

# A comparison of linear and mixture models for discriminant analysis under nonnormality

JOSEPH R. RAUSCH

University of Minnesota, Minneapolis, Minnesota

AND

KEN KELLEY

University of Notre Dame, Notre Dame, Indiana

Methods for discriminant analysis were compared with respect to classification accuracy under nonnormality through Monte Carlo simulation. The methods compared were linear discriminant analyses based both on raw scores and on ranks; linear logistic discrimination; and mixture discriminant analysis. Linear discriminant analysis and linear logistic discrimination were suboptimal in a number of scenarios with skewed predictors. Linear discriminant analysis based on ranks yielded the highest rates of classification accuracy in only a limited number of situations and did not produce a practically important advantage over competing methods. Mixture discriminant analysis, with a relatively small number of components in each group, attained relatively high rates of classification accuracy and was most useful for conditions in which skewed predictors had relatively small values of kurtosis.

Classification is of broad interest in science because it “permeates many scientific studies and also arises in the contexts of many applications” (Panel on Discriminant Analysis, Classification, and Clustering, 1989, p. 34). Examples in the educational, social, and behavioral sciences include identifying children in kindergarten at risk for future reading difficulties (Catts, Fey, Zhang, & Tomblin, 2001), identifying individuals at risk for addiction (Flowers & Robinson, 2002), and predicting the crimes that male juvenile offenders may commit according to their personality characteristics (Glaser, Calhoun, & Petrocelli, 2002). In the biological and medical sciences, applications of classification procedures include identifying patients with chronic heart failure (Udris et al., 2001), detecting lung cancer (Phillips et al., 2003), and determining whether certain breast masses are malignant or benign (Sahiner et al., 2004). In the management sciences, methods for classification have been used for such purposes as predicting bankruptcy (Jo, Han, & Lee, 1997) and investigating the product deletion process (Avlonitis, Hart, & Tzokas, 2000).

The primary goal of classification is to correctly sort objects into two or more mutually exclusive groups. Classification is often categorized into two subtypes—supervised and unsupervised (Hastie, Tibshirani, & Friedman, 2001; Panel, 1989). Supervised classification, also known as *discriminant analysis* (or, perhaps more appropriately, as *predictive discriminant analysis*; see Huberty, 1984, 1994), is used to correctly assign future objects to

groups that are already known to exist (Johnson & Wichern, 2002). Unsupervised classification, also known as *cluster analysis* (Panel on Discriminant Analysis, Classification, and Clustering, 1989), is used to assign objects to groups that are not known a priori.

We focus on methods for discriminant analysis in the present work—specifically, procedures based on linear and mixture models. With regard to linear methods, we investigate linear discriminant analysis (LDA) and linear logistic discrimination (LLD; Fan & Wang, 1999), along with an extension of LDA based on ranks (LDR; see, e.g., Conover & Iman, 1980). Furthermore, we investigate a lesser known method for discriminant analysis based on mixture models, which can be viewed as an extension of LDA (Fralely & Raftery, 2002). Mixture models are often used to model probability density functions through mixtures of normal distributions (Everitt & Hand, 1981; McLachlan & Basford, 1988; McLachlan & Peel, 2000; Titterton, Smith, & Makov, 1985). We use this approach to model the within-group multivariate densities for the predictors within a method known as *mixture discriminant analysis* (MDA; Fralely & Raftery, 2002; Hastie & Tibshirani, 1996; Taxt, Hjort, & Eikvil, 1991).

Previous research has been unclear concerning which of these four methods provides the highest rates of classification accuracy and under what conditions. For example, some comparisons of LLD and LDA have shown that these methods tend to provide similar results (Fan & Wang, 1999; Lim, Loh, & Shih, 2000; Press & Wilson,

1978; also see the remarks of Hastie et al., 2001, section 4.4.4), whereas others have demonstrated an advantage for one of these methods. Efron (1975), using asymptotic expansions, illustrated that LLD can be substantially less efficient than LDA when the assumptions of LDA are satisfied. McLachlan and Byth (1979) have shown, however, that in such situations, LLD and LDA provide similar classification accuracy when sample size is large relative to the number of predictors. Kiang (2003) demonstrated that LLD can yield more accurate classification rates than LDA can when the LDA assumptions of multivariate normality, linear relationships among the predictors, and equal covariance matrices do not hold. Furthermore, Lei and Koehly (2003) found that LDA and LLD can yield different rates of classification accuracy, depending on the assumed prior probabilities and the cut points for classifying objects, among other factors. Consequently, although LLD and LDA may often provide similar results with respect to classification accuracy, this is not necessarily true in a given research setting.

Even less information is available concerning the operating characteristics of LDR, although Barón (1991) compared this approach with LDA and LLD in a limited number of conditions. In particular, Barón demonstrated that LDR can provide more accurate classification than can LDA or LLD when the predictors are highly skewed in the same direction across groups. More research is needed, however, to thoroughly compare the accuracy of these methods under a wider variety of research scenarios.

Little work has been conducted comparing MDA with the more established classification procedures. Lim et al. (2000) investigated MDA in the context of a total of 33 methods for classification on 32 data sets. These authors found that MDA yielded proportions of classification accuracy comparable to those of LLD and LDA when averaging across data sets (all 3 methods were in the top 5 of the 33 methods and had average proportions of classification accuracy within .004 of one another), suggesting that MDA is comparable to these more popular approaches. Because the classification methods were primarily applied to real data by Lim et al., the characteristics of the underlying populations were generally unknown. Consequently, in the present study, we compared MDA with other discriminant analysis methods in conditions with prespecified population characteristics to fill this gap in the literature.

In particular, we were primarily concerned with comparing discriminant analysis methods for populations with nonnormal predictors, which represent realistic scenarios often encountered in applied research. For example, Micceri (1989) found that only 15% of the 440 distributions studied from the social and behavioral sciences had tail weights that approximated that of the normal distribution, and only approximately 28% of the distributions were relatively symmetric. With respect to skewness, Ostrander, Weinfurt, Yarnold, and August (1998) noted that, particularly in clinical research settings, "skewed variables [are] often encountered" (p. 661). Furthermore, Micceri found that more than half of the 125 psychometric measures studied had at least extreme asymmetry as defined by Micceri's criteria, and approximately 18% had asymmetry compa-

rable to the double exponential distribution, which exhibits skewness of 2.00. Consequently, it appears not only that nonnormality is common in applied research, but also that appreciable deviations from normality are not rare occurrences. Therefore, in the present study, we compared methods for discriminant analysis in these circumstances.

Previous research on the robustness of the classification accuracy of LDA given nonnormality has varied, depending on the type of nonnormal distributions investigated and the degree of nonnormality present (Ashikaga & Chang, 1981; Balakrishnan & Kocherlakota, 1985; Barón, 1991; Chinganda & Subrahmaniam, 1979; Lachenbruch, Sneeringer, & Revo, 1973; Nakanishi & Sato, 1985; Rawlings, Faden, Graubard, & Eckardt, 1986; Silva, Stam, & Neter, 2002; Subrahmaniam & Chinganda, 1978; see also McLachlan, 1992, chap. 5, who provides a review on the effects of nonnormality on LDA). LDA has generally been found to be less robust to more extreme deviations from normality with respect to skewness and kurtosis. Furthermore, skewness has been suggested to be "a more important factor than kurtosis in terms of misclassification of data" for LDA (Barón, 1991, p. 764). The results of Ashikaga and Chang suggested "that a more important issue than nonnormality is whether the distributions of the two populations are similar in shape" (p. 680). Finally, Nakanishi and Sato indicated that, for fixed levels of skewness, "[LDA] perform[s] best when the kurtoses are large" (p. 1190).

The robustness of LLD to nonnormality has typically been compared with that of LDA; LLD has generally been found to be more accurate, particularly when the distributions are highly skewed (Barón, 1991; Kiang, 2003; Silva et al., 2002). LDR, LLD, and LDA have been compared in a limited number of situations under nonnormality, and LDR has been shown to be potentially advantageous when the predictors are highly skewed in the same direction across groups (Barón, 1991). More research is needed, however, to fully explicate the utility of LDR relative to LDA and LLD in these situations. Given the lack of relevant research, MDA also needs to be compared with other classification methods for populations with nonnormal predictors.

Consequently, the primary aim of the present study was to systematically investigate the effects of nonnormal predictors on the proportions of classification accuracy for discriminant analysis methods based on linear and mixture models. Namely, we compared four methods for discriminant analysis through Monte Carlo simulation: LDA, LDR, LLD, and MDA. We manipulated three factors in the Monte Carlo study: group sample size relative to the number of predictors, group separation via the correlations among the predictors, and the type of nonnormality for the predictors. We will describe the exact specifications of the conditions within the Monte Carlo study momentarily, but first we will detail the four methods for discriminant analysis in the next section.

## METHODS FOR DISCRIMINANT ANALYSIS

Classification via discriminant analysis can be conceptualized through maximum posterior probability prin-

ciples (Huberty, 1994; Johnson & Wichern, 2002). That is, an object is classified into the group with the largest posterior probability, given the object's scores on the predictors and the prior probabilities. Although approaches to statistical classification have also been developed to account for the varying costs of classification errors (see, e.g., Huberty, 1994; Johnson & Wichern, 2002), we did not focus on this aspect of classification. If researchers have accurate information regarding classification error costs, it is wise to incorporate this information into the discriminant analysis (see, e.g., Huberty, 1994; Rudolph & Karson, 1988); however, such costs are often difficult to quantify (Huberty, 1994).

Methods for discriminant analysis are generally implemented through the following framework. A set of observations for which the group status of each object is assumed to be known, often referred to as the *training data*, is used to estimate the classification rule for allocating objects to groups. The classification rule obtained in this manner can be implemented on the training data to determine the proportion of observations that would be correctly classified; yet, such an approach is known to be too optimistic (e.g., Johnson & Wichern, 2002). Ideally, data not used to estimate the classification rule, often referred to as *test data*, should be used to provide a more realistic estimate of the accuracy of the discriminant analysis method. If test data are not available, other approaches are possible (see, e.g., Lachenbruch, 1967, for one such option), although the training–test data combination is ideal for investigating the classification accuracy of discriminant analysis methods.

