract

Classification is concerned with the development of rules for the allocation of observations to groups, and is a fundamental problem in machine learning. Much of previous work on classification models investigates two-group discrimination. Multi-category classification is less-often considered due to the tendency of generalizations of two-group models to produce misclassification rates that are higher than desirable. Indeed, producing "good" two-group classification rules is a challenging task for some applications, and producing good multi-category rules is generally more difficult. Additionally, even when the "optimal" classification rule is known, intergroup misclassification rates may be higher than tolerable for a given classification model. We investigate properties of a multi-category classification model that allows for the pre-specification of limits on intergroup misclassification rates. The mechanism by which the limits are satisfied is the use of a reserved judgment region, an artificial category into which observations are placed whose attributes do not sufficiently indicate membership to any particular group. The method is shown to be a consistent estimator of a classification rule with misclassification limits, and performance on simulated data is demonstrated.

Keywords: constrained discriminant analysis, mixed integer program, multi-category classification, consistency, reserved judgement.

2 Introduction

Classification is a fundamental machine learning task whereby rules are developed for the allocation of independent observations to groups. Classic examples of applications include medical diagnosis - the allocation of patients to disease classes based on symptoms and lab tests, and credit screening - the acceptance or rejection of credit applications based on applicant data. Data are collected concerning observations with known group membership. This *training data* is used to develop rules for the classification of future observations with unknown group membership.

Commonly-used methods for classification include linear discriminant functions (LDF), decision trees (CART, C4.5), support vector machines (SVM) and other math programming approaches, and artificial neural networks (ANN). These methods can be viewed as attempts at approximating a *Bayes optimal rule* for classification; that is, a rule which maximizes (minimizes) the total probability of correct classification (misclassification). Even if a Bayes optimal rule is known, intergroup misclassification rates may be higher than desired. For example, in a population that is mostly healthy, a Bayes optimal rule for medical diagnosis might misdiagnose sick patients as healthy in order to maximize total probability of correct diagnosis.

Anderson [2] presents the form of a classification rule that maximizes the total probability of correct classification subject to prespecified limits on misclassification probabilities. To satisfy the misclassification limits, an artificial group called the *reserved judgment group* is created for observations that do not sufficiently demonstrate their membership to any particular group. Allocation to the reserved judgment group can be interpreted as a signal to collect more information about an observation or an indication that a human expert should review the case individually. Classification with an option to reserve judgment on observations is called *constrained* or *partial* discrimination.

Gallagher, Lee, and Patterson [12] was the first to provide computational methods for estimating the parameters of an *Anderson optimal rule* based on the solution of integer programs. A recent paper [5] demonstrates that parameter estimation is an NP-Complete problem and develops solution methods for integer programming instances. In response to the computational complexity of the methods, which are based on the solution of integer programs, a linear programming approximation was developed [18]. The linear programming approximation has been applied successfully in drug delivery [17], DNA sequence analysis [10, 11], and cancer treatment [16] applications.

Other multi-category constrained discrimination methods include a stochastic approximation procedure due to Györfi et al. [13] and a histogram-based rule with tolerance regions proposed by Quesenberry and Gessaman [21]. Methods developed for two-group constrained discrimination include a ranking procedure introduced by Broffitt et al. [4] and enhanced by Beckman and Johnson [3]. Habbema et al. [14] introduce a two-group decision-theoretic method.

This work explores the ability of a constrained discrimination model in [12] to *generalize*; in other words, the ability of classification rules derived from the model to maximize the probability of correct classification of observations with unknown group membership while satisfying limits on intergroup misclassifications. To this end, we prove that the model is a *consistent* method for approximating an optimal classifier. A classification model is *consistent* if the performance of classifiers converge to an optimal classifier as the sample size increases. A proof of convergence for a model can help to describe the confidence that one may have in a classifier that is produced based on a given sample.

Vapnik-Chervonenkis Theory provides a means of evaluating classification rules that is independent of the distribution of any particular data set. In particular, the theory formalizes the notion of a tradeoff between the *complexity* of a classification rule and the ability of classifiers selected by a rule to generalize. The bounds provided by VC Theory are a measure of the confidence that we may have in a classifier; these bounds depend only on the sample size and the complexity of the classification rule. We will fuse Anderson's result with VC Theory to show that the constrained discrimination model in [12] is a consistent rule.

The remainder of this section provides an introduction to the Bayes optimal rule, the Anderson optimal rule, consistency, and VC Theory. Section 3 reviews the method for constrained discrimination presented by [12]. Section 4 contains a proof of the consistency of the method. Section 5 demonstrates the performance of the classifier on simulated data.

2.1 Bayes Optimal Rule

Let $(X, Y) \in \mathbb{R}^d \times \mathcal{G}$ be random variables where $\mathcal{G} = \{1, 2, \ldots, G\}$ is the set of groups and let $f(x)$ be the probability density function for X. The random variable Y is a discrete random variable defined by a conditional distribution $P\{Y = h|X = x\} = \pi_h \int f(x|h)dx$ for $h \in \mathcal{G}$ where π_h is the prior probability for membership to group h and $f(x|h)$ is the conditional group density function value for an observation x occurring given that it belongs to group h. A function $\phi : \mathbb{R}^d \to \mathcal{G}$ is a classifier. The probability of correct classification for the classifier is $P\{\phi(X) = Y\}$. The following development extends the treatment in [8] of the two-group case to multiple groups.

Let $\phi^*(x)$ be the *Bayes decision rule*, the function that assigns x to the group h for which $P\{Y = h|X = x\}$ is maximum, or equivalently, $\phi^*(x) = \arg\max_h P\{Y = h|X = x\}$.

Theorem 2.1. *The Bayes decision function ϕ^* is optimal for the optimization problem*[1]

$$\max_{\phi} P\{\phi(X) = Y\}$$

[1]Ordinarily, the Bayes decision rule is defined as that which minimizes the probability of misclassification. The definitions of *consistent* and *strongly consistent* are analogously defined.

Proof. For any classifier $\phi(x)$,

$$\begin{aligned} & P\{\phi^*(X) = Y|X = x\} - P\{\phi(X) = Y|X = x\} \\ = & \sum_{h=1}^{G} P\{Y = h, \phi^*(X) = h|X = x\} - \sum_{h=1}^{G} P\{Y = h, \phi(X) = h|X = x\} \\ = & \sum_{h=1}^{G} I_{\{\phi^*(x)=h\}} P\{Y = h|X = x\} - \sum_{h=1}^{G} I_{\{\phi(x)=h\}} P\{Y = h|X = x\} \\ \geq & \ 0 \end{aligned}$$

Integrating the first and third lines with respect to $f(x)dx$ gives

$$\begin{aligned} & \int (P\{\phi^*(X) = Y|X = x\} - P\{\phi(X) = Y|X = x\}) f(x)dx \\ = & \int P\{\phi^*(X) = Y|X = x\} f(x)dx - \int P\{\phi(X) = Y|X = x\} f(x)dx \\ = & \int P\{\phi^*(X) = Y, X = x\}dx - \int P\{\phi(X) = Y, X = x\}dx \\ = & P\{\phi^*(X) = Y\} - P\{\phi(X) = Y\} \\ \geq & \ 0 \end{aligned}$$

□

Let $\phi_n(X, D_n)$ be a classification rule selected based on an i.i.d. data set $D_n = \{(X_i, Y_i)\}_{i=1}^n$ of size n.

Definition 2.1. A decision rule is *consistent* if, as the sample size increases, the expected probability of correct classification converges to the probability of correct classification associated with the Bayes rule $P\{\phi^*(X) = Y\}$, or

$$\lim_{n\to\infty} EP\{\phi_n(X, D_n) = Y|D_n\} = P\{\phi^*(X) = Y\}$$

Definition 2.2. A decision rule is *strongly consistent* if

$$\lim_{n\to\infty} P\{\phi_n(X, D_n) = Y|D_n\} = P\{\phi^*(X) = Y\}, \text{ with probability 1}$$

Definition 2.3. A decision rule is *universally consistent* if the probability of correct classification converges to the probability of correct classification associated with Bayes rule for all possible distributions of X and Y.

