Journal of Educational and Behavioral Statistics
Summer 2004, Vol. 29, No. 2, pp. 219–240

Kernel-Based Discriminant Techniques for Educational Placement

Miao-hsiang Lin
Su-yun Huang
Yuan-chin Chang
Institute of Statistical Science, Academia Sinica

This article considers the problem of educational placement. Several discriminant techniques are applied to a data set from a survey project of science ability. A profile vector for each student consists of five science-educational indictors. The students are intended to be placed into three reference groups: advanced, regular, and remedial. Various discriminant techniques, including Fisher's discriminant analysis and kernel-based nonparametric discriminant analysis, are compared. The evaluation work is based on the leaving-one-out misclassification score. Results from the five school data sets and 500 bootstrap samples reveal that the kernel-based nonparametric approach with bandwidth selected by cross validation performs reasonably well. The authors regard kernel-based nonparametric procedures as desirable competitors to Fisher's discriminant rule for handling problems of educational placement.

Keywords: *classification, data-driven bandwidth selection approaches, educational placement, Fisher's discriminant analysis, generalized* k*th-nearest-neighbor method, science-education indicators*

Fisher's discriminant analysis (DA), known in the fields of education and applied psychology as a "trait-space model" (Cooley & Lohnes, 1971; Tatsuoka, 1971), has proven useful for identifying the structural dimensions along which groups differ. With respect to placement problems, however, it has not been very fruitful. The effectiveness of its classification role is jeopardized by two measurement caveats (Nunnally, 1978). One is associated with the difficulty involved in obtaining mutually exclusive groups when membership is designated by a quantitative attribute, while the other is associated with the use of a set of discriminant variables that measure only cognitive attributes, such as scores on ability, aptitude, and achievement tests. Thus, the classification effectiveness of Fisher's DA might be enhanced in a

This study was supported by the Mathematics, Science, and Technology Education Division of the National Science Council of the Republic of China, Taiwan. We would like to express sincere thanks to the editor and an associate editor for their careful reading and instructive comments, suggestions, and input on this article. Thanks are also extended to Professor Yeong-Jing Cheng for his help in data aggregation from five science education projects. All data used in this article are available electronically via e-mail at miao@stat.sinica.edu.tw.

measurement situation wherein exclusive groups are obtainable and the attributes measured by the discriminant variables are a combination of both cognitive and noncognitive ones. This suggests that the classification role of Fisher's DA should be reevaluated, since the proposed measurement situation has its application in the current practice of building an educational indicator system for monitoring schooling outcomes.

Monitoring students' school performance in science (Murnane & Raizen, 1988; Shavelson, Carey, & Webb, 1990; Shavelson, McDonnell, Oakes, Carey, & Picus, 1987; Smith, 1988) requires that the educational indicators used contain cognitive and noncognitive determinants that work together to influence performance. Such a set of indicators is considered crucial for tackling two central educational problems: learning diagnosis and instructional placement. The purpose of this study was to examine the usefulness of discriminant techniques in allocating students into instructional-curriculum coherence programs in compliance with their learning progress levels (Newmann, Smith, Allensworth, & Bryk, 2001).

For the purpose of instructional groupings, it is a common practice to place students into remedial-, regular-, and advanced-curriculum programs based on grade point average (GPA) distributions. GPAs, as accumulated overall performance on intellectual and qualitative measures, are considered suitable in designating group memberships. Moreover, the designation procedure is tailored to the long-standing recognition that most students are of average achievement level, and very few students have extremely high or extremely low achievement levels. Consequently, this study focuses on a three-group classification problem in which the sizes of the groups are grossly different. Therefore, within the indicator system just mentioned, we were interested in comparing a student's profile vector with three reference groups, where profile vector refers to the student's scores on a set of five science-education indicators (as described subsequently). Since the five indicators constitute an integrated composition within an educational indicator system, we treated them as a necessary set of discriminant variables, bypassing discussion of the variable selection problem.

As shown subsequently, the measurement levels associated with the five indicators are ordinal, resulting in the data vectors deviating from a multivariate normal distribution. Thus, we used Fisher's parametric as well as nonparametric discriminant rules for our placement problem, where nonparametric discriminant rules refer to procedures based on kernel-based density estimation ideas (Hand, 1982; Silverman, 1986). Titterington's work (Titterington et al., 1981) reveals that, when applied to a two-group medical diagnosis problem, the nonparametric kernel-based procedure does not perform well in terms of classification error rates. The specific characteristics of the class and data structures of our classification setting, however, might render perspectives different from those of Titterington.

For example, in the medical context of discriminant analysis, a patient's disease to some extent has a biological or physical root. The class structure of the training data set can achieve a relatively clear-cut division provided that diagnosis by clinicians or automatic examinations is correct. In contrast, students' school performance, as indicated by GPA records, largely reflects psychological or mental abilities. To obtain a mutually exclusive class structure, one must compromise some indifference

zones. In addition, as mentioned earlier, in the school setting only extremely slow and exceptionally able learners are considered for placement in remedial- or advanced-curriculum programs, respectively, and the bulk of students of average ability remain in the regular program. Therefore, the training data set of this study was designed to reflect such a class structure.

Characteristics of Data Structure

This study analyzed the data collected from a large-scale project commissioned by the National Science Council of the Republic of China for developing educational indicators to monitor and upgrade Taiwan's elementary and secondary science curricula (Cheng et al., 1994). The five indicators of science education, developed for assessing sixth-grade students, are the nature science test, four questionnaire scales on students' interest in and attitudes toward science, teachers' dedication, parents' educational levels, and hours spent on homework. Sixth graders from 30 classrooms were distributed throughout five public schools in Taipei City and Taipei County. Information on each sample school included students' scores on the five indicators and their GPAs. The distribution of GPAs was used to designate students' group membership according to the respective school norms.

The designation procedure was designed to obtain three relatively exclusive groups by separating students in the bottom 5% and top 5% of the GPA distribution from those in the remaining 90%. This procedure is frequently used in the selection of low- and high-ability students. In addition, students with GPAs in the indifference zones—between the 90% and 95% quantiles and between the 5% and 10% quantiles—were excluded from the study. This was done because no agreement could be reached on how to identify students falling in these zones (Glass, 1978). Following this procedure, the remaining students within each school were identified as the three reference groups in compliance with placement into the remedial-, regular-, or advanced-curriculum programs. Hence, the students whose group memberships were established were treated as a training data set with known membership status.

Table 1 provides a description of the classification and discriminating variables, including group categories and scale values. Table 2 presents training set sample sizes for the five schools, including cell frequencies broken down by groups.