Two restrictions were made to sufficiently limit the scope of the present work: continuous predictors and two-group classification settings. Neither of these restrictions is particularly problematic, because continuous predictor scenarios can be generalized to a mixture of continuous and categorical predictors through the described methods, and multiple-group classification settings can be conceptualized as multiple two-group classification settings. We describe the four discriminant analysis methods in the following subsections.

### LDA and LDR

LDA is one of the most popular methods of supervised classification. This procedure can be conceptualized as a nonparametric method (i.e., distributional assumptions are not explicitly made) because it maximizes between-group variability relative to within-group variability (Fisher, 1936). However, it can also be conceptualized as a parametric procedure for classification. In particular, Welch (1939) illustrated that LDA is optimal (i.e., it maximizes classification accuracy) under the assumptions that the within-group predictors follow multivariate normal distributions and that the population covariance matrices are equal across groups. On the basis of the latter framework, we assumed that the predictors follow multivariate normal distributions within Groups 1 and 2, in which the population mean vector for group  $g$  is  $\boldsymbol{\mu}_g$  ( $g = 1, 2$ ), and the group covariance matrices for the predictors are equal,

$\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ . Given these assumptions, the posterior probability for membership in Group 1 for LDA is

$$\pi_1 = \frac{p_1 \phi(\mathbf{x} | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{p_1 \phi(\mathbf{x} | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + p_2 \phi(\mathbf{x} | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})}, \quad (1)$$

where  $\mathbf{x}$  is a vector of values for the  $q$  predictors,  $\phi(\mathbf{x} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma})$  is the multivariate normal density function for  $\mathbf{x}$  given  $\boldsymbol{\mu}_g$  and  $\boldsymbol{\Sigma}$ , and  $p_1$  and  $p_2$  are the population group proportions for Groups 1 and 2, respectively (i.e., the prior probabilities; Johnson & Wichern, 2002). In practice,  $\boldsymbol{\mu}_g$  and  $\boldsymbol{\Sigma}$  in Equation 1 are generally not known and are estimated from the training data using ordinary least squares. Note that the posterior probability for membership in Group 2 is  $1 - \pi_1$ . Consequently, the maximum posterior probability approach in the case of two groups corresponds to classifying an object into Group 1 if Equation 1 is greater than .5, or into Group 2 otherwise.

A modification of LDA has also been developed to provide a robust version of this method. This robust method is a linear discriminant analysis on rank scores, rather than on the original scores for the predictors (Conover & Iman, 1980). That is, LDR is implemented in the manner described above for LDA, except that all the calculations are based on the predictors' rank scores. The impetus for developing LDR is to provide more accurate classification in situations with nonnormal predictors by relaxing the assumption of multivariate normality within the parametric version of LDA.

### LLD

Logistic regression, a statistical procedure subsumed by the generalized linear model, is similar to linear regression, except that the outcome variable is categorical rather than continuous (for a thorough discussion of logistic regression, see, e.g., Agresti, 2002). In logistic regression, the outcome variable is assumed to follow a binomial distribution, and the log odds (i.e., logit) is assumed to be appropriately described by a linear function of the logistic regression coefficients. When logistic regression is used for discriminant analysis, as is the case here, it is often referred to as *logistic discrimination* (Anderson, 1975; McLachlan, 1992, chap. 8).

LLD is a special case of logistic discrimination in which the logit is assumed to be appropriately characterized by a linear function of the predictors. The general formulation for LLD is

$$\log\left(\frac{\pi_1}{1 - \pi_1}\right) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}, \quad (2)$$

where  $\beta_0$  is the intercept,  $\boldsymbol{\beta}$  is a vector of logistic regression coefficients of length  $q$ ,  $\pi_1$  is the posterior probability for membership in Group 1 ( $\pi_2 = 1 - \pi_1$ ), and  $\log[\pi_1 / (1 - \pi_1)]$  is the logit [in which  $\log(\cdot)$  is the logarithm with base  $e$ ]*—*that is, the link function for logistic regression within the generalized linear model (Agresti, 2002).

Some further points regarding LLD are necessary. First, no explicit assumptions are made regarding the distributional forms of the predictors entered into the LLD model.

This is in contrast to the parametric version of LDA, which assumes multivariate normal within-group populations, since LLD estimates the marginal densities of the predictors “in a fully nonparametric and unrestricted fashion, using the empirical distribution function which places mass  $1/N$  at each observation” (Hastie et al., 2001, p. 104). Second, no explicit assumption of equal covariance matrices is made in LLD. Third, the intercept in LLD is a function of the log of the ratio of the prior probabilities (see, e.g., McLachlan, 1992, chap. 8):

$$\beta_0 = \beta_0^* + \log\left(\frac{p_1}{p_2}\right), \quad (3)$$

where  $\beta_0^*$  is the part of  $\beta_0$  that does not depend on the prior probabilities.

LLD classifies an object into the group with the largest posterior probability of group membership (see, e.g., Fan & Wang, 1999), where the posterior probability for membership in Group 1 is

$$\pi_1 = \frac{e^{\beta_0 + \beta'x}}{1 + e^{\beta_0 + \beta'x}}. \quad (4)$$

As is the case with LDA, an object is classified into Group 1 if Equation 4 is greater than .5, or is classified into Group 2 otherwise. Because they are generally not known, the population parameters in Equation 4 must be estimated, which is typically done with maximum likelihood estimation (MLE) procedures.

### MDA

MDA can be viewed as an extension of LDA that models the within-group multivariate density of the predictors through a mixture (i.e., a weighted sum) of multivariate normal distributions (Fraley & Raftery, 2002). In principle, this approach is useful for one of two purposes: (1) to model multivariate nonnormality or nonlinear relationships among variables within each group, allowing for more accurate classification; or (2) to determine whether latent/underlying subclasses may be present in each group. In the present work, we aimed to use MDA to improve classification accuracy.

For MDA, the posterior probability of membership in Group 1 is

$$\pi_1 = \frac{p_1 \sum_{c=1}^{C_1} \tau_{c1} \phi(\mathbf{x} | \boldsymbol{\mu}_{c1}, \boldsymbol{\Sigma}_{c1})}{p_1 \sum_{c=1}^{C_1} \tau_{c1} \phi(\mathbf{x} | \boldsymbol{\mu}_{c1}, \boldsymbol{\Sigma}_{c1}) + p_2 \sum_{c=1}^{C_2} \tau_{c2} \phi(\mathbf{x} | \boldsymbol{\mu}_{c2}, \boldsymbol{\Sigma}_{c2})}, \quad (5)$$

where  $\tau_{cg}$  is the mixing proportion for the  $c$ th component in group  $g$  ( $0 < \tau_{cg} < 1$ ;  $\sum_{c=1}^{C_g} \tau_{cg} = 1$  for a fixed value of  $g$ ;  $c = 1, 2, \dots, C_g$ ;  $C_g$  is the number of components in group  $g$ ), a component is defined as a multivariate normal distribution used to model the within-group density of the predictors, and  $\phi(\mathbf{x} | \boldsymbol{\mu}_{cg}, \boldsymbol{\Sigma}_{cg})$  is the multivariate normal density function for  $\mathbf{x}$  given the mean vector,  $\boldsymbol{\mu}_{cg}$ , and the covariance matrix,  $\boldsymbol{\Sigma}_{cg}$ . Note that, in contrast to LDA, there is no assumption of equal covariance matrices across groups for MDA in Equation 5. An object is classified into Group 1 if  $\pi_1$  in Equation 5 is greater than .5, or is classi-

fied into Group 2 otherwise. The population parameters in Equation 5 are typically estimated using MLE (McLachlan & Peel, 2000), where the EM algorithm is implemented for MLE in the present work (Fraley & Raftery, 2002, 2006).

The particular implementation of MDA used here is that of Fraley and Raftery (Banfield & Raftery, 1993; Fraley & Raftery, 1999, 2002, 2003, 2006) through the `mclustDAtrain()` and `mclustDAtest()` functions from the MCLUST library in the program R (R Development Core Team, 2007). This approach uses information criteria to simultaneously select (1) the number of within-group components, which can vary from 1 to the specified maximum value [the default maximum value in `mclustDAtrain()` is 9]; and (2) the best model from a set of models that constrains or varies different features of the within-group component covariance matrix in Equation 5,  $\boldsymbol{\Sigma}_{cg}$ , across within-group components. Both the number of within-group components and the within-group model for the component covariance matrices can vary across groups in `mclustDAtrain()`. The information criterion used in `mclustDAtrain()` is the Bayesian information criterion (Schwarz, 1978), because it has been shown to perform well in the MDA framework used by Fraley and Raftery (Biernacki & Govaert, 1999).

The MDA framework used by Fraley and Raftery's (1999, 2002, 2003, 2006) MCLUST software relies on three features of the within-group component covariance matrices: volume, shape, and orientation. These features are controlled through different parts of the eigenvalue decomposition of the component covariance matrices, an approach for model-based clustering proposed by Banfield and Raftery (1993):

$$\boldsymbol{\Sigma}_{cg} = \lambda_{cg} \mathbf{D}_{cg} \mathbf{A}_{cg} \mathbf{D}'_{cg}, \quad (6)$$

where  $\lambda_{cg}$  is a constant for component  $c$  in group  $g$  that controls the volume of the covariance matrix,  $\mathbf{D}_{cg}$  is the matrix of eigenvectors that controls the orientation of the covariance matrix, and  $\mathbf{A}_{cg}$  is a diagonal matrix with elements proportional to the eigenvalues of the covariance matrix that controls the shape of the covariance matrix. Varying or constraining these features across within-group components yields 14 models (Celeux & Govaert, 1995), 10 of which are used in `mclustDAtrain()` when two or more predictors are used. Table 1 presents these 10 models.