2.2 Anderson Optimal Rule

Anderson [2] considers the problem of *constrained discrimination* which is classification with limits on misclassification rates through the use of a reserved judgment group. He proposes allocation of observations to groups based on "modified posterior probabilities" of the form

$$L_h(x) = \pi_h f(x|h) - \sum_{i\neq h} f(x|i)\lambda_{ih}$$

An observation is placed in the group for which the modified posterior probability is largest. After determining the prior probabilities and conditional group density functions that characterize the Bayes optimal solution, $G(G-1)$ λ_{ih} parameters are determined so that the misclassification limits are satisfied. The modified posterior probabilities are not probabilities, as they do not sum to 1 and are not guaranteed to be greater than 0. In fact, having all modified posterior probabilities less than 0 is the condition for an observation to be placed in the reserved judgment region. The $L_h(x)$ functions are called "probabilities" because they give an indication as to the likelihood of events; namely, the likelihood that an observation belongs to a group.

Throughout this work, we will assume that the prior probabilities and conditional group density functions (i.e., the Bayes optimal rule) are known, and that we seek to determine the optimal constrained discrimination rule

$$\phi^{\dagger}(x) = \arg\max_{0,h\in\mathcal{G}} \{0, \pi_h f(x|h) - \sum_{i\neq h} f(x|i)\lambda_{ih}^{\dagger}\}$$

which we will call the Anderson optimal rule.

Theorem 2.2. *(Theorem 1A in [2]) Given the prior probabilities and conditional group density function associated with the Bayes optimal rule, the Anderson optimal rule $\phi^{\dagger}$ is optimal for the optimization problem*

$$\max_{\phi} P\{\phi(X) = Y\}$$

$$\textit{subject to } P\{\phi(X) = g, Y = h\} \leq \alpha_{hg},\ g,\ h \in \mathcal{G},\ g \neq h$$

The α_{hg} constants are the pre-specified limit on the probability of misclassification of observations from group h to group g. Note that the presence of the reserved judgment region ensures that the optimization problem always has a feasible solution. Notions of consistency are generalized here to accommodate misclassification limits.

Definition 2.4. For a classification problem with pre-specified misclassification limits, a classification rule is *consistent* if

$$\lim_{n\to\infty} EP\{\phi_n(X, D_n) = Y|D_n\} = P\{\phi^{\dagger}(X) = Y\}$$

and

$$\lim_{n\to\infty} EP\{\phi_n(X, D_n) = g, Y = h|D_n\} \leq \alpha_{hg},\ h \in \mathcal{G},\ g \neq h$$

The second condition requires that every sequence of classifiers produced by the method satisfies the misclassification constraints in the limit.

Definition 2.5. A classification rule is *strongly consistent* if

$$\lim_{n\to\infty} P\{\phi_n(X, D_n) = Y|D_n\} = P\{\phi^{\dagger}(X) = Y\}, \text{ with probability 1}$$

and

$$\lim_{n\to\infty} P\{\phi_n(X, D_n) = g, Y = h|D_n\} \leq \alpha_{hg},\ h \in \mathcal{G},\ g \neq h, \text{ with probability 1}$$

2.3 VC Theory

Vapnik and Chervonenkis developed much of the theory concerning the selection of a classifier from a class $\mathcal{C}$ of classifiers based on empirical performance [22, 23, 24]. In particular, they investigated the selection of a classifier based on minimizing *empirical loss*. Given a data set, the empirical loss of a classifier is the proportion of observations that are misclassified. VC Theory provides a means of proving bounds on the difference between the misclassification probabilities of a classifier selected by minimizing empirical loss and the best possible classifier in the class $\mathcal{C}$.

VC Theory is based on a result due to Vapnik and Chervonenkis [22] on the convergence of frequencies to their probabilities. The notation and development presented here follow from [8]. Let Z_j be n i.i.d. d-dimensional random variables. For measurable sets $A \subset \mathbb{R}^d$, let $\nu(A) = P\{Z_j \in A\}$ and

$$\nu_n(A) = \frac{\sum_{j=1}^{n} I_{\{Z_j \in A\}}}{n}$$

Theorem 2.3. *([22], as stated in Theorem 12.5 in [8]) For any probability measure ν and class of sets $\mathcal{A}$, and for any n and $\epsilon > 0$*

$$P\left\{\sum_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon\right\} \leq 8s(\mathcal{A}, n)e^{-n\epsilon^2/8}$$

where $s(\mathcal{A}, n)$ is the shatter coefficient *for the collection of sets $\mathcal{A}$.*

Definition 2.6. The *shatter coefficient* $s(\mathcal{A}, n)$ of a class of sets $\mathcal{A}$ for a set of n points is the maximal number of different subsets of points that can be selected by sets in $\mathcal{A}$.

As with the discussion of the Bayes decision rule, we will evaluate classifiers in terms of their ability to maximize the probability of correct classification, or *benefit*, rather than minimize the probability of misclassification, or *loss*. Accordingly, an optimal classifier ϕ^* is one that maximizes benefit $B(\phi) = P\{\phi(X) = Y\}$.

Let ϕ_n be the classifier selected based on a sample of size n. In an effort to minimize *approximation error*, $\sup_{\phi\in\mathcal{C}} B(\phi) - B(\phi_n)$, ϕ_n is selected from $\mathcal{C}$ such that the empirical benefit,

$$\widehat{B}_n(\phi) = \frac{\sum_{j=1}^{n} I_{\{\phi(X_j)=Y_j\}}}{n}$$

is maximized.

Theorem 2.3 provides us with bounds on the difference between empirical benefit and the true benefit of a classifier. For a set of classifiers, let $\mathcal{A}$ be the collection of sets

$$\{(x,h) : \phi(x) = h\},\ \phi \in \mathcal{C}$$

Then define $s(\mathcal{C},n) = s(\mathcal{A},n)$ as the n^{th} shatter coefficient of the set of classifiers $\mathcal{C}$. The shatter coefficient is a measure of the complexity of the set $\mathcal{C}$. Coupled with the following inequality (Lemma 8.2 in [8])

$$\sup_{\phi\in\mathcal{C}} B(\phi) - B(\phi_n) \leq 2 \sup_{\phi\in\mathcal{C}} |\widehat{B}_n(\phi) - B(\phi)|$$

a bound on approximation error is derived:

Corollary 2.4. *(Theorem 12.6 in [8])*

$$P\left\{\sup_{\phi\in\mathcal{C}} |\widehat{B}_n(\phi) - B(\phi)| > \epsilon\right\} \leq 8s(\mathcal{C},n)e^{-n\epsilon^2/8}$$

and therefore,

$$P\left\{\sup_{\phi\in\mathcal{C}} B(\phi) - B(\phi_n) > \epsilon\right\} \leq 8s(\mathcal{C},n)e^{-n\epsilon^2/32}$$

3 Estimating the Anderson Optimal Rule

Given a sample of size n, the problem of maximizing empirical benefit subject to limits on misclassification rates can be formulated as follows

$$\max_{\phi} \sum_{j=1}^{n} I_{\{\phi(X_j)=Y_j\}}$$

subject to

$$\begin{aligned}
\sum_{j=1}^{n} \frac{I_{\{\phi(X_j)=g, Y=h\}}}{n} &\leq \alpha_{hg} && g,\ h \in \mathcal{G},\ g \neq h \\
\phi(X_j) &= \arg\max_{0,h\in\mathcal{G}} \{0, \pi_h f(X_j|h) - \sum_{i\neq h} f(X_j|i)\lambda_{ih}\} && j = 1,\ldots,n \\
\lambda_{ih} &\geq 0 && i,\ h \in \mathcal{G},\ i \neq h
\end{aligned}$$

The problem can be formulated as a nonlinear mixed-integer program. Let

$$u_{hj} = \begin{cases} 1 & \text{if } L_h = \max_{h'}\{L_{h'j}, 0\} \\ 0 & o.w. \end{cases}$$