TABLE 1
Description of Classification and Discriminating Variables

Variable type		Description	Group category or scale value
Classification	GPA	Grade point average	1, 2, 3
Discriminating	NST	Nature science test	0–32
	ATT	Interest/attitude toward science	0–5
	TTE	Teachers' teaching endeavor	0–10
	HHS	Hours of homework study	0–12
	FME	Parents' education levels	1–6

TABLE 2
Training Sample Sizes of the Five Schools and Cell Frequencies Distributed in the Respective Three Groups

School	Group 1: remedial (bottom 5%) (n_1)	Group 2: regular (middle 10%–90%) (n_2)	Group 3: advanced (top 5%) (n_3)	N
1	6	84	9	99
2	9	101	10	120
3	12	166	22	200
4	14	202	14	230
5	22	309	26	357

As can be seen in Tables 1 and 2, the data structure of this study was characterized by ordinal data vectors and gross differences in group sizes. These two characteristics to some extent violate the normal assumptions underlying Fisher's DA. Therefore, the aim of this study was to apply the kernel-based nonparametric discriminant analysis to the educational placement data. The nonparametric approach neither lays ground on stringent distributional assumptions nor sets any parametric form of separating surface for groups. Practically, the nonparametric method allows a flexible separating surface driven by the training data.

The following section contrasts methods of density estimation involved in the parametric and nonparametric classification procedures. The theoretical aspects subsumed under both procedures have been well developed; therefore, this study, focusing on selecting methods for data analysis, addressed them in a more heuristic manner. Readers interested in detailed derivations may refer to the services listed.

In addition, the following description emphasizes the introduction of kernel-based density estimators, since the fields of education and applied psychology have yet to make much use of them. Moreover, for the sake of ease of use, we introduce three estimators based on data-driven bandwidth selection approaches, namely, methods of least squares and likelihood cross-validation and the generalized *k*th-nearest-neighbor method. Here we adopt classical methods of bandwidth selection rather than more recently developed counterparts (see the lists of Jones, Marron, & Sheather, 1996; Sain, Baggerly, & Scott, 1994; and Park & Turlach, 1992), because the asymptotic behaviors of those approaches might be only minimally in effect for such relatively small samples. (For example, we observed that the biased cross-validation score functions for the three groups in each school data set were strictly decreasing. Consequently, the minimizing h_g could not be located for the three groups.)

Comparison Framework

Brief Comparison of Density Estimation Methods

Corresponding to the terminology of discriminant analysis, classification of students into one of the three instructional groups was based on the largest value associated with the posterior probabilities. The posterior probability, denoted by $p(g|\mathbf{x})$, takes the form $p(g|\mathbf{x}) = f_g(\mathbf{x}) \Big/ \sum_{g=1}^{3} f_g(\mathbf{x})$, where g is a subscript to distinguish the groups, $\mathbf{x}$ is a p-element profile vector containing the discriminant variable scores of a student, and $f_g(\mathbf{x})$ is the group-conditional density function at **x.** Thus, computation of $p(g|\mathbf{x})$ relies on the methods used to estimate the group-conditional density at point **x,** $f_g(\mathbf{x})$.

Fisher's discriminant approach (Fisher, 1936; Rao, 1973) applies the linear or quadratic classification rule to compute an estimate $\hat{f}_g(\mathbf{x})$ of $f_g(\mathbf{x})$ from the training data set. Both criteria are based on the Mahalanobis generalized squared distance, denoted by $d_g^2 = (\mathbf{x} - \bar{\mathbf{x}}_g)^T \mathbf{C}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_g)$, where $\bar{\mathbf{x}}_g$ is the sample mean vector in group g and $\mathbf{C}$ refers to a pooled sample covariance matrix or within-group covariance matrices. Under a multivariate normal theory, the exact expression of the estimate $\hat{f}_g(\mathbf{x})$ is

$$\hat{f}_g(\mathbf{x}) = (2\pi)^{-p/2} |\mathbf{C}|^{-1/2} \exp\left(-0.5 d_g^2\right). \tag{1}$$

The nonparametric version of Fisher's DA used in this study refers to application of the kernel-based density estimator (Hand, 1981, 1982; Silverman, 1986) to estimate $f_g(\mathbf{x})$. Here we drop g for a discussion in the general setting. The multivariate kernel-based density estimator is defined as

$$\hat{f}(\mathbf{x}) = \frac{1}{n\, h^p} \sum_{i=1}^{n} K[(\mathbf{x} - \mathbf{X}_i)/h], \tag{2}$$

where $\mathbf{X}_i$ is a p-dimensional observed data vector, $K(\mathbf{z})$ is a p-dimensional kernel function, and h is the bandwidth (or smoothing parameter or window width). The density estimate $\hat{f}(\mathbf{x})$ yielded by this definition is composed of the smoothed version of the true density plus random error (Rosenblatt, 1956; Whittle, 1958), where the smoothed version of the true density is the expected value of density estimate, $E\hat{f}(\mathbf{x})$. Because the bias, $E\hat{f}(\mathbf{x}) - f(\mathbf{x})$, directly depends on the smoothing parameter h, the crucial work in the estimation process is to choose an appropriate h to minimize the approximate mean integrated square error, denoted by $\mathrm{MISE}[\hat{f}(\mathbf{x})] = E\int [\hat{f}(\mathbf{x})] - f(\mathbf{x})]^2$, which consists of the squared bias and variance components associated with the use of Equation 2 to estimate true density $f(\mathbf{x})$.

To avoid a choice of h that is too large or too small, methods of least squares cross validation (LSCV) and likelihood cross validation (LKCV) use only the data themselves to choose the smoothing parameter via different algorithms. The LSCV

method locates the best h by minimizing a score function in the sense of minimizing estimated MISE, where the score function is an estimate of $\int \hat{f}^2(\mathbf{x}) - 2\int \hat{f}(\mathbf{x}) f(\mathbf{x})$ (Bowman, 1984; Rudemo, 1982; Stone, 1984). The LKCV method, in contrast, arrives at the best choice of h that minimizes the Kullback-Leibler information distance between a density estimate and its true density, which is defined as $\int f(\mathbf{x}) \log[f(\mathbf{x})/\hat{f}(\mathbf{x})]$ (Duin, 1976; Habbema, Hermans, & van der Broek, 1974). The appealing aspect of these two methods is that the bandwidth yielded is asymptotically optimal in a certain sense (see Stone, 1982, 1984). For f having continuous second derivatives and satisfying $\int [\nabla^2 f(\mathbf{x})]^2 < \infty$, the optimal bandwidth for an Order 2 kernel estimate is of magnitude $O(n^{-1/(p+4)})$, where p is the dimension of the variable **x.** With a bandwidth of this order of magnitude, the MISE can achieve the rate of convergence $O(n^{-4/(p+4)})$.