The model notation used in Table 1 is the same as that used in `mclustDAtrain()`; the three letters in the first column represent information corresponding to the volume, shape, and orientation, respectively, of the within-group component covariance matrices: (1) “E” represents a particular feature that is equal across the within-group component covariance matrices, (2) “V” represents a particular feature that varies across the within-group component covariance matrices, and (3) “I” represents that the identity matrix is used instead of the  $\mathbf{A}$  or  $\mathbf{D}$  matrix within the eigenvalue decomposition. Model choice and the number of parameters estimated for the within-group component covariance matrices range from the EII model, in which one parameter (i.e.,  $\lambda$ ) is estimated, to the VVV model, in which  $C_g[q(q+1)/2]$  parameters are estimated

**Table 1**  
**Within-Group Models for Component Covariance Matrices in**  
**Mixture Discriminant Analysis Using `mclustDATrain()`**

Model	$\Sigma_c$	Volume	Shape	Orientation	$\Sigma_c$ Structure			
					Component 1		Component 2	
					Predictor 1	Predictor 2	Predictor 1	Predictor 2
EII	$\lambda \mathbf{I}$	equal	identity	identity	$a$		$a$	
					0	$a$	0	$a$
VII	$\lambda_c \mathbf{I}$	varies	identity	identity	$a$		$b$	
					0	$a$	0	$b$
E EI	$\lambda \mathbf{A}$	equal	equal	identity	$a$		$a$	
					0	$b$	0	$b$
VEI	$\lambda_c \mathbf{A}$	varies	equal	identity	$a$		$ca$	
					0	$b$	0	$cb$
EVI	$\lambda \mathbf{A}_c$	equal	varies	identity	$a$		$c$	
					0	$b$	0	$d$
VVI	$\lambda_c \mathbf{A}_c$	varies	varies	identity	$a$		$c$	
					0	$b$	0	$d$
EEE	$\lambda \mathbf{DAD}'$	equal	equal	equal	$a$		$a$	
					$b$	$c$	$b$	$c$
EEV	$\lambda \mathbf{D}_c \mathbf{AD}'_c$	equal	equal	varies	$a$		$d$	
					$b$	$c$	$e$	$f$
VEV	$\lambda_c \mathbf{D}_c \mathbf{AD}'_c$	varies	equal	varies	$a$		$d$	
					$b$	$c$	$e$	$f$
VVV	$\lambda_c \mathbf{D}_c \mathbf{A}_c \mathbf{D}'_c$	varies	varies	varies	$a$		$d$	
					$b$	$c$	$e$	$f$

Note— $\Sigma_c$  Structure, covariance structure for the particular model with two components and two predictors in which the lower diagonal is illustrated for the covariance matrices. For the EVI model, there is an additional constraint of  $ab = cd$ ; for the EEV model, there is an additional constraint of equal eigenvalues across within-group components; and for the VEV model, there is an additional constraint of proportional (ordered) eigenvalues across within-group components. See the text for further explanation.

for the within-group component covariance matrices. Consequently, the models for the component covariance matrices tend to range from more simple to more complex from the top to the bottom of Table 1.

The last four columns in Table 1 illustrate the models for the component covariance matrices for two components and two predictors. Different letters within a given model denote parameter estimates that are freely estimated, whereas the same letter within a given model denotes parameters that are constrained to be equal. Furthermore, note that some models require additional constraints, as is stated in the note to Table 1. In general, on the basis of the models in Table 1, MDA is used to attain relatively high rates of classification accuracy by balancing parsimony and flexibility when selecting models for the within-group predictor densities within discriminant analysis.

## METHOD

We used Monte Carlo simulation in R 2.6.0 (R Development Core Team, 2007) to systematically evaluate the four methods of discriminant analysis. Training data sets (10,000) with two groups and specified group sample sizes (these will be detailed momentarily) were generated in each condition to calculate a sample classification rule for each of the discriminant analysis methods within each training data set. A test data set with a sample size of 10,000 was also generated for each training data set, in which the test data had group sample sizes proportional to the training data. The test data were used to estimate the proportions of correct classifications (PCCs) across groups for each of the discriminant analysis methods. The PCC is bounded by 0 and 1, where larger PCCs are desirable. Note

that values greater than 0 for the PCC can be attained by simple allocation procedures, such as randomly allocating objects to groups according to the prior probabilities; consequently, PCC values for the discriminant analysis methods should generally be compared—at a minimum—with PCC values for simple allocation rules to ascertain their practical value. We assumed that the prior probabilities for the groups were equal to the group probabilities in the training sample, as is common in practice. We focused on the PCC as the primary outcome of the Monte Carlo study because it is commonly used to investigate the efficacy of discriminant analysis methods for classification, as can be seen in the studies cited in the introduction. In situations in which nonconvergence occurred when obtaining the classification rule for a particular method of discriminant analysis, new data sets were not generated; thus, the actual number of training–test data set combinations used for calculation of the mean PCC in a particular condition is less than or equal to 10,000. The nonconvergence rates are reported in the next section.

The four methods for discriminant analysis were implemented in R to estimate the classification rules from the training data. Three variations of MDA, implemented via `mclustDATrain()`, were included in the Monte Carlo study. The default value of 9 for the maximum number of within-group components was used for one variation of MDA (denoted MD9), although it could be argued that this value is too large for use in the social and behavioral sciences. Consequently, values of 2 and 4 (denoted MD2 and MD4, respectively) were also used for the maximum number of within-group components in MDA to determine their utility. The `lda()` function from the MASS library in R was used to implement LDA (Venables & Ripley, 2002). Also, the method of transforming raw scores to ranks based on linear interpolation for the test data (Conover & Iman, 1980; Huberty, 1994) was coded in R and used in combination with `lda()` to obtain the PCC for LDR. Furthermore, the `glm()` function was used to carry out LLD, in which the number of iterations was increased from a default of 25 to 1,000 to increase the likelihood of algorithmic convergence.

The factors manipulated in the Monte Carlo study were (1) group sample sizes relative to the number of predictors (4 levels), (2) group separation via the correlation among the predictors (2 levels), and (3) the type of nonnormality for the predictors (12 levels). A total of  $4 \times 2 \times 12 = 96$  conditions were investigated in the present study, in which six methods of discriminant analysis (including the two modifications of MDA) were compared with respect to the mean PCC within each condition. The levels of the variables used in the Monte Carlo study are described in more detail in the following subsections.

### Group Sample Size Relative to the Number of Predictors

The number of predictors was set to 8, and four sets of group sample sizes were used for each of Groups 1 and 2—32,32; 64,64; 24,72; and 72,24—which yielded ratios for the group sample sizes relative to the number of predictors of 4,4; 8,8; 3,9; and 9,3. Unequal group sample sizes were investigated in order to evaluate their effect on the mean PCC, along with the corresponding unequal prior probabilities. The values for the number of predictors and the group sample sizes were chosen to represent realistic applied research scenarios.

### Group Separation Through Predictor Correlations

Group separation was chosen as a design factor for this study because of its direct relation to the PCC. Group separation can be quantified through the Mahalanobis distance (Mahalanobis, 1936):

$$\Delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}, \quad (7)$$

where  $\Delta$  is a population distance measure for multivariate scenarios that takes into account different variances for the predictors along with bivariate covariation among the predictors. All other things being equal, larger  $\Delta$ s yield greater group separation, making it more likely that objects are correctly classified into groups.

The population standardized mean differences were equal across predictors, set at a value of .65, where Group 1 always had the highest means on the predictors. A value of .65 for the standardized mean difference corresponds to an effect halfway between medium and large using Cohen's (1988) standards. The population variances of the predictors were all set to 1, whereas the population correlations among the predictors were the same for a given covariance matrix and were either .1 or .5, corresponding to small or large correlations using Cohen's standards. The manipulation of the correlation among the predictors yielded  $\Delta$ s of approximately 1.41 and 0.87 for predictor intercorrelations of .1 and .5, respectively.

### Nonnormal Predictors

Multivariate nonnormal distributions for the predictors were generated using the Vale and Maurelli (1983) procedure, which is a multivariate extension of Fleishman's (1978) method. The Vale and Maurelli procedure was coded into R by the authors. Using standardized measures of skewness and (excess) kurtosis— $\gamma_1$  and  $\gamma_2$ , respectively, as defined in Stuart and Ord (1994, p. 109; note that, for a normal distribution, both measures are equal to 0)—the marginal distributions corresponding to the multivariate distribution for a particular group are described below. Illustrations of the specified marginal distributions in each condition are shown in Figures 1 and 2 for conditions with small and large values for kurtosis, respectively.

Twelve conditions constituted the nonnormal predictor factor, in which the condition names correspond to the marginal distributions in Group 1, a comma, and the marginal distributions in Group 2:

- Norm,PsSk, with small (7) and large (8) kurtosis, consisted of normal marginal distributions in Group 1 and positively skewed marginal distributions in Group 2 ( $\gamma_1 = 1.25, \gamma_2 = 1.5$  with small kurtosis;  $\gamma_1 = 1.25, \gamma_2 = 3.75$  with large kurtosis).
- PsSk,PsSk, with small (3) and large (4) kurtosis, consisted of positively skewed marginal distributions in both groups ( $\gamma_1 = 1.25, \gamma_2 = 1.5$  for Groups 1 and 2 with small kurtosis;  $\gamma_1 = 1.25, \gamma_2 = 3.75$  for Groups 1 and 2 with large kurtosis).
- PsSk,PsSk, with small (5) and large (6) kurtosis, consisted of positively skewed marginal distributions in Group 1 ( $\gamma_1 = 1.25, \gamma_2 = 1.5$  with small kurtosis;  $\gamma_1 = 1.25, \gamma_2 = 3.75$  with large kurtosis) and normal marginal distributions in Group 2.
- Norm,PsSk, with small (7) and large (8) kurtosis, consisted of normal marginal distributions in Group 1 and positively skewed marginal distributions in Group 2 ( $\gamma_1 = 1.25, \gamma_2 = 1.5$  with small kurtosis;  $\gamma_1 = 1.25, \gamma_2 = 3.75$  with large kurtosis).
- PsSk,NgSk, with small (9) and large (10) kurtosis, consisted of positively skewed marginal distributions in Group 1 ( $\gamma_1 = 1.25, \gamma_2 = 1.5$  with small kurtosis;  $\gamma_1 = 1.25, \gamma_2 = 3.75$  with large kurtosis) and negatively skewed marginal distributions in Group 2 ( $\gamma_1 = -1.25, \gamma_2 = 1.5$  with small kurtosis;  $\gamma_1 = -1.25, \gamma_2 = 3.75$  with large kurtosis).
- NgSk,PsSk, with small (11) and large (12) kurtosis, consisted of negatively skewed marginal distributions in Group 1 ( $\gamma_1 = -1.25, \gamma_2 = 1.5$  with small kurtosis;  $\gamma_1 = -1.25, \gamma_2 = 3.75$  with large kurtosis) and positively skewed marginal distributions in Group 2 ( $\gamma_1 = 1.25, \gamma_2 = 1.5$  with small kurtosis;  $\gamma_1 = 1.25, \gamma_2 = 3.75$  with large kurtosis).