The u_{hj} variables indicate the allocation of observation X_j to the group h for which the (estimated) modified posterior probability is largest. The nonlinear mixed-integer program $(DAMIP)$ is then

$$\max \sum_{j=1}^{n} u_{Y_j j} \qquad (DAMIP)$$

subject to

$$\begin{aligned}
L_{hj} &= \pi_h f(X_j|h) - \sum_{i \neq h} f(X_j|i)\lambda_{ih} && j = 1, \ldots, n;\ h \in \mathcal{G} \\
u_{hj} &= \begin{cases} 1 & \text{if } L_h = \max_{h'}\{L_{h'j}, 0\} \\ 0 & o.w. \end{cases} && j = 1, \ldots, n;\ h \in \mathcal{G} \\
\sum_{0, h \in \mathcal{G}}^{n} u_{hj} &= 1 && j = 1, \ldots, n \\
\sum_{j: Y_j = h} u_{gj} &\leq \lfloor \alpha_{hg} n_h \rfloor && h,\ g \in \mathcal{G}
\end{aligned}$$

$$-\infty < L_{hj} < \infty,\ \lambda_{ih} \geq 0$$

The nonlinearity arises in the definition of the u_{hj} variables, where the maximum of a set of linear functions must be determined. When appropriately chosen constants are introduced, the problem can be converted into an equivalent linear mixed integer program ([12] Model 3, and [5]). The computational complexity and solution methods for solving ($DAMIP$) are discussed in [12] and [5].

4 Consistency of the Constrained Discrimination Model

We now combine Anderson's result [2] with VC Theory to show that a classifier selected by solving (DAMIP) is consistent in the sense that it converges to an Anderson optimal rule, given that we know a Bayes optimal rule. In estimating the parameters of an Anderson optimal rule, we may think of the posterior probabilities and conditional group density function values as input data and the solution of (DAMIP) as the selection of a classifier based on sample data.

Let ϕ_n be the classifier selected by solving (DAMIP) for a sample of size n and $\phi^\dagger$ be the Anderson optimal classifier. Let $\widehat{B}_n(\phi)$ be the number of correctly classified observations in a sample of size n by classifier ϕ; similarly, let $\widehat{M}_n^{hg}(\phi)$ be the number of observations from group h misclassified as belonging to group g by ϕ. Let $B(\phi) = P\{\phi(X) = Y\}$ and $M^{hg}(\phi) = P\{\phi(X) = g, Y = h\}$. The following lemma demonstrates that the approximation error is zero.

Lemma 4.1.

$$B(\phi^\dagger) - B(\phi_n) = \sup_{\phi \in \mathcal{C}} B(\phi) - B(\phi_n)$$

where $\mathcal{C}$ is the set of classifiers that allocate observations according to modified posterior probabilities.

Proof.

$$\begin{aligned}
B(\phi^\dagger) - B(\phi_n) &= (B(\phi^\dagger) - \sup_{\phi \in \mathcal{C}} B(\phi)) + (\sup_{\phi \in \mathcal{C}} B(\phi) - B(\phi_n)) \\
&= \sup_{\phi \in \mathcal{C}} B(\phi) - B(\phi_n)
\end{aligned}$$

The result follows directly from Theorem 2.2 which implies that allocating observations to groups according to optimally-calculated modified posterior probabilities will maximize the probability of correct classification subject to pre-specified limits on misclassification probabilities. □

Lemma 4.2. *If $\mathcal{C}$ is the set of classifiers that partition the feature space using t hyperplanes, then*

$$s(\mathcal{C}, n) \leq (2(n-1)^d + 2)^t$$

Proof. According to Corollary 13.1 of [8], an upper bound for the shatter coefficient for the class $\mathcal{C}$ of half spaces $\{x : ax \geq b\}$ is

$$s(\mathcal{C}, n) \leq 2(n-1)^d + 2$$

where d is the dimension of the feature space. Theorem 13.5(iii) of [8] implies that if a class $\mathcal{C}$ is derived from the intersection of sets from two classes, $\mathcal{C}_1$ and $\mathcal{C}_2$, then an upper bound on the shatter coefficient for $\mathcal{C}$ is

$$s(\mathcal{C}, n) \leq s(\mathcal{C}_1, n)s(\mathcal{C}_2, n)$$

These two results together imply that the shatter coefficient for the class $\mathcal{C}$ consisting of classifiers that partition the feature space using t hyperplanes has an upper bound

$$s(\mathcal{C}, n) \leq (2(n-1)^d + 2)^t$$

□

Lemma 4.3. *For any pair of groups g and h,*

$$M^{hg}(\phi_n) - \alpha_{hg} \leq \sup_{\phi \in \mathcal{C}} |M^{hg}(\phi) - \widehat{M}_n^{hg}(\phi)|$$

Proof.

$$\begin{aligned} M^{hg}(\phi_n) - \alpha_{hg} &= M^{hg}(\phi_n) - \widehat{M}_n^{hg}(\phi_n) + \widehat{M}_n^{hg}(\phi_n) - \alpha_{hg} \\ &\leq M^{hg}(\phi_n) - \widehat{M}_n^{hg}(\phi_n) \\ &\leq \sup_{\phi \in \mathcal{C}} |M^{hg}(\phi) - \widehat{M}_n^{hg}(\phi)| \end{aligned}$$

The first inequality is due to the constraints in (DAMIP) that restrict the selection of ϕ_n. □

Theorem 4.4. *Given prior probabilities π_h and conditional group density functions $f(x|h)$, allocation according to modified posterior probabilities defined by the solution to $(DAMIP)$ is a universally strongly consistent method for classification.*

Proof. Considering the prior probabilities π_h and conditional group density function values $f(X_j|h)$ as input data for G-group discrimination, solving $(DAMIP)$ is equivalent to selecting a classifier from a set $\mathcal{C}$ defined by the intersection of $G+1$ hyperplanes $\{0, \pi_h f(x|h) - \sum_{i \neq h} f(x|i)\lambda_{ih}\}$. The dimension of the feature space is $G(G-1)$, and using Lemma 4.2 gives the following upper bound on the shatter coefficient

$$s(\mathcal{C}, n) \leq (2(n-1)^{G(G-1)} + 2)^{G+1}$$

This bound, together with Corollary 2.4, imply that

$$P\left\{\sup_{\phi \in \mathcal{C}} B(\phi) - B(\phi_n) > \epsilon\right\} \leq 8(2(n-1)^{G(G-1)} + 1)^{G+1} e^{-n\epsilon^2/32}$$

and thus the estimation error converges uniformly to zero. As Lemma 4.1 shows, the approximation error is zero; therefore, the benefit of an empirically selected classifier converges uniformly to that of an Anderson optimal rule.

Let $\mathcal{A}^c$ be the following collection of sets

$$\{(x, h) : \phi(x) \neq h\}, \ \phi \in \mathcal{C}$$

then $s(\mathcal{A}^c, n) = s(\mathcal{C}, n)$ (see also Theorem 13.5(ii) in [8]). Then Lemma 4.3 and Theorem 2.3 imply that

$$\begin{aligned} P\left\{M^{hg}(\phi_n) - \alpha_{hg} > \epsilon\right\} &\leq P\left\{\sup_{\phi \in \mathcal{C}} |M^{hg}(\phi) - \widehat{M}_n^{hg}(\phi)| > \epsilon\right\} \\ &\leq 8s(\mathcal{C}, n)e^{-n\epsilon^2/8} \end{aligned}$$

Therefore, the intergroup misclassification rates for classifiers selected by solving $(DAMIP)$ converge uniformly to satisfy the misclassification limits α_{hg}. The convergence results for the benefit $B(\phi)$ and misclassification probabilities $M^{hg}(\phi)$ depend only on the structure of the set of classifiers and not on the distribution of the data. The constrained discrimination model defined by solving $(DAMIP)$ is universally strongly consistent method for estimating the Anderson optimal rule. □

5 Simulation Results

The constrained method is tested on simulated data generated from bivariate normal distributions. The simulations in Section 5.1 compare the performance of the constrained method to various standard methods for classification. Section 5.2 investigates the effect of using equal prior probabilities versus unequal priors, and Section 5.3 considers the effect of varying the proportions of training observations from each group.