The generalized kth-nearest-neighbor (GKNN) estimator is defined by

$$\hat{f}(\mathbf{x}) = \frac{1}{n r_k(\mathbf{x})^p} \sum_{i=1}^{n} K[(\mathbf{x} - \mathbf{X}_i)/r_k(\mathbf{x})], \tag{3}$$

where k, an integer, is the smoothing parameter playing the same role as h to govern the smoothing quantity $r_k(\mathbf{x})$, which is the Euclidean distance from **x** to the kth nearest data point. The GKNN method, as contrasted to Equation 2, possesses a self-adjusting property in that the distance $r_k(\mathbf{x})$ varies with data points, resulting in placing a flatter kernel on data points in regions of low density and a narrower kernel on those in regions of high density. In addition, literature results have shown that, under some conditions on the functional form of k, the GKNN estimator is asymptotically unbiased and consistent (Devroye & Wanger, 1977; Moore & Yackel, 1977).

To summarize, this study employed the estimators of Equations 1, 2, and 3 to compute the group-conditional density estimate $\hat{f}_g(\mathbf{x})$ for the calculation of posterior probability estimates $\hat{p}(g|\mathbf{x})$ evaluated at point **x.** Through comparisons of the relative magnitudes of the three posterior probabilities, a student was assigned as a member of a group g if the largest value of $\hat{p}(g|\mathbf{x})$ was associated with that group.

Generally speaking, the Fisher classification procedure based on Equation 1 is capable of arriving at an optimal decision surface for two normal groups, in the sense of minimizing the number of misclassifications. Moreover, the approach was found robust with respect to its tolerance of some amount of deviation from the normal distribution assumption (Lachenbruch, 1975). However, for three nonnormal groups of grossly different sizes, the extent to which Fisher's approach converges to the desired optimal decision surface is unknown. In this case, the borderline cases might be prone to misclassification.

To ensure a given degree of accuracy in five-dimensional density estimates by the kernel-based estimators (Equation 2 or 3), a moderate sample size is required. Although the sample sizes shown in Table 2 might not be sufficient to ensure a given degree of accuracy in density estimates, this study aimed to compare performances in terms of classification accuracy rather than in terms of accuracy in density esti-

mates. As mentioned earlier, the classification procedure is intended to assign a student to the group to which he or she has the highest probability of belonging. According to the formula $\hat{p}(g|\mathbf{x}) = \hat{f}_g(\mathbf{x}) \Big/ \sum_{g=1}^{3} \hat{f}_g(\mathbf{x})$, what is required is to select the group with the highest density estimate from the trio $\hat{f}_1(\mathbf{x})$, $\hat{f}_2(\mathbf{x})$, and $\hat{f}_3(\mathbf{x})$ rather than to study how close these three density estimates are to their respective true densities. In other words, so long as the effect of the small sample size does not jeopardize the probability of selecting the highest density estimate among the three groups, the decision to assign students to the associated groups is legitimate.

With the use of kernel-based estimators (Equations 2 and 3) to estimate densities, the choice of the kernel $K(\mathbf{z})$ could be made subjectively. The kernels associated with Equations 2 and 3 were chosen as the normal and uniform functions, respectively. This choice was made to allow an easier calculation in terms of locating the window width h or k. Moreover, this different choice between the uniform and normal kernels is legitimate, since the efficiencies of the two kernel functions are almost the same on the basis of the MISE (Epanechnikov, 1969).

As to the choice of h, the computational score functions under the LSCV and LKCV methods were used to obtain a triplet of h_g for the three-group discriminant problem of this study. That is, the values of the bandwidth were adjusted to within-group data structures, but the same single h_g was used in the five dimensions for each group. For the choice of k, the asymptotic value of k minimizing the MISE was proportional to $n_g^{4/(p+4)}$ (Mack & Rosenblatt, 1979), where the constant of proportionality depends on the unknown density function of the target group. Since this theoretical value of k is intractable, we used the selection criterion that k yields the lowest classification error rate based on the training data set (Hand, 1982). Within this analogy, the GKNN method is also a data-driven bandwidth selection approach, since the location of a best triplet of k_g is based on the error rate function calculated from the empirical data set.

Criteria for Performance Evaluation

Evaluation work under the discriminant analysis is based on the misclassification rate, obtained by applying the classification rules derived from the training sample to a test data set or to leaving-one-out cross-validation samples (Lachenbruch & Mickey, 1968). This study treated the samples of the individual schools as shown in Table 2 as the training samples. Since the sample sizes were small, we adopted the leaving-one-out cross-validation procedure via the training sample to calculate different measures of classification error rate (as described subsequently). These performance measures were carried over for each of the five school data sets. In addition, 500 bootstrap samples from the original sample data of the fifth school—which had the largest sample ($n = 357$)—were used to obtain the variances of these error-rate estimates, including the bootstrap means and the 95% confidence intervals.

Here we would like to point out that, in the school setting, the major concern is identifying students suitable for remedial- and advanced-curriculum programs. Therefore, evaluation work based on the error rates of Groups 1 and 3 will be more

important than that based on the Group 2 error rate or the overall error rate. As a result, the performance of the Fisher and kernel-based methods was evaluated via this particular aspect, as well as through an average of individual group error rates.

Here we also mention that evaluation work via the assignment formula $\hat{p}(g|\mathbf{x}) = \hat{f}_g(\mathbf{x}) \Big/ \sum_{g=1}^{3} \hat{f}_g(\mathbf{x})$, does not take any loss function into consideration. This is to carry out an educational policy in which more students with higher-than-average ability may be accepted into advanced-curriculum programs. Another reason is to enhance educational accountability, in that more students with lower-than-average ability may have increased opportunities to take part in remedial-curriculum programs.

Preliminary Analyses and Computional Formulas

Preliminary Data Analyses

The Fisher classification decision was made for a specific student by comparing the distance between the student's position and each of the three group mean vectors to locate the group to which the student most likely belonged. Since the mean vector (the centroid in the five-dimensional space) was used to summarize the position of a group, Table 3 lists the values of the means and standard deviations of the five variables for each of the three groups.

Can be seen in the top part of Table 3, Group 3 (the advanced-curriculum group) has the highest mean scores, while Group 1 (the remedial-curriculum group) has the lowest mean scores and Group 2 (the regular group) falls in the middle. Although the centroids of the groups are distinctive, an overlap of individual students is expected given the large differences in the standard deviations among the three groups (see Table 3, bottom), particularly those associated with variable NST. A visualization of the students' group positions in the pairwise scatterplots of the five variables provided evidence that the three groups' students were not well separated, since there was a great deal of overlap. Given such an overlapping data structure, we were interested in examining the extent to which the Fisher and nonparametric discriminant rules could accomplish a better separation of the students in these groups.