Although positive skewness was predominantly used in the simulation study, the results of these conditions are also indicative of similar conditions with negative skewness. Furthermore, the skewness value of 1.25 was chosen in order to represent realistic scenarios in applied research on the basis of Micceri's (1989) findings. The kurtosis value of 3.75 was chosen because it was the maximum value in Fleishman's (1978) table, and a kurtosis value of 1.5 was chosen because it was the minimum value for kurtosis in Fleishman's table, given a skewness value of 1.25.

## RESULTS

The primary aim of the Monte Carlo study was to systematically compare the methods for discriminant analysis with respect to the mean PCC in scenarios with nonnormal predictors. In particular, we aimed to determine when the commonly used methods of LDA and LLD were suboptimal and under what conditions, if any, MDA or LDR were optimal relative to the other methods investigated.

With respect to algorithmic convergence, none of the methods exhibited serious issues. The only methods that failed to converge were those based on MDA. That is, MD9 failed to converge more often than MD4, which failed to converge more often than MD2. Still, the lowest proportion of convergence for MD9 in a particular condition was .992. Furthermore, MD2 failed to converge only for three of the training data sets within the entire simulation study. In general, nonconvergence was rare and posed no problems for our purposes.

The mean PCC for each of the discriminant analysis methods within each condition is illustrated in Figures 3–6. These figures distinguish equal versus unequal group sample size conditions and small versus large kurtosis nonnormality conditions. In principle, a line with a slope of 0 in a particular condition in Figures 3–6 corresponds to equal mean PCCs across discriminant analysis methods. Furthermore, larger deviations from such a line tend to correspond to larger differences between methods.

LDA and LLD yielded similar results across all conditions in the Monte Carlo study. Note that differences

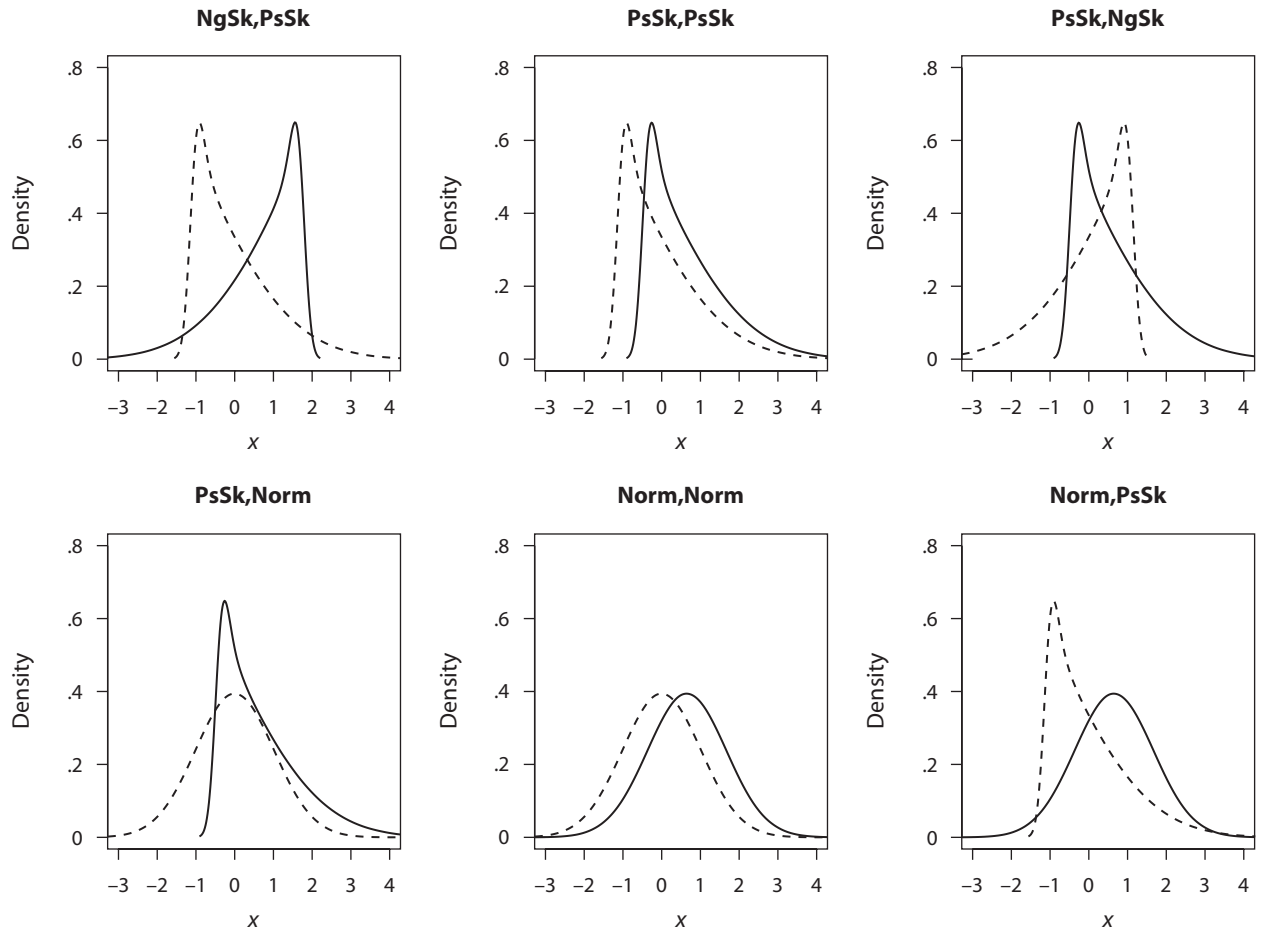


Figure 1. Illustrations of marginal distributions for predictors in nonnormal conditions with small values for kurtosis. The group with the higher means—Group 1—is represented by a solid line, whereas Group 2 is represented by a dotted line. For all distributions,  $SD = 1$ . Condition names consist of the marginal distributions in Group 1 followed by a comma and the marginal distribution in Group 2. NgSk, negative skew; PsSk, positive skew; Norm, normality.

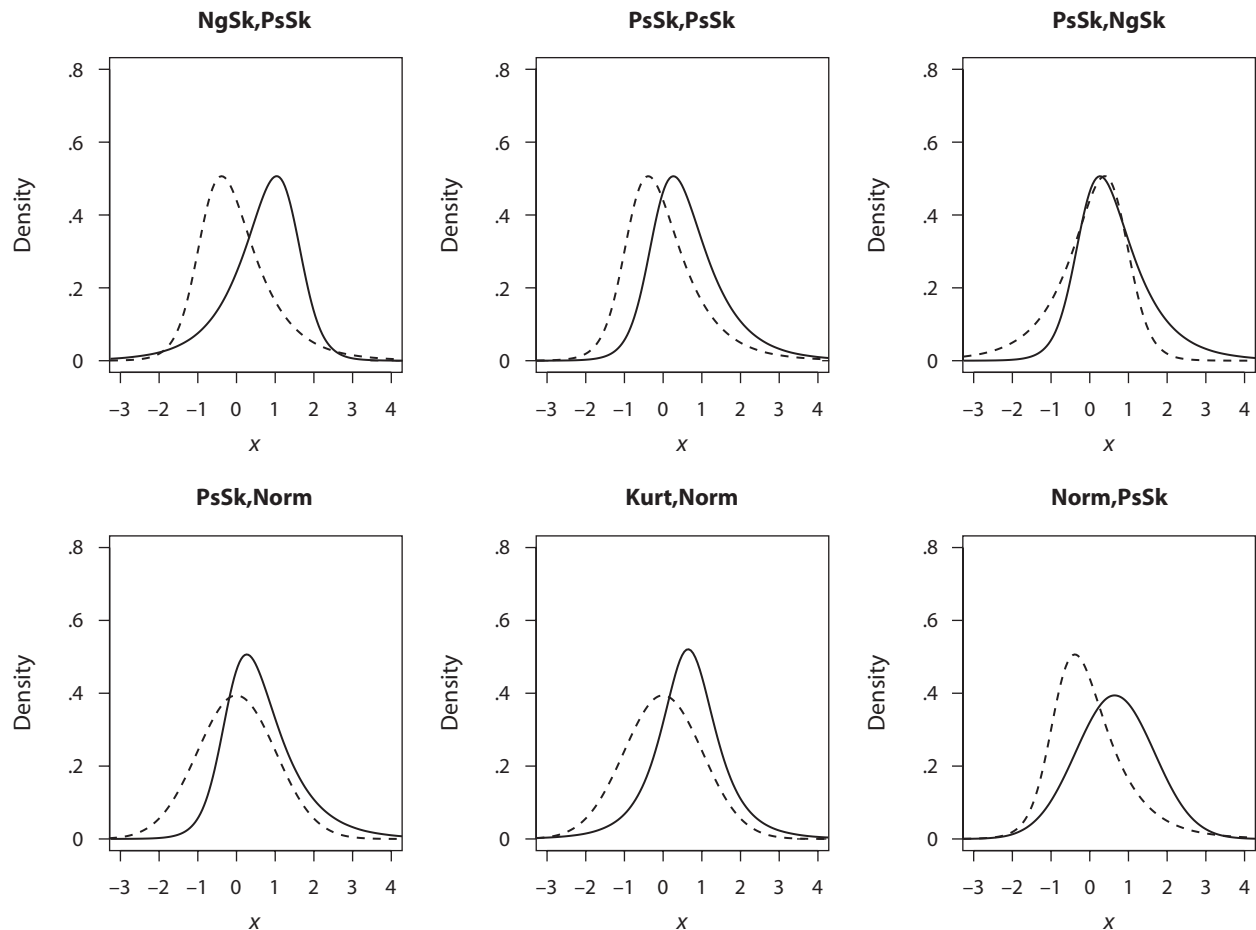
between methods on the mean PCC near .05 or larger were considered to be practically important in the present study. Given this criterion, the largest difference between LDA and LLD on the mean PCC of .009 was not considered practically important, and the majority of the differences between LDA and LLD were much smaller than this maximum value. Consequently, we focus only on the results of LDA in this section, since they are also representative of LLD. For the three versions of MDA, MD9 was generally no better than—and, at times, was substantially less accurate than—MD4. MD2 and MD4 tended to yield similar results across all conditions, although each of these methods yielded higher rates of classification accuracy than did the other in certain conditions. Consequently, we focus on the results of MDA based on maxima of both two and four within-group components.