5.1 Comparison to Standard Methods

The data for the first simulation follows the procedure in [18]. Briefly, the data are generated from bivariate (2 attributes) normal and contaminated normal distributions with different mean and variance configurations. For each run, 40 training observations from each of 3 groups are generated. The rules are then tested on 1000 test observations from each group. The process is repeated 200 times for each of 8 mean-variance configurations (Table 1). The configurations represent different ways of arranging three groups of data in the attribute space; i.e., two groups close together and far from the other ($E4$), all three groups close together ($E1$), all three groups far apart ($E2$). The means and covariances are chosen such that the Mahalanobis distance is approximately 1 for groups close together and approximately 3 for groups far apart. The Mahalanobis distance between groups i and j is

$$(\mu_i - \mu_j)^T V^{-1} (\mu_i - \mu_j)$$

where μ_i, μ_j are the group means and V is the covariance matrix. For configurations $E1 - E5$, the 2×2 identity matrix is used as the covariance matrix for each group. For configurations $U1 - U3$, different covariance matrices are used for different groups. In $U1 - U3$, the first group's covariance matrix is diagonal $(1, 0.25)$ and the second and third groups' covariance matrices are diagonal $(0.25, 1)$. When calculating the Mahalanobis distance for configurations, $U1 - U3$, V is computed as the average of V_i and V_j, the covariance matrices for groups i and j. For the data from contaminated normal distributions, the same configurations are used, with the exception that 10% of the data is derived from a normal distribution with the covariance matrix multiplied by 100.

The proportions of training observations from each group are used as estimates for the prior probabilities π_h. Therefore, for this set of simulations, $\pi = [1/3, 1/3, 1/3]$. The misclassification rate is calculated as the number of misclassified observations divided by the total number of observations. The non-reservation rate is the proportion of observations not placed in the reserved judgment group.

Splus 6.0 for Linux/Unix and ILOG CPLEX Callable Library 8.1 are used for generating simulated data and solving mixed-integer programs, respectively. The data are generated on Sun 280R's, each with 2x900MHz UltraSparc-III-Cu CPU's and 2 GB RAM. The mixed-integer programs are solved on machines with 2×2.4GHz Intel Xeon processors and 2GB RAM. If a provably-optimal solution is not obtained after 7200 CPU seconds, the best-known integer feasible solution is used.

The performance of the constrained method is compared to linear discriminant functions (LDF), quadratic discriminant functions (QDF), classification trees (CART), and support vector machines (SVM). LDF and QDF can be derived by considering the normal distribution as the form for the conditional group density functions $f(x|h)$ (also known as the likelihood function). In the two-group case, LDF is equivalent to Fisher's

Table 1: The mean-variance configurations for the normal distributions used in the simulation study. Configurations $E1-E5$ use equal covariance matrices and configurations $U1-U3$ use unequal covariance matrices. The values of the covariance matrices are given in the text.

Config.	Means			Mahalanobis Distances		
	Group 1	Group 2	Group 3	d(1,2)	d(1,3)	d(2,3)
$E1$	(0,0)	(-0.500, 0.868)	(0.500, 0.868)	1	1	1
$E2$	(0,0)	(-1.500, 2.598)	(1.500, 2.598)	3	3	3
$E3$	(0,0)	(-1.000, 0.000)	(1.000, 0.000)	1	1	2
$E4$	(0,0)	(-0.500, 2.968)	(0.500, 2.958)	3	3	1
$E5$	(0,0)	(0.000, 2.000)	(2.905, -0.750)	2	3	4
$U1$	(0,0)	(-0.250, 0.750)	(0.250, 0.750)	1	1	1
$U2$	(0,0)	(0.000, 0.791)	(1.990, 0.395)	1	2.6	4
$U3$	(0,0)	(0.000, 0.000)	(2.000, 0.000)	0	2.5	4

linear discriminant analysis (LDA).

LDF and QDF are carried out using the lda() and qda() functions, respectively, in the MASS library of R 2.4.1. CART is performed using the rpart() function of the rpart library of R 2.4.1. Support vector machines with a linear kernel and radial basis function (rbf) kernel are trained and tested using SVM^{multiclass} [15], an extension of SVM^{light}[15]. The values 0.001, 0.01, 0.1, 1.0, and 10.0 are tested for the parameter C determining the relative emphasis of margin and error; the results reported here use $C = 0.1$ for the linear kernel and $C = 0.01$ for the rbf kernel. The values 0.5, 1.0, and 2.0 are tested for the parameter g determining the width of the rbf kernel; the results reported here use $g = 1.0$. The likelihood values for the constrained model are generated using LDF.

Figure 1 contains the results of the simulations for data sampled from normal and contaminated normal distributions. For the constrained discrimination method, the misclassification rates of test observations increases as the misclassification limits for training observations is raised. The misclassification rates on test observations is slightly higher than the limits imposed on the training data, but one can see that tight control of the misclassification rates is possible.

Compared with standard methods, the misclassification rates using the constrained method are consistently lower when limits of 0, 5, and 15% are placed on the training set. The low misclassification rates are made possible by higher rates of placement in the reserved judgment group. With limits of 15%, misclassification rates are as low or lower than all of the standard methods and 65-100% of the test observations are classified.

The constrained method with 50% limits is comparable to to LDF, the base classifier, and rarely makes use of the reserved judgment region.

The average misclassification rate of the standard methods is 9.6% higher than for the data derived from contaminated normal distributions than for data derived from non-contaminated normal distributions. In contrast, the misclassification rates of the constrained model rise by an average of 2.1%. The constrained method compensates for the contaminated data by increasing the rate of placement in the reserved judgment region by an average of 10.3%.

5.2 Sensitivity to Prior Probabilities

To investigate the the effect of prior probabilities on classification performance, we generate data from bivariate normal distributions $E1$ and $U1$ as described in Table 1. Simulations are conducted as described in Section 5.1 with the exception that the training data consists of 100 observations from group 1, 100 observations from group 2, and 10 observations from group 3. This setup reflects a situation where observations from group 3 are rare; for example, groups 1 and 2 could represent healthy populations in a medical diagnosis setting while group 3 represents a population with a rare disease. The constrained discrimination method is compared to standard methods for which prior probabilities can be specified: LDF, QDF, and CART.

Figure 2 contains classification matrices for the constrained method and standard methods where the prior probabilities are specified as the proportion of observations from each group. In this case, the prior probabilities are $\pi = [100/210, 100/210, 10/210]$. The standard methods tend to place all test observations in groups 1 and 2, and misclassify 97.9-99.9% of test observations from group 3. The constrained method utilizes the reserved judgment region to lower misclassification rates for observations from group 3 to 5.4-36.4% when the misclassification limits are set between 0 and 15% on the training data. The correct classification rates for groups 1 and 2 are lower for the constrained method, as approximately 30-90% of these observations are placed in the reserved judgment region.

Figure 3 contains classification matrices for simulations where the prior probabilities are equal ($\pi = [1/3, 1/3, 1/3]$). The standard methods demonstrate dramatic improvement in the ability to correctly classify observations from group 3, with misclassification rates of 52.9-64.0%. This improvement is accompanied by a drop in the correct classification rates for groups 1 and 2 of approximately 10-15%. In contrast, the constrained method produces classification matrices that are almost identical to those obtained using unequal prior probabilities. Misclassification rates and correct classification rates change by less than 7.0% for each group. These simulations provide evidence that the constrained method is insensitive to the specification of prior probabilities, yet is capable of reducing misclassification rates when the number of training observations from each group is different.

5.3 Effect of Varying Group Sizes

To further investigate the effect of underrepresented groups in the training data, we again use data generated from $E1$ and $U1$ distributions as described in Table 1. We seek the relationship between the proportions of training observations from each group and the classification performance of the constrained model. Two sets of 3-group simulations are conducted as described in Section 5.1. The first set of simulations consists of training data with a larger number of observations from group 1 than from groups 2 and 3. The second set of simulations consists of training data with a larger number of observations from groups 1 and 2 than from group 3. The constrained method is compared to LDF, the base classifier.