To decide whether to use pooled covariance or within-group covariance matrices in density estimates, we conducted a likelihood ratio test (Morrison, 1976) of the homogeneity of the three within-group covariance matrices for each of the five school data sets. The chi-square values associated with the three larger data sets (3, 4, and 5) were all significant at the .10 level, whereas the chi-square values of the two smaller data sets (1 and 2) were not significant.

Because the likelihood ratio test is not robust to nonnormality (Anderson, 1984; Perlman, 1980), we looked further into the similarity in entries on the group covariance matrices. We relied on a comparison made with heuristic information rather than the individual entries in each cell of the matrices of the large shape. This heuristic information included the natural log of the determinant of the individual within-group and pooled covariance matrices, as well as the generalized squared distances between group mean vectors calculated via the two types of covariance matrices, all of which are listed in Table 4.

TABLE 3

Values of the Group Mean Vectors on the Five Variables for Each of the Five School Data Sets, Including Subgroup Sizes and Standard Deviations

		School 1 ($N = 99$)			School 2 ($N = 120$)			School 3 ($N = 200$)			School 4 ($N = 230$)			School 5 ($N = 357$)		
	Discriminating variable	1 [6]	2 [84]	3 [9]	1 [9]	2 [101]	3 [10]	1 [12]	2 [166]	3 [22]	1 [14]	2 [202]	3 [14]	1 [22]	2 [309]	3 [26]
Mean values	Nature science test (NST)	16.3	21.7	25.9	13.4	19.5	25.0	10.7	19.0	25.2	13.4	20.2	25.1	12.9	20.5	23.7
	Parents' education levels (FME)	4.4	4.5	4.6	3.1	3.7	4.1	2.9	3.6	3.9	2.5	3.8	4.5	3.5	3.7	3.8
	Interest/attitude toward science (ATT)	1.8	3.3	3.6	2.8	2.5	3.2	2.3	2.3	3.1	2.8	2.8	2.9	2.4	3.0	3.3
	Teachers' teaching endeavor (TTE)	7.8	9.1	9.2	6.7	6.6	7.2	7.1	7.8	8.3	7.4	7.8	7.7	6.7	8.0	8.3
	Hours of homework study (HHS)	7.0	7.6	7.1	6.4	7.0	7.3	6.4	6.6	7.5	6.5	6.9	7.4	5.6	7.1	7.8
Standard deviations	Nature science test (NST)	2.1	4.4	3.4	5.2	4.3	3.0	2.5	5.0	3.5	3.4	4.5	1.9	4.7	4.4	3.7
	Parents' education levels (FME)	0.6	0.8	0.8	0.8	0.9	1.1	0.9	1.0	1.1	0.6	1.0	0.6	1.3	1.0	0.8
	Interest/attitude toward science (ATT)	1.3	1.2	1.0	1.0	1.2	1.1	1.5	1.2	1.4	1.2	1.2	1.4	1.3	1.3	1.3
	Teachers' teaching endeavor (TTE)	1.7	1.0	0.7	1.0	1.2	0.9	1.3	1.3	1.1	1.9	1.2	1.3	1.8	1.0	0.7
	Hours of homework study (HHS)	1.9	1.5	1.5	1.9	1.6	1.3	2.2	1.5	1.1	2.3	1.6	1.1	1.8	1.5	1.4

Note. 1 stands for the remedial group, 2 for the regular group, and 3 for the advanced group. Values in brackets are subgroup sizes.

TABLE 4

Natural Log of Determinants of the Within-Group and Pooled Covariance Matrices and Pairwise Generalized Squared Distances Between Group Mean Vectors for Five School Data Sets

			Within-group cov. matrix				Pooled cov. matrix			
			1:	2:	3:	Inf.	1:	2:	3:	Inf.
Data set	Group	n_g	remedial	regular	advanced	$\ln \lvert \mathbf{C}_g \rvert$	remedial	regular	advanced	$\ln \lvert \mathbf{C} \rvert$
School 1	1. Remedial	6	0.3	8.0	14.0	0.3	0.0	4.5	8.3	
($N = 99$)	2. Regular	84	77.3	3.4	3.8	3.4	4.5	0.0	1.1	3.4
	3. Advanced	9	187.4	4.4	1.0	1.0	8.3	1.1	0.0	
School 2	1. Remedial	9	3.3	6.7	23.9	3.3	2.6	2.6	8.6	
($N = 120$)	2. Regular	101	5.3	4.1	5.9	4.1	8.6	0.0	2.3	4.1
	3. Advanced	10	12.7	6.3	1.7	1.7	0.0	2.3	0.0	
School 3	1. Remedial	12	3.1	8.3	23.2	3.1	3.9	3.9	10.4	
($N = 200$)	2. Regular	166	34.2	4.6	7.3	4.6	10.4	0.0	2.0	4.7
	3. Advanced	22	112.9	6.5	3.3	3.3	0.0	2.0	0.0	
School 4	1. Remedial	14	3.5	7.9	73.5	3.5	4.1	4.1	10.4	
($N = 230$)	2. Regular	202	13.3	4.2	12.1	4.2	10.4	0.0	1.6	4.2
	3. Advanced	14	28.2	5.6	−0.7	−0.7	0.0	1.6	0.0	
School 5	1. Remedial	22	5.0	9.3	19.6	5.0	5.1	5.1	9.9	
($N = 357$)	2. Regular	309	8.5	3.9	3.2	3.9	9.9	0.0	0.8	4.0
	3. Advanced	26	12.3	4.8	2.0	2.0	5.1	0.8	0.0	
			$D^2(i \mid j) = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \, \mathbf{C}_j^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) + \ln \lvert \mathbf{C}_j \rvert$				$D^2(i \mid j) = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \, \mathbf{C}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$			

As shown in Table 4, the magnitudes of the determinants of the three within-group covariance matrices (see the column labeled ln $|\mathbf{C}_g|$) were quite different among the three groups. Moreover, the log values of the determinants of the pooled covariance matrices (see the column labeled by ln $|\mathbf{C}|$) were much closer to those of Group 2 than to those of Groups 1 and 3. In addition, a lack of similarity in the entry pattern was found by comparing the values of generalized distances between group mean vectors associated with the within–group and pooled covariance matrices (see Table 4).

The analyses just described indicated that the underlying covariance structures among the three groups were, to some extent, not identical. Since a classification equation could be distorted by invalid use of the pooled covariance matrix, we employed the within-group covariance matrices ($\mathbf{C}_g$) for the distance measure involved in the density estimation for all of the five data sets.