### Equal Group Sample Sizes

For equal group sample sizes and nonnormality conditions with small kurtosis, Figure 3 illustrates that, as is typically the case, larger sample sizes and larger  $\Delta$ s (through a smaller value of  $\rho$ , the predictor intercorrelation) tend

to correspond to larger mean PCCs, everything else being equal. Furthermore, for some conditions with  $\rho = .5$ , the mean PCCs are near .60 and are thus relatively close to a mean PCC of .50 attained from randomly classifying objects to groups on the basis of the prior probabilities. Given that the parameter values for the Monte Carlo study were selected to be representative of applied research settings, the mean PCCs illustrated in Figure 3 should not necessarily be considered unusual. Moreover, it can be argued that a .10 increase over random classification can be considered large enough that discriminant analysis methods are indeed useful relative to this simple approach.

In Figure 3, LDA provided an advantage over MD2 and MD4 in the Norm, Norm conditions with  $\rho = .5$ , whereas LDR and LDA had similar mean PCCs in these same conditions. These conditions were the only situations in Figure 3, however, in which LDA provided more classification accuracy than did both MD2 and MD4. LDR provided the largest mean PCCs within the NgSk, PsSk and Norm, PsSk conditions with  $\rho = .1$  in Figure 3. The differences between LDR and the other discriminant analysis methods, however, were small in these conditions, practi-



**Figure 2. Illustrations of marginal distributions for predictors in nonnormal conditions with large values for kurtosis. The group with the higher means—Group 1—is represented by a solid line, whereas Group 2 is represented by a dotted line. For all distributions,  $SD = 1$ . Kurt, leptokurtic marginal distributions. All other abbreviations are as in Figure 1.**

cally speaking. Consequently, LDR was of limited utility in Figure 3.

MD2 or MD4 provided substantially larger mean PCCs than did LDA, LLD, and LDR in a number of scenarios in Figure 3. Specifically, either MD2 or MD4 yielded the largest mean PCCs in the NgSk,PsSk conditions with  $\rho = .5$ ; in the PsSk,Norm conditions; in the PsSk,PsSk conditions, with the exception of the  $n_1 = 32, n_2 = 32, \rho = .1$  condition, in which MD2 was slightly less accurate than LDR; and in the PsSk,NgSk conditions with  $\rho = .5$ . For example, MD4 yielded a mean PCC just over .75, compared with a mean PCC just under .65 for LDA, in the PsSk,PsSk condition with  $n_1 = 64, n_2 = 64, \rho = .5$ . A similar difference in mean PCCs was obtained when comparing MD4 and LDA with  $n_1 = 64, n_2 = 64$ , and  $\rho = .5$  in the PsSk,Norm and NgSk,PsSk conditions. Thus, MD4 provided a clear advantage over linear methods for discriminant analysis in these scenarios within Figure 3, especially in scenarios with  $\rho = .5$ . In general, MD4 tended to perform nearly as well as MD2 in conditions in which MD2 was optimal, whereas MD4, at times, provided substantially larger mean PCCs than did MD2 in Figure 3. Thus, within the different modifications of MDA investigated, MD4 often yielded

the highest, or near the highest, rates of classification accuracy. In general, Figure 3 provides evidence that MDA can be used to substantially increase classification accuracy in a number of conditions with nonnormal predictors.

The results in Figure 4 for scenarios with equal group sample sizes and nonnormal conditions with large values of kurtosis were generally consistent with the results in Figure 3, with one primary exception. Namely, for scenarios in Figure 4 in which MD2 or MD4 yielded an appreciable advantage over competing methods, this advantage was generally smaller compared with Figure 3. Moreover, the differences among methods observed in Figure 4 were rarely practically significant, with the possible exception of the advantage for MD4 over competing methods in the NgSk,PsSk condition with  $n_1 = 64, n_2 = 64$ , and  $\rho = .5$ . Thus, larger values of kurtosis for skewed predictors tended to attenuate the advantage gained by using methods other than linear methods for discriminant analysis, everything else being equal.

**Unequal Group Sample Sizes**

Figures 5 and 6 contain the results from the Monte Carlo study for the conditions with unequal group sample