Figures 4 and 5 contain the results of the simulations. The training data for the simulations of Figure 4 is comprised of 100 observations from group 1, and varying numbers of observations from groups 2 and 3. LDF tends to place most test observations in group 1, misclassifying large numbers of test observations from groups 2 and 3. As relative proportion of observations from groups 2 and 3 is increased, the misclassification rates for groups 2 and 3 drops, while the misclassification rate for group 1 increases. The constrained method, particularly with misclassification limits of 15% or less, makes use of the reserved judgment region to maintain stable misclassification rates with respect to changing proportions of training observations. The classification performance of the constrained method with very few training observations from groups 2 and 3 is very similar to the performance with larger numbers of observations.

The results in Figure 5 further support the claim that the constrained method generates stable classification rules regardless of the proportions of training observations from each group. The training data is comprise of 100 observations from each of groups 1 and 2, and varying numbers of observations from groups 2 and 3. Again, LDF gives preference to the groups with larger numbers of training observations, misclassifying many of the group 3 test observations. The constrained method is able to control misclassification rates for each of the groups. The plot of non-reservation rates for data generated from distribution $U1$ in Figures 4b and 5b shows that more observations from group 1 are placed in the reserved judgment region. This effect is due to the fact that the group 1 observations are generated using a different covariance matrix from the matrix used for groups 2 and 3.

6 Performance on Real-World Data

The constrained method based on solving ($DAMIP$) is compared to standard methods on nine real-world data sets. Eight of the data sets are from the UCI machine learning database repository [1]; the remaining data set, *FNlnVN* consists of cell motility data [25]. Five of the data sets (*wine*, *iris*, *new-thyroid*, *sepal*, and *FNlnVN*) are 3-group data sets, and four are 5-group data sets (*va*, *switzerland*, *hungarian*, *cleveland* [7]).

Table 2: Misclassification rates on real-world data sets.

	CD 0%	CD 5%	CD 15%	LDF	QDF	CART	SVMlinear	SVMrbf
FNlnVN	14.3	21.4	34.3	51.4	51.4	38.6	50.0	71.4
iris	1.3	3.3	2.7	2.0	2.7	4.0	3.3	5.5
new-thyroid	1.4	4.3	5.7	8.6	3.8	5.7	2.4	5.7
sepal	2.7	7.3	12.0	20.7	22.7	29.3	35.3	30.0
wine	1.8	1.8	1.8	1.8	1.2	8.2	3.5	35.3
cleveland	5.9	12.8	20.0	41.7	43.8	43.1	43.8	48.6
hungarian	8.1	13.5	20.4	36.5	40.4	38.5	33.8	41.2
switzerland	20.0	37.0	59.0	60.0	69.0	65.0	59.0	77.0
va	16.9	23.1	43.1	73.8	76.2	71.5	70.0	74.6

The data set *sepal* is derived from *iris* as described in [12].

Computation of classification rules for each method is carried out as described in Section 5. The rules are evaluated using ten-fold cross-validation. In ten-fold cross-validation, the data set is randomly partitioned into ten sets. Ten rules are generated for each model, witholding one of the partitions of data on each iteration. The witheld partition is used as a test set, and the cumulative performance on all test sets is reported.

The results are contained in Tables 2 and 3. As was seen on the simulated data, constrained discrimination via solution of ($DAMIP$) provides lower misclassification rates than other methods with the misclassification rates are restricted. The price paid for decreased misclassification is placement of observations in the reserved judgement region. For the "easier" data sets, *iris*, *new-thyroid*, and *wine*, misclassification rates are controlled by the constrained method (CD), with little use of the reserved judgment region. For the more "difficult" datasets, 12-35% of observations are placed in the reserved judgment region with 15% misclassification limit. The use of the reserved judgment region increases as misclassification limits are lowered, with up to 83.1% of observations placed in the reserved judgment region.

7 Conclusion and Discussion

We have shown that a method for constrained discrimination is strongly universally consistent, given that the Bayes optimal rule for classification is known. An interesting open question is to investigate the conditions under which the method is consistent when various methods for estimating the Bayes optimal rule are used. Several consistent methods for estimating the Bayes optimal rule exist, but due to the so-called No Free

Table 3: Non-reservation rates rates for constrained discrimination (CD) on real-world data sets.

	CD 0%	CD 5%	CD 15%
FNlnVN	40.0	54.3	72.9
iris	96.7	100.0	100.0
new-thyroid	92.4	99.0	100.0
sepal	51.3	64.0	75.3
wine	100.0	100.0	100.0
cleveland	25.2	36.6	49.7
hungarian	16.9	48.1	65.8
switzerland	30.0	56.0	88.0
va	26.9	35.4	65.4

Table 4: Correct classification rates on real-world data sets.

	CD 0%	CD 5%	CD 15%	LDF	QDF	CART	SVMlinear	SVMrbf
FNlnVN	64.3	60.6	52.9	51.4	48.6	61.4	50.0	28.6
iris	98.7	96.7	97.3	2.0	97.3	96.0	96.7	94.5
new-thyroid	98.5	95.7	94.3	8.6	96.2	94.3	97.6	94.3
sepal	94.7	88.6	84.1	20.7	77.3	70.7	64.7	70.0
wine	98.2	98.2	98.2	1.8	98.8	91.8	96.5	64.7
cleveland	76.6	65.0	59.8	41.7	56.2	56.9	56.2	51.4
hungarian	52.1	71.9	69.0	36.5	59.6	61.5	66.2	58.8
switzerland	33.3	33.9	33.0	60.0	31.0	35.0	41.0	23.0
va	37.2	34.7	34.1	73.8	23.8	28.5	30.0	25.4

Lunch Theorem ([6], Theorem 9.1 in [9], Theorem 7.2 in [8]), there will always exist distributions of the data for which convergence is arbitrarily slow. Uniform convergence to an optimal rule with a guaranteed rate of convergence requires assumptions on the distributions of X and Y.

The constrained method that we investigate here involves the solution of a mixed-integer program where the number of correctly classified observations is maximized. Support vector machines (SVMs) are another math programming-based classification algorithm. The objective for SVMs attempts to minimize error while maximizing the margin between groups in the training data. One might suggest adding a quadratic term to the constrained method, serving the role of the margin term in a SVM formulation, as a regularization term to prevent overfitting. However, the classification performance of the constrained method on simulated data suggests that the use of the reserved judgment region provides sufficient stability to the classification rules. The constrained method is able to effectively control error when differing numbers of training observations are available from each group. Further Lee has demonstrated the success of this MIP-based constrained discriminant approach [12, 18] to a broad class of biomedical applications [19, 20].

A limitation of the constrained method is the need to establish the tradeoff between the costs of misclassification and placement in the reserved judgment group. Lee, Gallagher, and Patterson [18] have provided some insight into the effect of weights on classification rates verus the placement into the reserved judgement region. We expect this tradeoff to be dependent on each particular application, and likely requires testing several sets of misclassification limits on the training data. Further simulation tests are necessary in order to characterize the relationship between misclassification and reservation rates; in particular, tests with data generated from non-normal distributions are of interest.

Constrained discrimination provides a means of controlling intergroup misclassification rates. The introduction of a reserved judgment group is a natural way to handle observations for which the model can effectively say, "I don't know yet," and signal the need to collect more data.

Acknowledgement

This research is partially supported by the National Science Foundation.

References

[1] A. Asuncion and D.J. Newman. UCI Machine Learning Repository [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 2007.

[2] J.A. Anderson. Constrained discrimination between k populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31:123–139, 1969.

[3] R.J. Beckman and M.E. Johnson. A ranking procedure for partial discriminant analysis. *Journal of the American Statistical Association*, 76:671–675, 2006.

[4] J.D. Broffitt, R.H. Randles, and R.V. Hogg. Distribution-free partial discriminant analysis. *Journal of the American Statistical Association*, 71:934–939, 1976.

[5] J.P. Brooks and E.K. Lee. Solving an MIP formulation of a classification model with limits on misclassification probabilities. Working paper.

[6] T. Cover. Rates of convergence for nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on System Sciences*, pages 413–415, Honolulu, 1968.

[7] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64:304–310, 1989.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[9] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2001.

[10] F.A. Feltus, E.K. Lee, J.F. Costello, C. Plass, and P.M. Vertino. Predicting aberrant CpG island methylation. *Proceedings of the National Academy of Sciences*, 100:12253–12258, 2003.