Computational Procedures for Group-Conditional Density Estimates

Given that the distance measure is based on the within-group covariance matrix, $\mathbf{C}_g$, the computational formula for Fisher's parametric group-conditional density estimator is of the form

$$\hat{f}_g(\mathbf{x}) = (2\pi)^{-5/2}\left|\mathbf{C}_g\right|^{-1/2}\exp\left(-0.5d_g^2\right), \tag{4}$$

where

$$d_g^2 = \left(\mathbf{x} - \bar{\mathbf{x}}_g\right)'\mathbf{C}_g^{-1}\left(\mathbf{x} - \bar{\mathbf{x}}_g\right).$$

The computational formula corresponding to the normal kernel-based density estimator of Equation 2 takes the following form:

$$\hat{f}_g(\mathbf{x}) = \frac{1}{n_g h_g^5}\sum_{i=1}^{n_g} K_N\left[(\mathbf{x} - \mathbf{X}_i)/h_g\right], \tag{5}$$

where

$$K_N\left[(\mathbf{x} - \mathbf{X}_i)/h_g\right] = (2\pi)^{-5/2}\left|\mathbf{C}_g\right|^{-1/2} \exp\left[-0.5\left(h_g^{-2}D_g^2\right)\right]$$

and

$$D_g^2 = (\mathbf{x} - \mathbf{X}_i)'\mathbf{C}_g^{-1}(\mathbf{x} - \mathbf{X}_i),\ i = 1, 2, \ldots, n_g.$$

As mentioned earlier, this study pinpointed the LSCV and LKCV methods to be associated with the kernel–based estimator of Equation 2 as the procedure for

choosing window width h. The computational score functions corresponding to the LSCV and LKCV algorithms are presented in Equations 6 and 7, respectively:

$$LSCV(h_g) = \frac{1}{n_g^2 h_g^5} \sum_i \sum_j K^* \left[(\mathbf{X}_i - \mathbf{X}_j)/h_g \right] + \frac{2}{n_g h_g^5} K(\mathbf{0}), \tag{6}$$

where

$$K^*(\mathbf{z}) = K^{(2)}(\mathbf{z}) - 2K(\mathbf{z}),$$

and

$$K(\mathbf{z}) = (2\pi)^{-5/2} \exp(-\mathbf{z}'\mathbf{z}/2) \quad K(\mathbf{0}) = (2\pi)^{-5/2},$$

and

$$K^{(2)}(\mathbf{z}) = \left(\frac{1}{2\sqrt{\pi}} \right)^5 \exp(-\mathbf{z}'\mathbf{z}/4).$$

$$LKCV(h_g) = \frac{1}{n_g} \sum_{i=1} \log \left[\frac{1}{(n_g - 1)h_g^5} \sum_{j \neq i} K\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h_g} \right) \right]. \tag{7}$$

In other words, we first employed Equations 6 and 7 to obtain h_g and then used them as the best values to calculate the group-conditional density estimates via Equation 5. Hereafter, for the sake of simplicity, let LSCV and LKCV respectively stand as the abbreviations for the kernel-based estimator of Equation 5 using the LSCV and LKCV bandwidth selectors, respectively.

The computational formula corresponding to the GKNN estimator of Equation 3 takes the form

$$\hat{f}_g(\mathbf{x}) = \frac{1}{n_g r_{k_g}(\mathbf{x})^5} \sum_{i=1}^{n_g} K_U\left[(\mathbf{x} - \mathbf{X}_i)/r_{k_g}(\mathbf{x}) \right], \tag{8}$$

where

$$K_U\left[(\mathbf{x} - \mathbf{X}_i)/r_{k_g}(\mathbf{x}) \right] = \frac{1}{|\mathbf{C}_g|^{1/2} \pi^{5/2} / \Gamma[(5/2) + 1]}$$

$$\text{if } (\mathbf{x} - \mathbf{X}_i)' \mathbf{C}_g^{-1} (\mathbf{x} - \mathbf{X}_i) \leq r_{k_g}^2(\mathbf{x}).$$

Equation 8 represents a uniform kernel-based density estimator in the calculation of group-conditional densities via a choice of the group smoothing parameter k_g.

The preceding computational formulas were used for the five school data sets, as well as for the 500 bootstrap samples of the School 5 data set. The bootstrap sample held the stratified characteristic of the School 5 data set. That is, each of

the 500 bootstrap samples was of size 357, with subgroup sizes (n_g) equal to 22, 309, and 26 for the first, second, and third groups, respectively. In other words, we conducted a simple random sampling from $\mathbf{X}_1, \ldots, \mathbf{X}_{ng}$ with replacement for each of the three subgroup samples, since $\mathbf{X}_1, \ldots, \mathbf{X}_{ng}$ are iid random variables from the associated subgroup populations (Efron, 1982).

As mentioned earlier, the five school data sets were the training data sets from which the corresponding discriminant rules were derived and classification error rates were calculated on the bases of the cross-validation procedure. The parameters estimated according to Fisher's quadratic discriminant rules were $\bar{\mathbf{x}}_g$ and $\mathbf{C}_g$ in Equation 4, whereas the parameters estimated according to nonparametric discriminant rules were h_g and $\mathbf{C}_g$ in Equation 5 and k_g and $\mathbf{C}_g$ in Equation 8. The smoothing parameters h_g were obtained via the computational score functions of Equations 6 and 7 for the respective individual school data sets broken down by the three groups. As for the three integers of k_g, they were obtained by assessing different sets of values and choosing the set that minimized the cross-validated estimate of the total error rate.

Note that the Fisher and nonparametric discriminant rules were also applied to evaluate the 500 bootstrap samples of the School 5 data set. Therefore, the parameter estimates ($\bar{\mathbf{x}}_g$, $\mathbf{C}_g$, h_g, k_g) obtained from the School 5 data set remained constant when used to analyze each of the 500 bootstrap samples.

Table 5 lists the resulting values of h_g and k_g corresponding to the respective bandwidth selection algorithms for the three subgroups within each of the five school data sets. As can be seen in Table 5, as expected, the best sets of h_g and k_g varied with the five school data sets, which differed in terms of sample size. One feature that the GKNN, LSCV, and LKCV methods have in common is that they assign a much smaller value of h or k for Group 2 than for Group 1 or 3. This is also in compliance with theoretical expectations, since the ideal value of h or k decreases as group size increases.

To summarize, the major difference in density estimation between the Fisher and kernel-based procedures resides in the carrying out of different distance measures, whereas the main difference among the LSCV, LKCV, and GKNN methods is in the choice of different values of bandwidth. Therefore, we investigated the effect of varying the distance measure and window width on the density estimates, as well as the resulting effect on the posterior probability estimates.

We used the following classification error rate measures to evaluate performance goodness among the four methods: (a) a group-specific error rate ($\hat{e}_g$); (b) an overall error rate ($\hat{e}$), which was an average of the group-specific error rate; and (c) a tau index, which was an error reduction rate relative to a prediction made by pure random assignment. The latter is expressed as tau $= \left(n_c - \sum_{g=1}^{3} n_g/3\right) \Big/ \left(N - \sum_{g=1}^{3} n_g/3\right)$, where n_c is the total number of students correctly classified and N is the sum of the three subgroup sizes. The value of tau also indicates whether an improvement is accomplished by the classification based on the discriminating variables (Klecka, 1984).