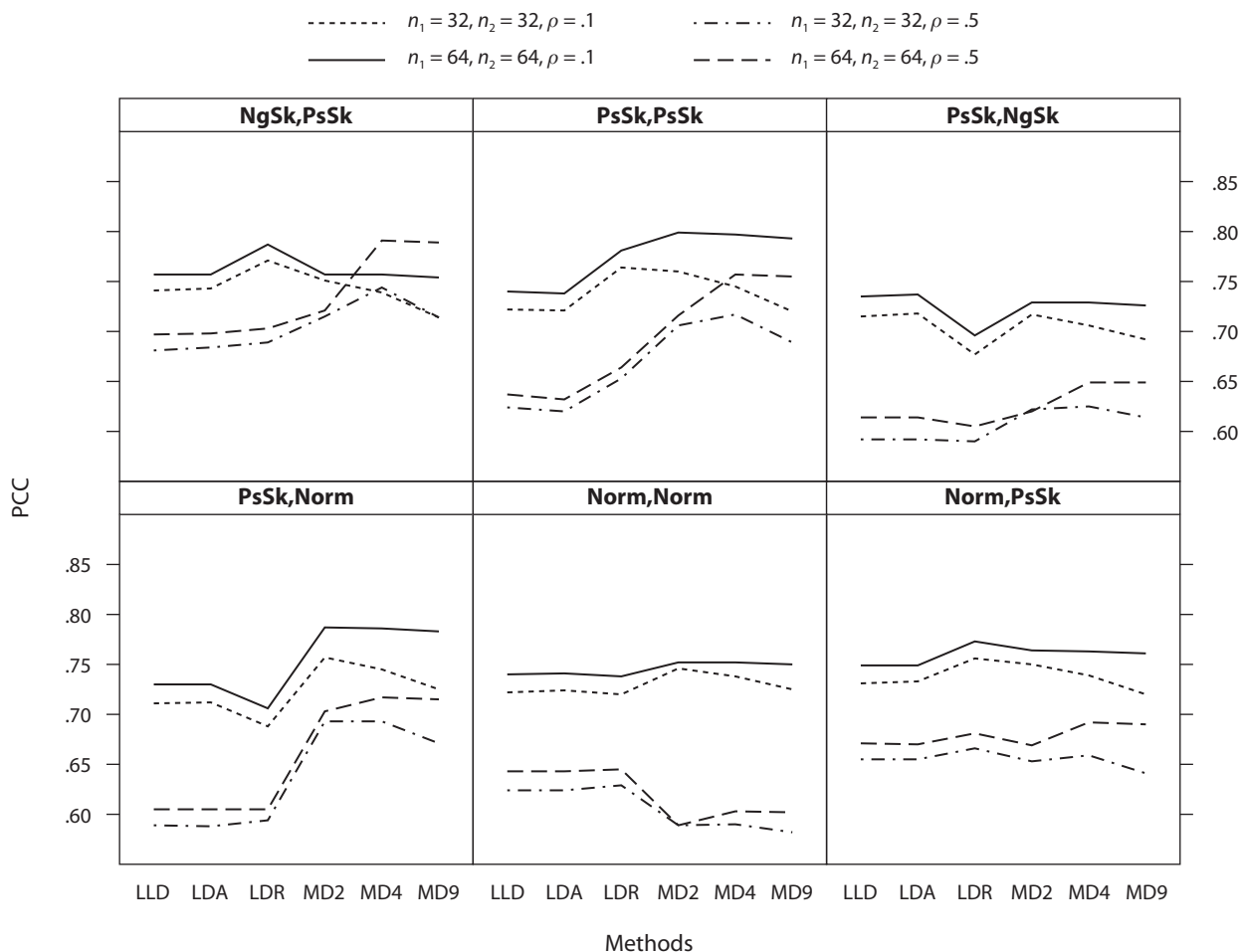


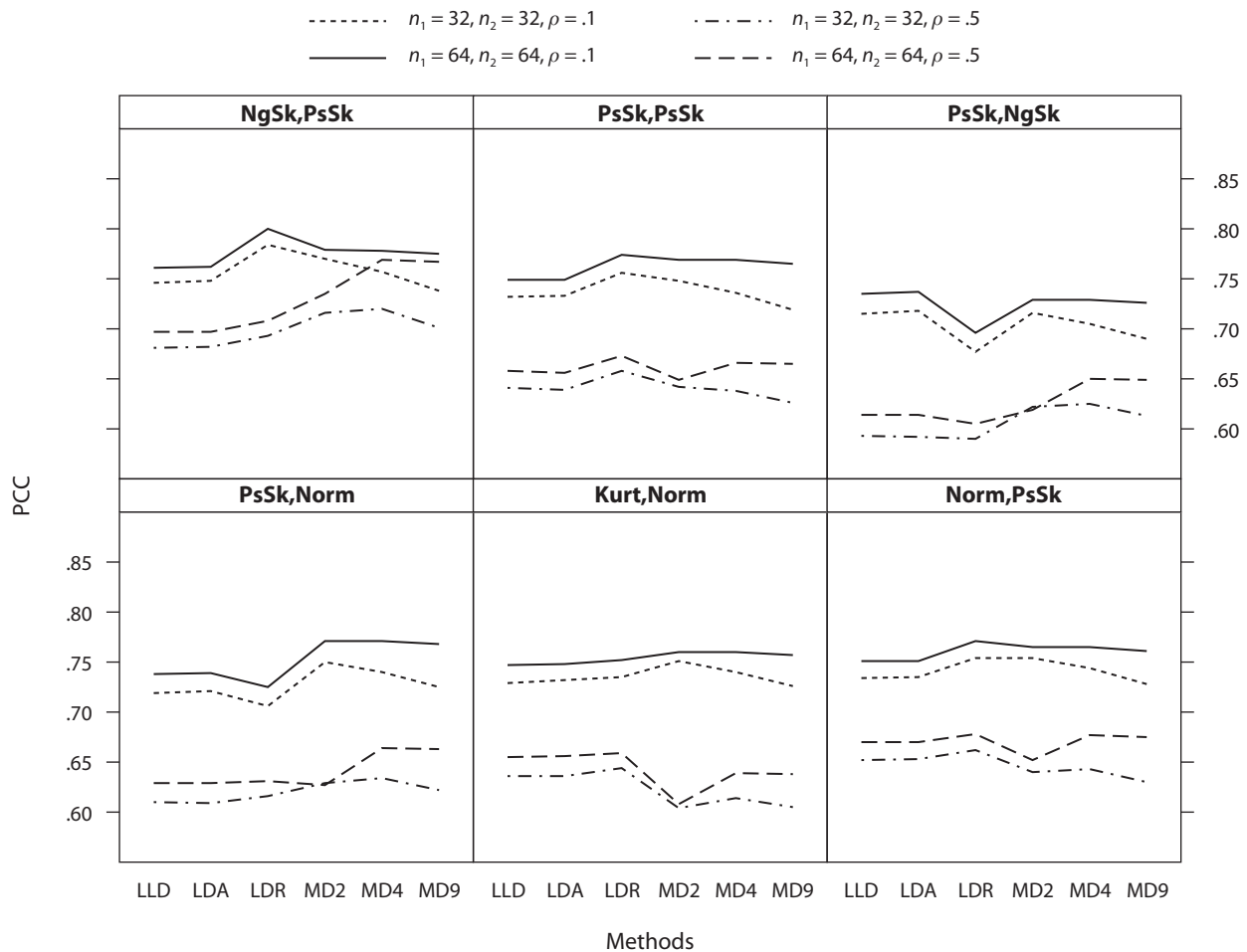
Figure 3. Mean proportions of correct classifications (PCCs) in nonnormality conditions with small values of kurtosis as a function of equal sample sizes across Groups 1 and 2 ( $n_1$  and  $n_2$ , respectively), predictor intercorrelations ( $\rho$ ), and methods for discriminant analysis. LLD, linear logistic discrimination; LDA, linear discriminant analysis; LDR, linear discriminant analysis based on ranks; MD2, mixture discriminant analysis (MDA) with maximum of two within-group components; MD4, MDA with maximum of four within-group components; MD9, MDA with maximum of nine within-group components. Condition abbreviations are as in Figure 1.

sizes in scenarios with small and large values of kurtosis, respectively. In Figure 5, the mean PCCs are generally larger when compared with the corresponding results in Figure 3. Recall that in the simulation conditions in Figure 5, the prior probabilities are assumed equal to the sample proportions, and this information is capitalized on to yield larger mean PCCs for the conditions in Figure 5 relative to those in Figure 3. That is, because one group is known to be more likely a priori within the conditions in Figure 5, this information can be used to generally attain higher PCCs relative to those in Figure 3. Also, the differences on the mean PCC are not as large for different values of  $\rho$  when compared with the equal-sample-size conditions in Figures 3 and 4. Thus, Figures 5 and 6 provide evidence that different predictor intercorrelations are not as consequential with respect to the PCC when sample sizes and prior probabilities are more discrepant.

In Figures 5 and 6, a number of scenarios, especially in the conditions in which  $\rho = .5$ , yield PCCs that are near that attained by random classification to groups on the

basis of prior probabilities. That is, given the prior probability of .75 for the larger group in Figures 5 and 6, a PCC of .75 could be obtained by this simple classification method. Consequently, results for the mean PCC in these figures of .75 or less should be treated with caution. Given that the selected parameter values are indeed representative of research scenarios in applied research, in some scenarios in Figures 5 and 6, methods for discriminant analysis appear to be of little use over random classification. Other scenarios in Figures 5 and 6 do yield values for the PCC appreciably above .75, however, and we focus on these situations in the rest of this section.

Note that, in Figure 5, the  $n_1 = 24, n_2 = 72$  condition for each value of  $\rho$  appears to be absent from the plots for the following conditions: NgSk,PsSk; PsSk,NgSk; and Norm, Norm. In fact, the  $n_1 = 24, n_2 = 72$  conditions are present in Figure 5, but are overlapping with their corresponding  $n_1 = 72, n_2 = 24$  conditions. Thus, whether Group 1 or 2 had the smaller sample size had no effect on the mean PCC in these scenarios. The general pattern



**Figure 4.** Mean proportions of correct classifications (PCCs) in nonnormality conditions with large values of kurtosis as a function of equal sample size across Groups 1 and 2 ( $n_1$  and  $n_2$ , respectively), predictor intercorrelations ( $\rho$ ), and methods for discriminant analysis. Kurt, leptokurtic marginal distributions. All other abbreviations are as in Figure 3.

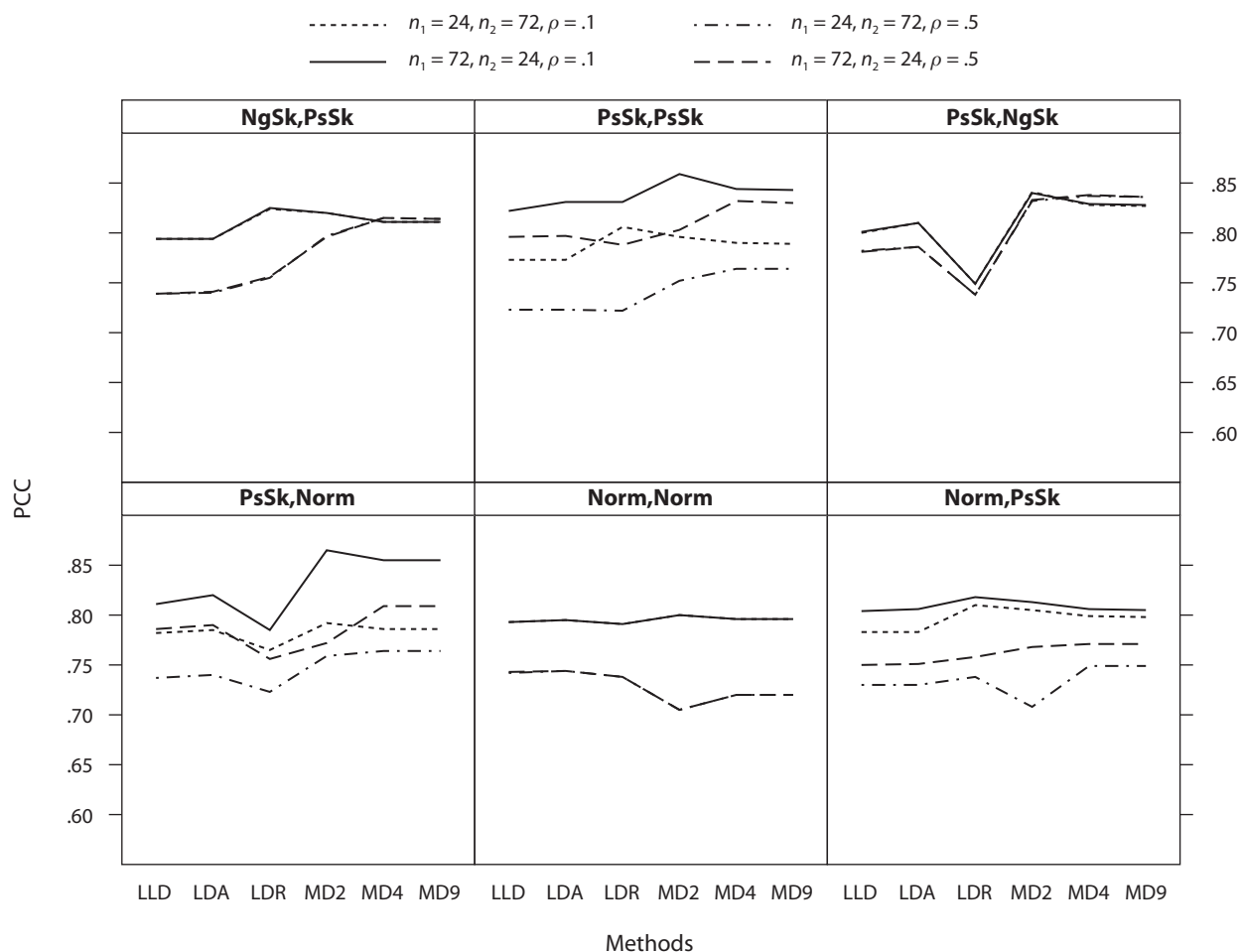
of results in Figure 5 is similar to the pattern of results for equal group sample sizes in Figure 3, with the exception that the differences among methods on the mean PCC are generally smaller across conditions. LDA and LLD again yielded higher mean PCCs in the Norm, Norm condition with  $\rho = .5$  relative to MD2 and MD4. Although LDR yielded the highest mean PCC in some conditions in which  $\rho = .1$ , its advantage over competing methods for discriminant analysis was again not important from a practical perspective. As was the case in Figure 3, an appreciable advantage for MD2 or MD4 over linear methods for discriminant analysis was observed for  $\rho = .5$  in the NgSk,PsSk conditions, PsSk,PsSk conditions, and PsSk,NgSk conditions, and for  $\rho = .1$  in the PsSk, Norm conditions. Consequently, MDA still provides an advantage over competing methods for some scenarios with non-normal predictors when group sample sizes are unequal.