[11] F.A. Feltus, E.K. Lee, J.F. Costello, C. Plass, and P.M. Vertino. Dna motifs associated with aberrant CpG island methylation. *Genomics*, 87:572–579, 2006.

[12] R.J. Gallagher, E.K. Lee, and D.A. Patterson. Constrained discriminant analysis via 0/1 mixed integer programming. *Annals of Operations Research*, 74:65–88, 1997.

[13] L. Györfi, Z. Györfi, and I. Vajda. Bayesian decision with rejection. *Problems of Control and Information Theory*, 8:445–452, 1979.

[14] J.D.F. Habbema, J. Hermans, and A.T. Van Der Burgt. Cases of doubt in allocation problems. *Biometrika*, 61:313–324, 1974.

[15] T. Joachims. *Advances in Kernel Methods - Support Vector Learning, B*, chapter Making large-scale SVM learning practical. B. Schölkopf, C. Burges, and A. Smola (ed.). MIT-Press, 1999.

[16] E.K. Lee, A.Y.C. Fung, J.P. Brooks, and M. Zaider. Automated planning volume definition in soft-tissue sarcoma adjuvant brachytherapy. *Biology in Physics and Medicine*, 47:1891–1910, 2002.

[17] E.K. Lee, R.J. Gallagher, A.M. Campbell, and M.R. Prausnitz. Prediction of ultrasound-mediated disruption of cell membranes using machine learning techniques and statistical analysis of acoustic spectra. *IEEE Transactions on Biomedical Engineering*, 51:1–9, 2004.

[18] E.K. Lee, R.J. Gallagher, and D.A. Patterson. A linear programming approach to discriminant analysis with a reserved-judgment region. *INFORMS Journal on Computing*, 15:23–41, 2003.

[19] E.K. Lee. Large-scale optimization-based classification models in medicine and biology. *Annals of Biomedical Engineering, Systems Biology and Bioinformatics*. 2006. In press.

[20] E.K. Lee. Optimization-Based Predictive Models in Medicine and Biology. *Optimization in Medicine*, Springer Netherlands, Computer Science Series, 2006. In press.

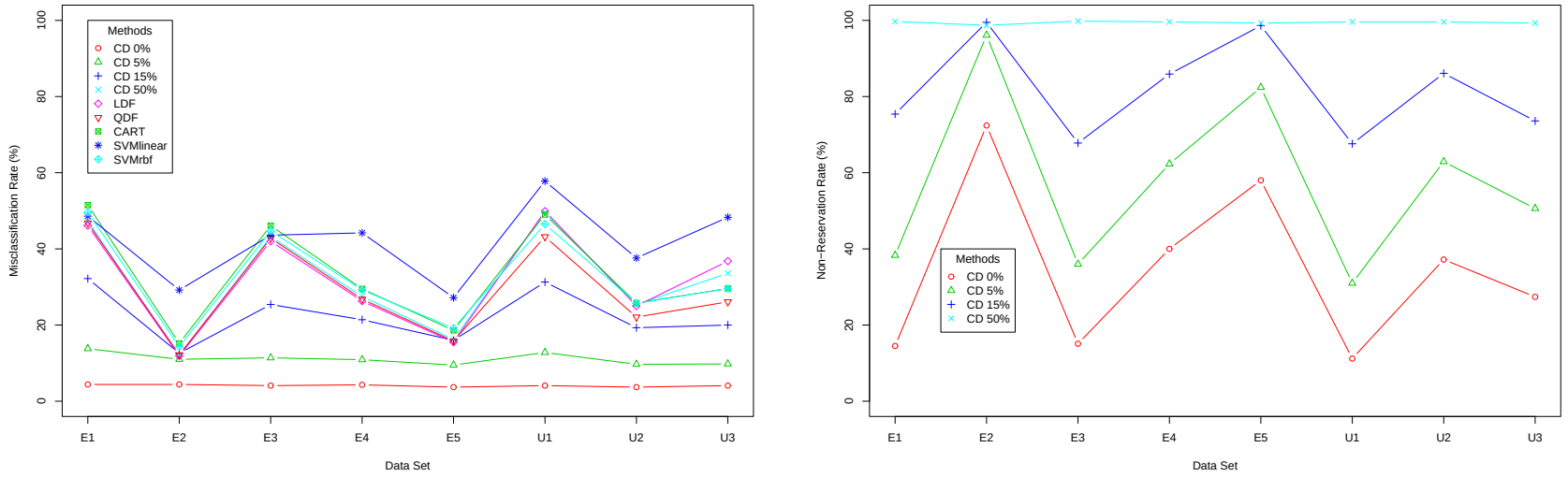
[21] C.P. Quesenberry and M.P. Gessaman. Nonparametric discrimination using tolerance regions. *Annals of Mathematical Statistics*, 39:664–673, 1968.

[22] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–270, 1971.

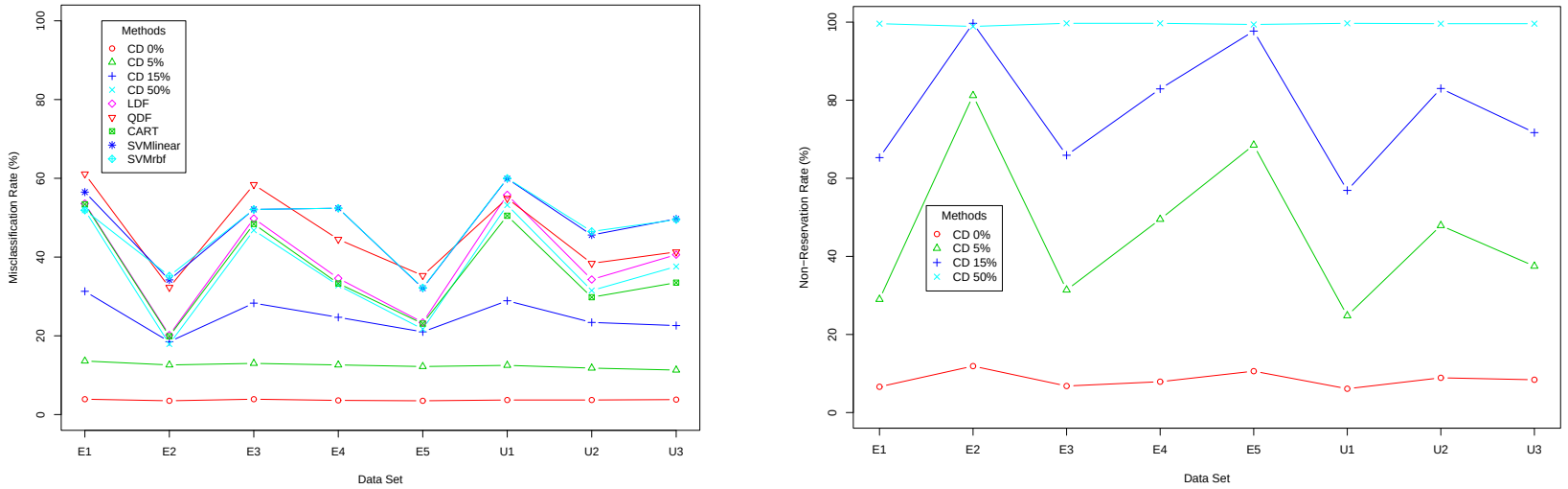
[23] V.N. Vapnik and A.Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, 26:532–553, 1981.

[24] V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[25] Adele Hart Wright. *The role of integrins in the differential upregulation of tumor cell motility by endothelial extracellular matrix proteins*. PhD thesis, Georgia Institute of Technology, December 1999.