TABLE 5
Values of the Smoothing Parameters via Data-Driven Bandwidth Selection Methods for the Three Groups of Individual School Data Sets

		Data-driven bandwidth selectors: kernel method with		
		least squares cross validation	likelihood cross validation	generalized *k*th-nearest neighbor
Data set	Group	(LSCR-*h*)	(LKCR-*h*)	(GKNN-*k*)
School 1	1. Remedial	1.64	1.41	3
	2. Regular	0.58	0.74	1
	3. Advanced	0.84	1.12	4
School 2	1. Remedial	1.33	1.20	6
	2. Regular	0.56	0.68	1
	3. Advanced	1.13	1.10	8
School 3	1. Remedial	0.98	0.96	7
	2. Regular	0.63	0.61	1
	3. Advanced	1.09	1.00	17
School 4	1. Remedial	0.75	0.89	8
	2. Regular	0.24	0.60	1
	3. Advanced	0.84	0.99	10
School 5	1. Remedial	0.93	0.92	9
	2. Regular	0.32	0.55	1
	3. Advanced	0.91	0.85	9

Results

Findings for the Five School Data Sets

Table 6 lists the resulting error rate estimates based on the classification matrices via the leaving–one-out cross-validation procedure for each of the five school data sets.

As shown in Table 6, by comparing the overall error rate ($\hat{e}$), the performance of Fisher's approach is inferior to that of its nonparametric counterparts. Fisher's approach resulted in the highest overall error rate estimate for each of the five school data sets. In terms of the group error rate ($\hat{e}_g$), Fisher's approach also yielded the largest average values of $\hat{e}_1$ and $\hat{e}_3$ (remedial- and advanced-curriculum groups), regardless of the school data set. However, that approach did perform as well as the LSCV and LKCV methods on the value of $\hat{e}_2$ (regular-curriculum group); only slight discrepancies in magnitudes were found over the five data sets.

Comparisons of the performance of the three kernel-based procedures (Table 6) clearly revealed that the GKNN method produced the substantially lowest value of $\hat{e}_2$ for each school data set. For example, the five values of $\hat{e}_2$ associated with the GKNN method were all below .10, whereas those obtained with the LSCV method

TABLE 6
Measures of Error Rates and Tau Indexes Obtained From the Fisher and Kernel-Based Procedures for the Five School Data Sets

Data set	Method	$\hat{e}$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}$	$(\hat{e}_1 + \hat{e}_3)/2$	Hit rate = $1 - \hat{e}$	Tau
School 1	Fisher	1.00	.38	.78	.72	.89	.28	.32
($N = 99$)	LSCV	.50	.43	.67	.53	.58	.47	.32
	LKCV	.50	.43	.44	.46	.47	.54	.35
	GKNN	.83	.06	.67	.52	.75	.48	.76
School 2	Fisher	.56	.44	.70	.56	.63	.44	.30
($N = 120$)	LSCV	.67	.34	.50	.50	.58	.50	.44
	LKCV	.44	.34	.30	.36	.37	.64	.49
	GKNN	.33	.07	.40	.27	.37	.73	.83
School 3	Fisher	.42	.34	.32	.36	.37	.64	.49
($N = 200$)	LSCV	.17	.30	.36	.28	.27	.72	.55
	LKCV	.17	.30	.32	.26	.24	.74	.56
	GKNN	.08	.05	.32	.15	.20	.85	.88
School 4	Fisher	.43	.24	.29	.32	.36	.68	.62
($N = 230$)	LSCV	.07	.50	.14	.24	.11	.76	.32
	LKCV	.14	.31	.14	.20	.14	.80	.57
	GKNN	.00	.03	.64	.22	.32	.78	.90
School 5	Fisher	.41	.47	.42	.44	.42	.57	.31
($N = 357$)	LSCV	.23	.50	.35	.36	.29	.64	.30
	LKCV	.27	.41	.38	.36	.33	.64	.40
	GKNN	.14	.09	.62	.28	.38	.72	.80

Note. LSCV stands for kernel method with least squares cross validation, LKCV for likelihood cross validation, and GKNN for generalized *k*th-nearest neighbor. $\text{Tau} = \left(n_c - \sum_{g=1}^{3} n_g/3\right) \Big/ \left(N - \sum_{g=1}^{3} n_g/3\right)$

ranged from .30 to .50 and those produced by the LKCV method ranged from .30 to .43. Comparing the mean values of $\hat{e}_1$ and $\hat{e}_3$, however, revealed that the GKNN method was not superior to the LSCV and LKCV methods, given that the GKNN method, on average, produced higher mean values of $\hat{e}_1$ and $\hat{e}_3$ than those yielded by the other two methods.

The LSCV and LKCV methods yielded similar results, with negligible differences for the data sets involving larger sample sizes (Schools 3, 4, and 5). This was true regardless of whether the comparison was made in terms of overall or group-specific error rates. For example, the absolute differences in values of $\hat{e}$ for the five data sets were .07, .14, .02, .04, and .00, respectively, while the respective absolute

discrepancies in values of $(\hat{e}_1 + \hat{e}_3)/2$ were .11, .21, .03, .03, and .04. These figures showed that the LSCV and LKCV methods arrived at a more compatible performance for larger data sets than for smaller data sets.

Because the tau index refers to a proportional reduction in error rate relative to that expected with random assignment, we examined the values of tau obtained under the four methods. As shown in Table 6, the average magnitudes of tau over the five data sets within each of the four methods were .41, .39, .47, and .83, respectively. These figures indicated that classification based on the five indicators resulted, on average, in 41%, 39%, 47%, and 83% fewer errors than the expected accurate percentage produced by random assignment. In terms of the tau measure, the GKNN method outperformed the other three methods. This superiority, however, was associated mainly with its overwhelming accuracy in classifying second-group students.

Bootstrap Samples of the School 5 Data

To understand the effect of sampling variation on the resulting error rates, we conducted an identical set of analyses on the 500 bootstrap samples from the School 5 data set. That is, for each of the 500 samples, we obtained the associated classification results using the cross-validation procedure under the Fisher and kernel-based discriminant rules. Then we used the resulting 500 classification matrices to calculate bootstrap means, standard errors, and 95% intervals of error rates, all of which are summarized in Table 7 and Figure 1.