The results in Figure 6 are generally consistent with those in Figure 5, except that the differences between methods tended to be smaller for scenarios in which MDA provided an appreciable advantage over competing meth-

ods. That is, larger values of kurtosis tended to decrease any differences between the investigated methods with respect to classification accuracy. Consequently, all the methods for discriminant analysis yielded similar results in Figure 6, at least from a practical perspective, with the possible exception of the advantage for MD4 over competing methods in the NgSk,PsSk conditions with  $\rho = .5$ .

**Summary**

In general, the results in Figures 3–6 illustrate that, many times, LDA and LLD are not optimal in research scenarios with nonnormal predictors, especially for skewed predictors with relatively small values for kurtosis. LDR can provide an advantage over competing methods in a limited number of conditions with smaller predictor intercorrelations, although any advantage exhibited by LDR in the present study was not practically important. MDA, via MD2 or MD4, was demonstrated to be a useful alternative to LDA, LDR, and LLD for obtaining high mean PCCs in a number of conditions with skewed predictors, especially when kurtosis was relatively small.



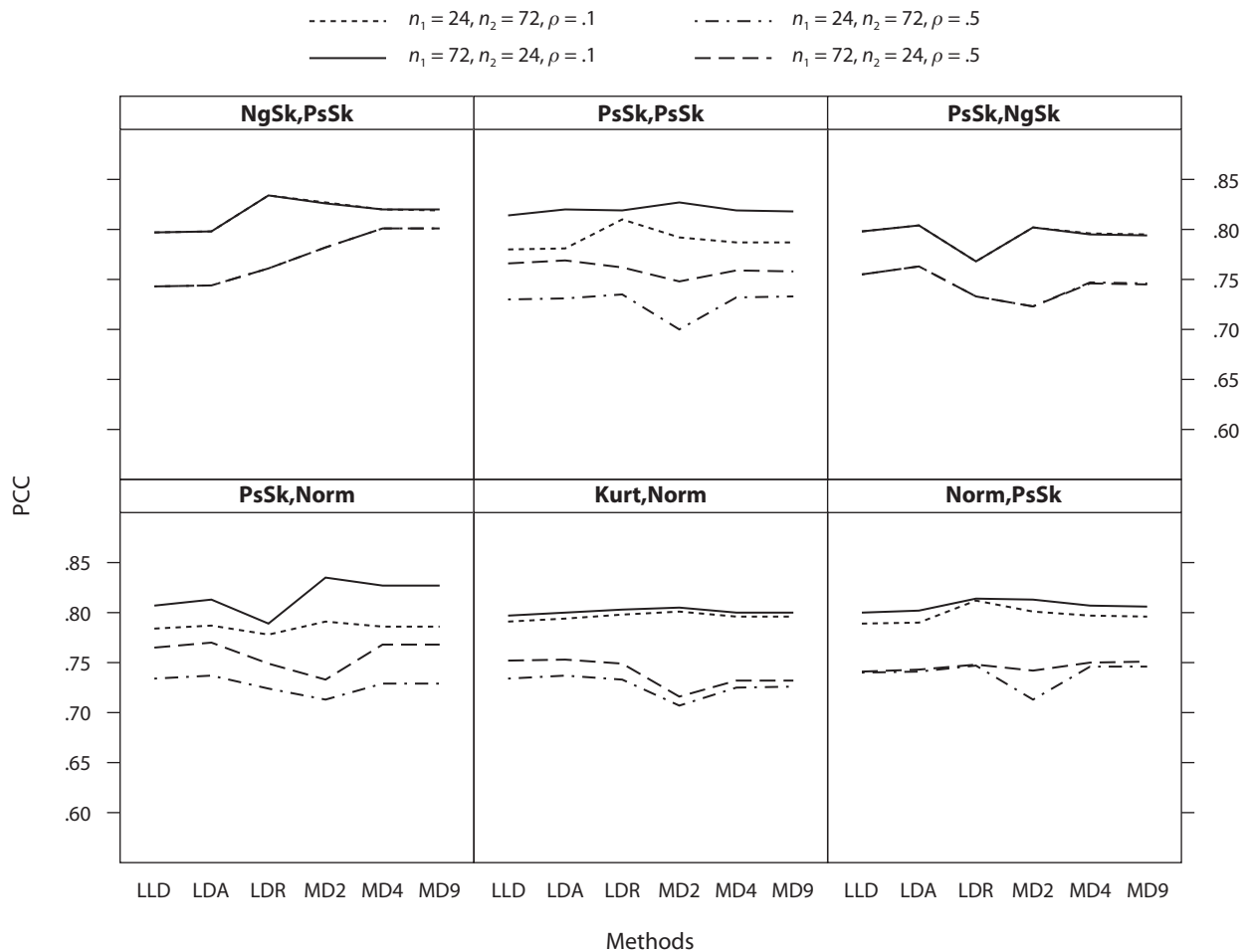
**Figure 5.** Mean proportions of correct classifications (PCCs) in nonnormality conditions with small values of kurtosis as a function of unequal sample sizes across Groups 1 and 2 ( $n_1$  and  $n_2$ , respectively), predictor intercorrelations ( $\rho$ ), and methods for discriminant analysis. All abbreviations are as in Figure 3.

## DISCUSSION

In the present study, we compared linear and mixture models for discriminant analysis with respect to classification accuracy under nonnormality. It was generally shown that LDA and LLD are not optimal procedures for discriminant analysis in the presence of skewed predictors, particularly when kurtosis is relatively small. Although LDR may be useful in a limited number of situations, our results indicate that MDA is generally a more viable approach for discriminant analysis in scenarios with non-normal predictors, given that the number of within-group components is relatively small. That is, if the default maximum value of 9 for the number of within-group components is used in `mclustDAtrain()`, problems of classification inaccuracy can occur, whereas maximum values of 2 or 4 yielded more accurate results in the scenarios investigated here. Interestingly, it appears that relying solely on the Bayesian information criterion for the selection of the number of within-group components without specifying an appropriate maximum for the number of within-group components can yield inaccurate predictions, particularly

when the group sample sizes are small relative to the number of predictors. Consequently, researchers should provide clear evidence for choosing a large number of within-group components when using mixture models for prediction in discriminant analysis.

The results of previous research on the optimality of LDA in situations with nonnormal predictors were generally consistent with the results in the present study. That is, greater deviations from normality tended to be more problematic for this method, at least with respect to skewness, and, given fixed values of skewness, larger kurtoses tended to yield situations in which LDA was more robust, because such distributions become increasingly less asymmetric for larger values of kurtosis. Other results in the present study were not necessarily consistent with previous research and conclusions. First, the results for LDA and LLD did not differ, for all practical purposes, in the present study. Some, although not all, previous research has suggested that LLD is more robust than LDA with respect to nonnormality, particularly in the presence of substantial skewness. Although the nonnormality simulated in the present study may not have been extreme enough



**Figure 6.** Mean proportions of correct classifications (PCCs) in nonnormality conditions with large values of kurtosis as a function of unequal sample sizes across Groups 1 and 2 ( $n_1$  and  $n_2$ , respectively), predictor intercorrelations ( $\rho$ ), and methods for discriminant analysis. All abbreviations are as in Figure 4.

to observe such differences between LDA and LLD, the skewness simulated was reasonable, given the results of Micceri (1989). Furthermore, the sample sizes used in the Monte Carlo study may not have been large enough for LLD to yield an advantage in classification accuracy over LDA. The sample sizes used were reasonable, however, for studies based on supervised classification in many applied research settings. Consequently, the results of the present study suggest that LLD and LDA may often provide comparable results in applied research in the presence of nonnormal predictors.

The results for LDR were especially interesting, given the scarce body of empirical results on the utility of this method. Our results indicated that LDR was most advantageous, relative to LDA and LLD, for situations with small predictor intercorrelations in which the predictors were either skewed in the same direction across groups, consistent with the study of Barón (1991), or skewed in opposite directions across groups, with the negatively skewed population having the higher mean. Even in these situations, however, LDR yielded only modest gains over MDA. Consequently, the present study indicated that

there is little evidence to recommend LDR over LDA, LLD, or MDA, depending on the underlying characteristics of the populations of interest, for the conditions investigated here.

As mentioned previously, the results of Ashikaga and Chang (1981) suggested “that a more important issue than nonnormality is whether the distributions of the two populations are similar in shape” (p. 680) when assessing the robustness of LDA with respect to classification accuracy. The results of the present study suggest that this statement is true within the Norm, Norm and Kurt, Norm conditions, because little change in the mean PCCs for LDA occurred across these conditions (see, e.g., Figures 3 and 4). The results of the present study also suggest that, for example, this is true for the NgSk, PsSk conditions with  $\rho = .5$ , which had distributions skewed in opposite directions when comparing LDA on the mean PCC with the corresponding Norm, Norm condition. The difference between the mean PCCs for LDA in the NgSk, PsSk and Norm, Norm conditions was not as substantial, however, for  $\rho = .1$ . Also, when distributions were both highly skewed in the same manner, as in the PsSk, PsSk conditions, the mean

PCC for LDA changed little relative to the Norm, Norm condition. However, LDA can still be substantially suboptimal in the NgSk,PsSk and PsSk,PsSk conditions, at least when compared with MDA. That is, the conclusions of Ashikaga and Chang held true in the present study when comparing LDA across nonnormal conditions; yet, this did not necessarily lead to LDA being the optimal method in such situations. Rather, other methods, such as MDA, can provide more accurate classification for nonnormal predictors, even across conditions in which the mean PCC for LDA does not change appreciably.