(a)



(b)

Figure 1: Misclassification and non-reservation rates for various methods on simulated data sets derived from bivariate (a) normal and (b) contaminated normal distributions. The mean and covariance configurations ($E1 - E5$, $U1 - U3$) are described in Table 1 and in the text. The constrained discrimination model (CD) is tested with misclassification limits of 0, 5, 15, and 50%. Other methods tested are linear discriminant functions (LDF), quadratic discriminant functions (QDF), classification trees (CART), support vector machines with a linear kernel (SVMlinear), and support vector machines with a radial basis function kernel (SVMrbf). **Non-reservation reates are only reported for CD because it is the only method with a reserved judgment region.**

	CD 0%				CD 5%					CD 15%					CD 50%			
	Res	1	2	3		Res	1	2	3		Res	1	2	3		Res	1	2	3
1	88.0	9.7	1.7	0.7	1	65.7	23.7	6.1	4.5	1	28.6	43.6	15.6	12.1	1	0.1	66.0	31.6	2.2
2	87.6	1.6	10.1	0.6	2	66.0	6.0	23.8	4.2	2	28.6	15.8	43.5	12.1	2	0.2	30.2	67.6	2.0
3	90.6	2.6	2.8	4.0	3	67.7	7.6	7.6	17.1	3	29.7	18.3	18.1	33.9	3	0.7	44.5	45.4	9.4

	LDF				QDF				CART			
	1	2	3		1	2	3		1	2	3	
1	68.6	31.2	0.2	1	68.4	31.3	0.3	1	64.3	35.5	0.2	
2	30.9	68.9	0.2	2	31.2	68.6	0.2	2	34.3	65.4	0.2	
3	48.8	49.6	1.6	3	48.8	49.6	1.6	3	47.9	51.4	0.7	

(a) $E1$

	CD 0%				CD 5%					CD 15%					CD 50%			
	Res	1	2	3		Res	1	2	3		Res	1	2	3		Res	1	2	3
1	97.0	1.5	0.8	0.7	1	84.3	9.0	3.9	2.8	1	47.8	29.8	12.0	10.4	1	0.7	78.7	18.6	2.0
2	82.7	1.1	15.4	0.8	2	66.9	5.3	23.1	4.7	2	33.1	14.8	40.6	11.5	2	0.4	39.1	59.0	1.5
3	86.9	1.1	8.8	3.3	3	69.9	5.3	11.5	13.3	3	34.6	14.5	22.2	28.7	3	0.8	44.4	49.1	5.7

	LDF				QDF				CART			
	1	2	3		1	2	3		1	2	3	
1	75.3	24.7	0.1	1	82.8	17.1	0.1	1	73.3	26.7	0.1	
2	34.4	65.6	0.0	2	31.6	68.0	0.4	2	30.6	69.3	0.1	
3	40.4	59.4	0.1	3	39.4	58.5	2.1	3	37.1	62.4	0.5	

(b) $U1$

Figure 2: Classification matrices for (a) $E1$ and (b) $U1$ distributions (Table 1) with 100 training observations each in groups 1 and 2 and 10 training observations in group 3. Classification matrices for linear discriminant functions (LDF), quadratic discriminant functions (QDF), classification trees (CART), and constrained discrimination (CD) with training misclassification limits of 0, 5, 15, and 50%. The rows of each matrix correspond to the known group membership of observations, and the columns correspond to the allocated group. For the constrained method, the first column corresponds to the proportion of test observations from each group placed in the reserved judgment region. For each method, prior probabilities are calculated using the proportion of training observations from each group ($\pi = [100/210, 100/210, 10/210]$).

	CD 0%				CD 5%					CD 15%					CD 50%			
	Res	1	2	3		Res	1	2	3		Res	1	2	3		Res	1	2	3
1	88.2	9.5	1.6	0.7	1	65.8	24.1	6.2	3.9	1	30.2	43.9	15.7	10.2	1	0.2	66.4	31.4	1.9
2	87.9	1.6	9.8	0.7	2	66.0	6.1	24.0	4.0	2	30.0	15.7	43.8	10.4	2	0.2	30.7	67.2	1.9
3	90.7	2.5	2.6	4.2	3	68.5	7.9	7.7	15.9	3	32.7	18.4	18.3	30.6	3	1.0	44.6	45.5	8.9

	LDF				QDF				CART			
	1	2	3		1	2	3		1	2	3	
1	54.9	23.4	21.8	1	53.8	23.2	23.0	1	52.1	26.5	21.3	
2	23.3	54.6	22.1	2	23.0	53.9	23.1	2	26.4	52.1	21.5	
3	24.5	24.5	50.9	3	26.3	26.6	47.1	3	30.7	33.3	36.1	

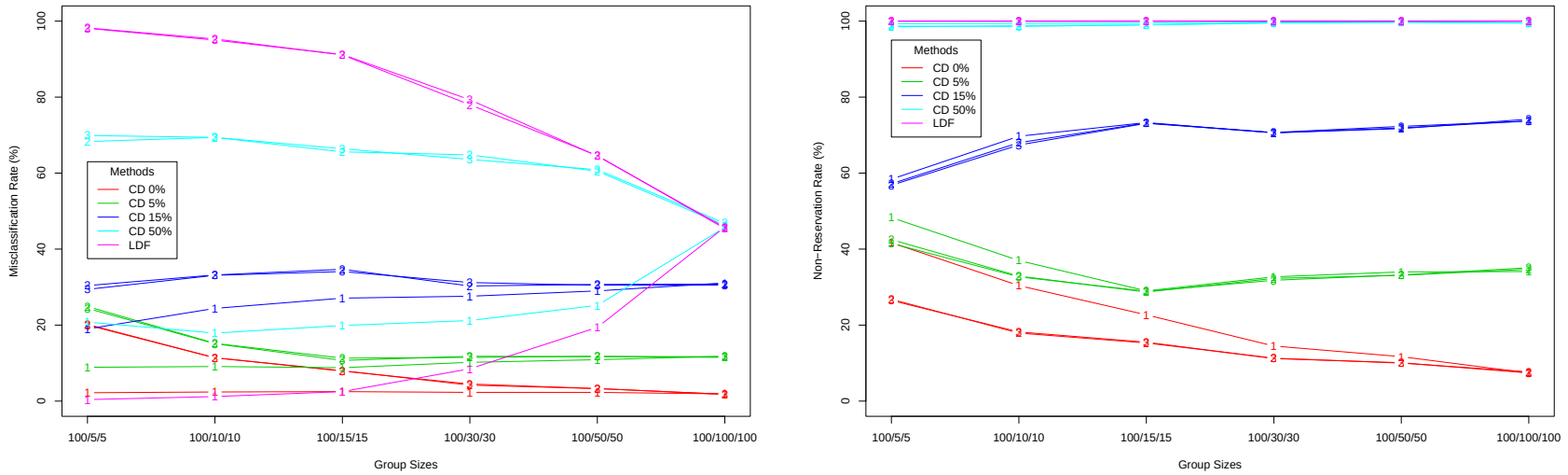
(a) $E1$

	CD 0%				CD 5%					CD 15%					CD 50%			
	Res	1	2	3		Res	1	2	3		Res	1	2	3		Res	1	2	3
1	97.0	1.5	0.8	0.7	1	83.7	9.6	4.0	2.7	1	49.5	30.2	12.8	7.5	1	0.9	78.5	19.1	1.6
2	82.7	1.0	15.5	0.9	2	65.9	5.4	24.7	4.0	2	33.1	15.0	43.1	8.8	2	0.5	38.9	58.8	1.7
3	86.8	1.1	8.6	3.5	3	70.3	5.5	12.5	11.8	3	38.4	14.9	24.6	22.1	3	1.3	44.4	48.6	5.7