As shown in the top part of Table 7, Fisher's approach relative to the other three kernel-based counterparts yielded the largest bootstrap means of all error-rate mea-

TABLE 7
Bootstrap Means and Standard Errors of Error Rates Based on 500 Bootstrap Samples of the School 5 Data Set

			Kernel method with		
		Fisher	least squares cross validation (LSCV)	likelihood cross validation (LKCV)	generalized *k*th-nearest neighbor (GKNN)
Bootstrap means of error rates	$\hat{e}_1^*$	.27	.15	.19	.14
	$\hat{e}_2^*$	.41	.21	.25	.03
	$\hat{e}_3^*$	.32	.22	.18	.41
	$\hat{e}^*$	.33	.19	.21	.19
	$(\hat{e}_1^* + \hat{e}_3^*)/2$	.29	.18	.19	.27
Bootstrap standard errors of error rates	$\hat{e}_1^*$	.09	.08	.09	.07
	$\hat{e}_2^*$	.05	.02	.04	.01
	$\hat{e}_3^*$	.09	.09	.07	.11
	$\hat{e}^*$	.05	.04	.04	.04

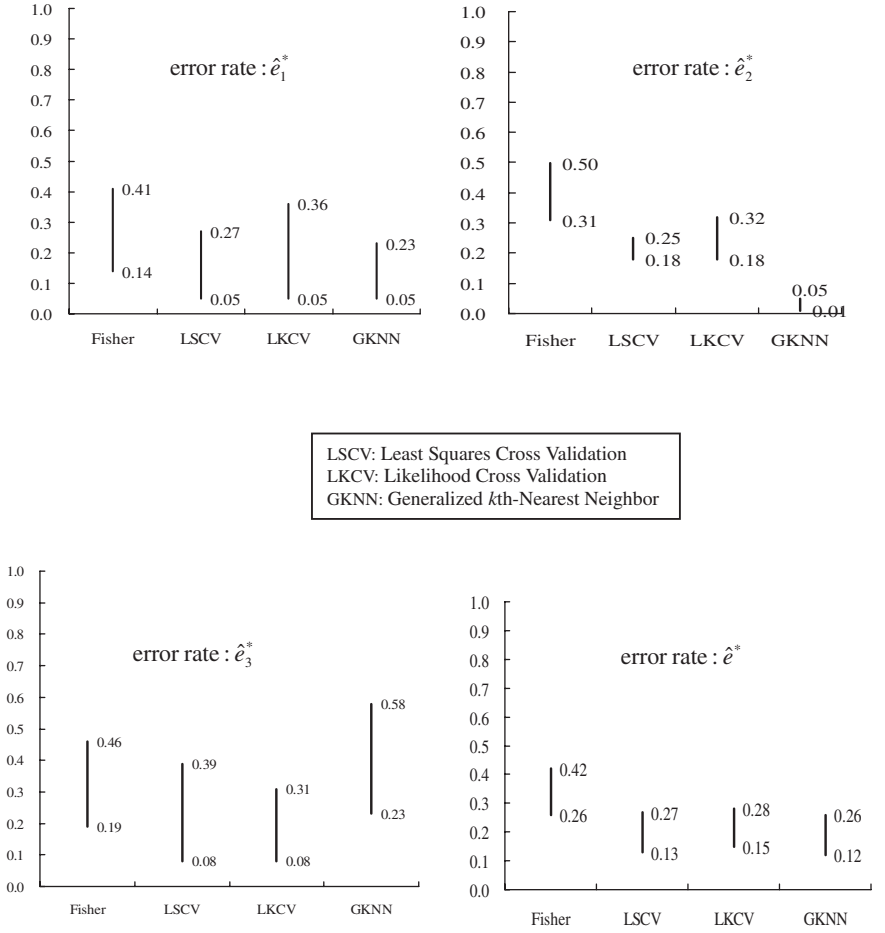

FIGURE 1. *95% confidence intervals of error rates based on 500 bootstrap samples of the School 5 data set. The depiction is repeated for each of the four error rates based on the Fisher and kernel methods with least squares cross validation (LSCV), likelihood cross validation (LKCV), and generalized* k*th-nearest neighbor (GKNN).*

sures, with one exception. Among the four methods, the GKNN method yielded the lowest bootstrap mean of $\hat{e}_2^*$ (.03) and the highest bootstrap mean of $\hat{e}_3^*$ (.41), whereas the LSCV and LKCV methods both produced bootstrap means of $\hat{e}_1^*$ and $\hat{e}_3^*$ that were much smaller than those yielded by the GKNN and Fisher methods.

The bottom part of Table 7 lists the bootstrap standard errors of error-rate measures. Comparison of the standard errors of error-rate measures showed that, except for the standard error of $\hat{e}_3^*$, Fisher's approach produced slightly higher standard errors than the three kernel-based methods. Again, there were only slim variations

in standard errors of error rates between the LSCV and LKCV methods. Also, the GKNN method produced the lowest standard error of $\hat{e}_2^*$ (.01) and the highest standard error of $\hat{e}_3^*$ (.11).

Figure 1 displays the interval estimates of the four error-rates measures. The interval estimates of $\hat{e}_g^*$ and $\hat{e}^*$ were compared by computing relative ratios using the performance of Fisher's approach as a baseline. These comparisons were made for both upper- and lower-interval estimates. Results showed that, with one exception, the relative ratios of the upper and lower limits of the three kernel-based methods were well below 1.00 with respect to Fisher's approach, regardless of whether the comparison involved group-specific or overall error rates. The exception, again, was in the comparison of interval estimates of $\hat{e}_3^*$ between the GKNN and Fisher methods. The relative ratios of the upper and lower limits were both above 1.00: 1.26 for the GKNN method and 1.21 for Fisher's method.

In terms of means and interval estimates of the overall error rates ($\hat{e}^*$), the performance differences among the three kernel-based methods can be considered negligible. Moreover, in terms of group-specific error rates ($\hat{e}_g^*$), the GKNN method performed better than the LSCV and LKCV methods, but only with respect to its accuracy in classifying second-group students. Its associated $\hat{e}_2^*$ statistics were found to be substantially lower than those yielded by the LSCV and LKCV methods. However, the reverse results were found for the comparison made on the statistics of $\hat{e}_3^*$. With respect to the performance of $\hat{e}_1^*$, only slight discrepancies in the means or interval estimates were found among the three methods. Among the three kernel-based procedures, the LSCV and LKCV methods demonstrated compatible performances on the entire set of bootstrap statistics. Both methods produced similar patterns of interval estimates as well as similar mean values on group-specific or overall error rates.

Discussion and Suggestions

This study investigated the effectiveness of Fisher's parametric discriminant rule relative to its nonparametric discriminant counterparts with respect to placing students into three instructional-curriculum coherence programs based on information carried by five science-education indicators measuring both cognitive and noncognitive attributes. The nonparametric discriminant counterparts examined in our study referred to kernel-based density estimation procedures using the LSCV, LKCV, and GKNN algorithms for bandwidth selection. We let LSCV, LKCV, and GKNN stand as abbreviations for the respective methods under kernel-based procedures. Results from the five school data sets and 500 bootstrap samples displayed the same patterns regarding relative performance goodness among the four methods. These findings can be summarized as follows.