Another alternative for dealing with nonnormal predictors was not investigated in the present study—that is, transforming predictors to normality or near normality (see, e.g., Beauchamp, Folkert, & Robson, 1980). Transformation to normality may be a reasonable option if the researcher knows the appropriate transformation (see, e.g., Beauchamp et al., 1980), but it is also important to heed the question posed by Chinganda and Subrahmaniam (1979): “Is it preferable not to transform at all to normality or should one resort to transformation with the hope that it is the correct one?” (p. 76). Furthermore, in the present study, situations that yielded the largest advantage for MDA over the other methods—such as the NgSk,PsSk; PsSk,NgSk; and PsSk, Norm conditions—are not amenable to a single transformation. Rather, in these conditions, different transformations would be needed depending on the group of interest, raising an important practical issue for prediction, because the group that an individual is in is not known a priori. In our opinion, these issues related to predictor transformation make this approach less than optimal, at least when the goal is to improve classification accuracy.

Future research using mixture models in the context of discriminant analysis could include an approach that allows for nonlinear relationships among the predictors. Although another viable method for modeling nonlinear relationships consists of classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984), MDA could be a useful approach to this problem, given the ability of mixture models to approximate nonlinear relationships among variables (see Bauer & Curran, 2004, for a recent example of this). Further study of this problem could be especially interesting in the context of classification issues.

It is clear that classification is an important procedure, given the need to correctly sort objects into groups in various scientific domains. Consequently, the methods used for this problem should be efficient and robust as different substantive questions arise. We have shown that nonnormality can appreciably affect the optimality of popular methods for discriminant analysis, such as LDA and LLD, with respect to classification accuracy. Moreover, we have provided evidence that MDA, with a relatively small number of within-group components, can achieve relatively high rates of classification accuracy in the presence of nonnormal predictors, particularly when the predictors are skewed with kurtoses that are relatively small. Consequently, in such situations, applied researchers should seriously consider MDA with a relatively small number of within-group components as a method for increasing classification accuracy.

## AUTHOR NOTE

Correspondence concerning this article should be addressed to J. R. Rausch, Cincinnati Children's Hospital Medical Center, MLC 3015, 3333 Burnet Ave., Cincinnati, OH 45229-3039 (e-mail: joseph.rausch@cchmc.org).

## REFERENCES

- AGRESTI, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.
- ANDERSON, J. A. (1975). Quadratic logistic discrimination. *Biometrika*, *62*, 149-154.
- ASHIKAGA, T., & CHANG, P. C. (1981). Robustness of Fisher's linear discriminant function under two-component mixed normal models. *Journal of the American Statistical Association*, *76*, 676-680.
- AVLONITIS, G. J., HART, S. J., & TZOKAS, N. X. (2000). An analysis of product deletion scenarios. *Journal of Product Innovation Management*, *17*, 41-56.
- BALAKRISHNAN, N., & KOCHERLAKOTA, S. (1985). Robustness to non-normality of the linear discriminant function: Mixtures of normal distributions. *Communications in Statistics—Theory & Methods*, *14*, 465-478.
- BANFIELD, J. D., & RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*, 803-821.
- BARÓN, A. E. (1991). Misclassification among methods used for multiple group discrimination—The effects of distributional properties. *Statistics in Medicine*, *10*, 757-766.
- BAUER, D. J., & CURRAN, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3-29.
- BEAUCHAMP, J. J., FOLKERT, J. E., & ROBSON, D. S. (1980). A note on the effect of logarithmic transformation on the probability of misclassification. *Communications in Statistics—Theory & Methods*, *9*, 777-794.
- BIERNACKI, C., & GOVAERT, G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation & Simulation*, *64*, 49-71.
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.
- CATTS, H. W., FEY, M. E., ZHANG, X., & TOMBLIN, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, & Hearing Services in Schools*, *32*, 38-50.
- CELEUX, G., & GOVAERT, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*, 781-793.
- CHINGANDA, E. F., & SUBRAHMANIAM, K. (1979). Robustness of the linear discriminant function to nonnormality: Johnson's system. *Journal of Statistical Planning & Inference*, *3*, 69-77.
- COHEN, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- CONOVER, W. J., & IMAN, R. L. (1980). The rank transformation as a method of discrimination with some examples. *Communications in Statistics—Theory & Methods*, *9*, 465-487.
- EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, *70*, 892-898.
- EVERITT, B. S., & HAND, D. J. (1981). *Finite mixture distributions*. London: Chapman & Hall.
- FAN, X., & WANG, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *Journal of Experimental Education*, *67*, 265-286.
- FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179-188.
- FLEISHMAN, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.
- FLOWERS, C. P., & ROBINSON, B. (2002). A structural and discriminant analysis of the Work Addiction Risk Test. *Educational & Psychological Measurement*, *62*, 517-526.
- FRALEY, C., & RAFTERY, A. E. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification*, *16*, 297-306.
- FRALEY, C., & RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611-631.

- FRALEY, C., & RAFTERY, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *Journal of Classification*, **20**, 263-286.
- FRALEY, C., & RAFTERY, A. E. (2006). *MCLUST version 3 for R: Normal mixture modeling and model-based clustering* (Tech. Rep. No. 504). Seattle: University of Washington, Department of Statistics.
- GLASER, B. A., CALHOUN, G. B., & PETROCELLI, J. V. (2002). Personality characteristics of male juvenile offenders by adjudicated offenses as indicated by the MMPI-A. *Criminal Justice & Behavior*, **29**, 183-201.
- HASTIE, T., & TIBSHIRANI, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society B*, **58**, 155-176.
- HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- HUBERTY, C. J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin*, **95**, 156-171.
- HUBERTY, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.
- JO, H., HAN, I., & LEE, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems With Applications*, **13**, 97-108.
- JOHNSON, R. A., & WICHERN, D. W. (2002). *Applied multivariate statistical analysis* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
- KIANG, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems*, **35**, 441-454.
- LACHENBRUCH, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, **23**, 639-645.
- LACHENBRUCH, P. A., SNEERINGER, C., & REVO, L. T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics—Theory & Methods*, **1**, 39-56.
- LEI, P., & KOEHLI, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *Journal of Experimental Education*, **72**, 25-49.
- LIM, T.-S., LOH, W.-Y., & SHIH, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, **40**, 203-228.
- MAHALANOBIS, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, **2**, 49-55.
- MCLACHLAN, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.
- MCLACHLAN, G. J., & BASFORD, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Dekker.
- MCLACHLAN, G. J., & BYTH, K. (1979). Expected error rates for logistic regression versus normal discriminant analysis. *Biometrical Journal*, **21**, 47-56.
- MCLACHLAN, G. J., & PEEL, D. (2000). *Finite mixture models*. New York: Wiley.
- MICCERI, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, **105**, 156-166.
- NAKANISHI, H., & SATO, Y. (1985). The performance of the linear and quadratic discriminant functions for three types of non-normal distribution. *Communications in Statistics—Theory & Methods*, **14**, 1181-1200.
- OSTRANDER, R., WEINFURT, K. P., YARNOLD, P. R., & AUGUST, G. J. (1998). Diagnosing attention deficit disorders with the Behavioral Assessment System for Children and the Child Behavior Checklist: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting & Clinical Psychology*, **66**, 660-672.
- PANEL ON DISCRIMINANT ANALYSIS, CLASSIFICATION, AND CLUSTERING (1989). Discriminant analysis and clustering. *Statistical Science*, **4**, 34-69.
- PHILLIPS, M., CATANEO, R. N., CUMMIN, A. R. C., GAGLIARDI, A. J., GLEESON, K., GREENBERG, J., ET AL. (2003). Detection of lung cancer with volatile markers in the breath. *Chest*, **123**, 2115-2123.
- PRESS, S. J., & WILSON, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of American Statistical Association*, **73**, 699-705.
- RAWLINGS, R. R., FADEN, V. B., GRAUBARD, B. I., & ECKARDT, M. J. (1986). A study on discriminant analysis techniques applied to multivariate lognormal data. *Journal of Statistical Computation & Simulation*, **26**, 79-100.
- R DEVELOPMENT CORE TEAM (2007). *R: A language and environment for statistical computing* (Version 2.6.0). Vienna: R Foundation for Statistical Computing.
- RUDOLPH, P. M., & KARSON, M. (1988). The effect of unequal priors and unequal misclassification costs on MDA. *Journal of Applied Statistics*, **15**, 69-83.
- SAHINER, B., CHAN, H. P., ROUBIDOUX, M. A., HELVIE, M. A., HADJISKI, L. M., RAMACHANDRAN, A., ET AL. (2004). Computerized characterization of breast masses on three-dimensional ultrasound volumes. *Medical Physics*, **31**, 744-754.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.
- SILVA, A. P. D., STAM, A., & NETER, J. (2002). The effects of misclassification costs and skewed distributions in two-group classification. *Communications in Statistics—Simulation & Computation*, **31**, 401-423.
- STUART, A., & ORD, J. K. (1994). *Kendall's advanced theory of statistics: Distribution theory* (Vol. 1, 6th ed.). New York: Wiley.
- SUBRAHMANIAM, K., & CHINGANDA, E. F. (1978). Robustness of the linear discriminant function to nonnormality: Edgeworth series distribution. *Journal of Statistical Planning & Inference*, **2**, 79-91.
- TAXT, T., HJORT, N. L., & EIKVIL, L. (1991). Statistical classification using a linear mixture of two multinormal probability densities. *Pattern Recognition Letters*, **12**, 731-737.
- TITTERINGTON, D. M., SMITH, A. F. M., & MAKOV, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, U.K.: Wiley.
- UDRIS, E. M., AU, D. H., McDONELL, M. B., CHEN, L.-W., MARTIN, D. C., TIERNEY, W. M., & Fihn, S. D. (2001). Comparing methods to identify general internal medicine clinic patients with chronic heart failure. *American Heart Journal*, **142**, 1003-1009.
- VALE, C. D., & MAURELLI, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, **48**, 465-471.
- VENABLES, W. N., & RIPLEY, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
- WELCH, B. L. (1939). Note on discriminant functions. *Biometrika*, **31**, 218-220.

(Manuscript received March 25, 2008;  
revision accepted for publication September 13, 2008.)