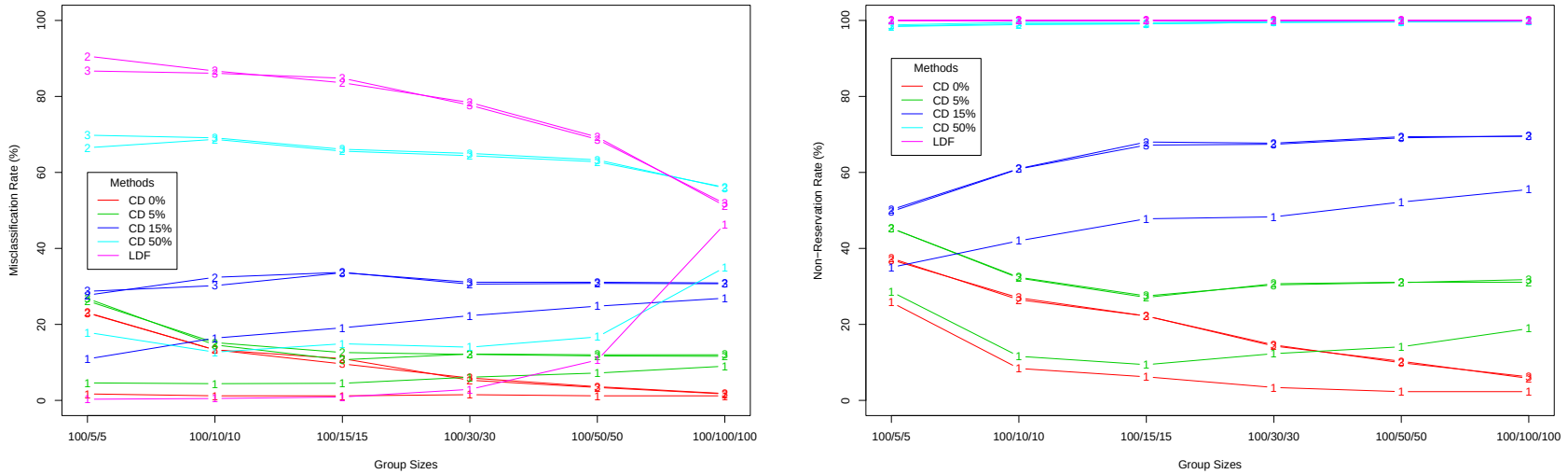
	LDF				QDF				CART			
	1	2	3		1	2	3		1	2	3	
1	59.8	20.2	20.0	1	71.9	12.7	15.3	1	64.6	17.9	17.5	
2	31.1	47.1	21.8	2	26.1	49.5	24.3	2	27.3	48.9	23.8	
3	31.1	25.7	43.2	3	26.6	27.4	46.0	3	29.6	33.1	37.3	

(b) $U1$

Figure 3: Classification matrices for (a) $E1$ and (b) $U1$ distributions (Table 1) with 100 training observations each in groups 1 and 2 and 10 training observations in group 3. Classification matrices for linear discriminant functions (LDF), quadratic discriminant functions (QDF), classification trees (CART), and constrained discrimination (CD) with training misclassification limits of 0, 5, 15, and 50%. The rows of each matrix correspond to the known group membership of observations, and the columns correspond to the allocated group. For the constrained method, the first column corresponds to the proportion of test observations from each group placed in the reserved judgment region. For each method, equal prior probabilities ($\pi = [1/3, 1/3, 1/3]$) are used for training.

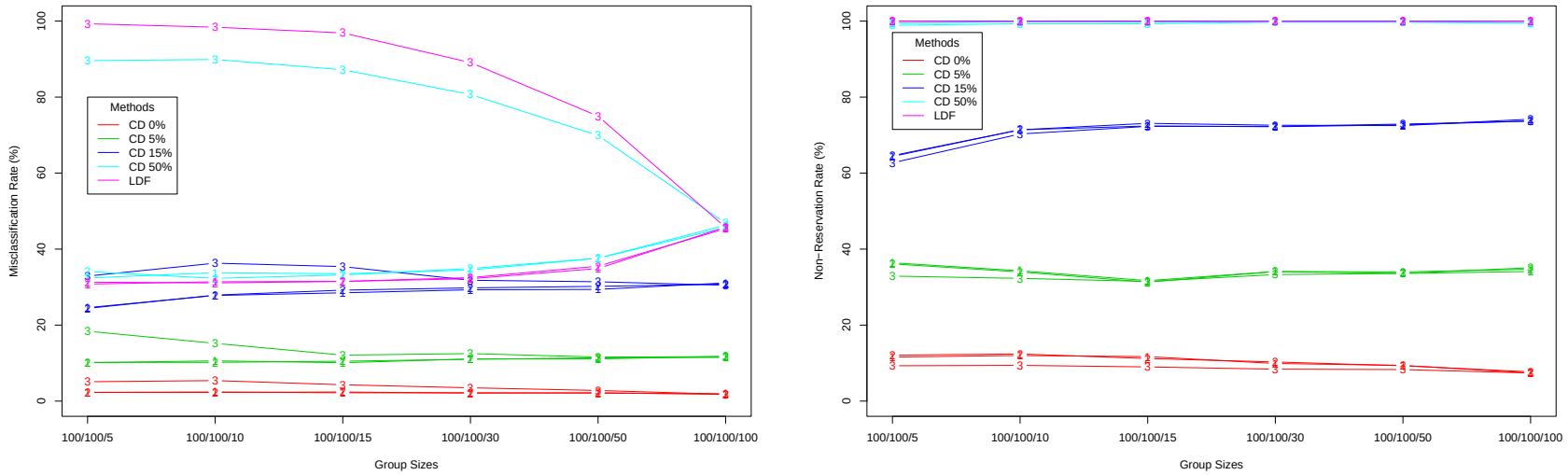


(a)

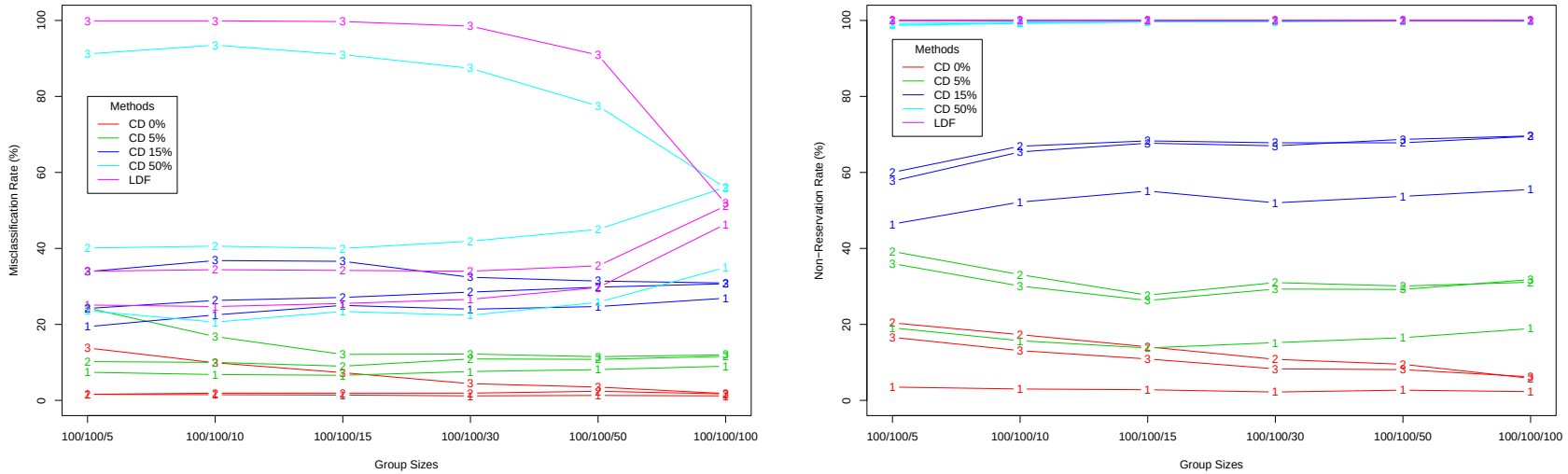


(b)

Figure 4: Misclassification and non-reservation rates for the constrained decision model (CD) with misclassification limits of 0, 5, 15, and 50% compared to linear discriminant functions (LDF) for data generated from bivariate normal distributions (a) $E1$ and (b) $U1$ as described in Table 1 and in the text. The three-group training data consisted of 100 observations from group 1 and varying numbers of observations from groups 2 and 3, as indicated on the x-axis of the graphs.



(a)



(b)

Figure 5: Misclassification and non-reservation rates for the constrained decision model (CD) with misclassification limits of 0, 5, 15, and 50% compared to linear discriminant functions (LDF) for data generated from bivariate normal distributions (a) $E1$ and (b) $U1$ as described in Table 1 and in the text. The three-group training data consisted of 100 observations each from groups 1 and 2 and varying numbers of observations from group 3, as indicated on the x-axis of the graphs.