Given that the benchmark for comparison is performance in terms of proportion of students correctly classified, Fisher's approach performed worse than its three kernel-based counterparts in that it resulted in greater overall error rates. Performance goodness among the three kernel-based procedures also varied with the different benchmarks adopted for comparison. For example, the GKNN procedure out-

performed its LSCV and LKCV counterparts with respect to its overwhelming accuracy in classifying students belonging to the regular-instructional program. The LSCV and LKCV procedures, however, both made fewer decision errors in classifying students belonging to the remedial and advanced instructional programs.

Although the GKNN method resulted in the highest overall hit rate among the four methods, this superiority was mainly due to the substantial reduction in misclassification of students in regular-instructional programs. Moreover, this superiority was an anticipated result, since the choice of k_g was set to minimize overall misclassification rate, producing an inflated hit rate. On the other hand, the major concern in the school setting is to pinpoint students who belong in remedial- or advanced-curriculum programs. We found that the advantage of the LSCV and LKCV procedures is in providing a higher hit rate associated with students in these two groups. Therefore, we regard both the LSCV and LKCV procedures as desirable competitors to Fisher's parametric discriminant rule.

The LSCV and LKCV procedures demonstrated compatible performances in terms of the entire set of evaluation measures used in this study. Moreover, both procedures arrived at a more compatible performance for larger data sets than for smaller data sets. Thus, the two methods seem equally effective in this three-group educational placement problem.

As evidenced by the higher overall hit rate and tau values accomplished by the LSCV and LKCV kernel-based procedures, Fisher's parametric discriminant rule should no longer be considered robust in dealing with a data structure that deviates from a multivariate normal distribution. The underlying principle of both the LSCV and LKCV kernel-based procedures can alleviate the need to meet such a rigid requirement. Therefore, the LSCV and LKCV kernel-based procedures are both useful and suitable for tackling the placement problem within an educational indicator system. In addition, the LSCV and LKCV procedures can both be easily implemented, since the associated h bandwidth selection is characterized by a fully automated data-driven process that does not need human assistance or intervention.

Note that the LSCV and LKCV procedures failed to accomplish a complete separation of Group 1 and Group 3 students, resulting in one or two students being placed erroneously in the opposite direction. This occurrence, however, might be the result of the use of a relatively small data set, since the kernel-based procedure is prone to produce large variations for density estimates evaluated at the tail points in highly dimensional settings. Thus, these low error counts still fall within acceptable ranges. This occurrence also indicates that further implementation or study should rely on a much larger sample to take the greatest advantage of the two procedures.

We now turn to the discriminant power contained in the five science indicators. With the three-group problem, the meaningful total hit rate must be far beyond a base rate of 33%. This study found that the hit rate given by the LSCV or LKCV method was over 33% in all five data sets, whereas Fisher's was over 33% in only four of the five data sets, and it was never as far over as the LSCV and LKCV methods. In terms of standardized measure of improvement, all three methods disclosed

the high values of the tau index. These high values give sufficient evidence that the classification based on the five science indicators achieved a meaningful improvement in prediction accuracy.

Although this study bypassed the problem of variable selection, we did find that, in terms of the classification function coefficients obtained via the use of the pooled covariance matrix, the NST measure (nature science test) appeared to be much more discriminating than the other four indicators (ATT, TTE, HHS, and FME). Because the numbers of items involved in these four science indicators are quite small, there remains much room for increasing the scale length of the four measures for the sake of building an effective educational indicator system.

This study shows the usefulness of the nonparametric kernel-based density estimation procedure with respect to its discriminant role in educational placement. Here we would like to mention as a caveat that the strength found is associated with the case wherein the underlying data structure is legitimately characterized by unequal group covariance matrices. Provided that the assumption of equal group covariance matrices holds, this study showed that in terms of total hit rate the nonparametric kernel-based density estimators, LSCV and LKCV, were unable to outperform the Fisher linear classification procedure.

Finally, we would also like to point out that the use of GPA records to establish group memberships for the training set by no means constitutes the best approach to defining students' achievement levels. A better alternative would incorporate experienced teachers' longitudinal observations of students. However, the establishment of a clear-cut group status for the training set demands a lengthy verification process as well. For this concern, we further underscore that the kernel-based procedure might be helpful in identifying numbers of clusters without reference to a training set. Future research could explore the goodness of the cluster role of the kernel-based procedure in relation to the designation of group membership.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika, 71,* 353–360.

Cheng, Y. J., Mao, S. L., Guo, H. M., Fang, T. s., Lin, J. H., & Lin, J. T. (1994). *Study of indicators of science education: Learning progress* (NSC Report No. 83-0111-S001-001). Taipei, Taiwan: National Science Council.

Cooley, W. W., & Lohnes, P. R. (1971). *Multivariate data analysis.* New York: Wiley.

Devroye, L. P., & Wanger, T. J. (1977). The strong uniform consistency of nearest neighbor density estimates. *Annals of Statistics, 5,* 536–540.

Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput., C25,* 1175–1179.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia: SIAM.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Probability and Its Applications, 14,* 153–158.

Fisher, R. A. (1936). The use of multiple measurement in taxonomic problems. *Annals of Eugenics, 7,* 179–188.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237–261.

Habbema, J. D. F., Hermans, J., & van der Broek, K. (1974). A stepwise discriminant program using density estimation. In G. Bruckman (Ed.), *Compstat 1974* (pp. 100–110). Vienna: Physica Verlag.

Hand, D. J. (1981). *Discrimination and classification.* New York: Wiley.

Hand, D. J. (1982). *Kernel discriminant analysis.* Chichester England: Research Studies Press.

Hora, S. C., & Wilcox, J. B. (1982). Estimation of error rates in several population discriminant analysis. *Journal of Marketing Research, XIX,* 57–61.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association, 91,* 401–407.

Klecka, R. W. (1984). *Discriminant analysis.* Beverly Hills, Sage.

Lachenbruch, P. A. (1975). *Discriminant analysis.* New York: Hafner.

Lachenbruch, P. A., & Mickey, M. A. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10,* 1–10.

Mack, Y. P., & Rosenblatt, M. (1979). Multivariate *K*-nearest neighbor density estimates. *Journal of Multivariate Analysis, 9,* 1–15.

Moore, D. S., & Yackel, J. W. (1977). Consistency properties of nearest neighbor density function estimators. *Annals of Statistics, 5,* 143–154.

Morrison, D. F. (1976). *Multivariate statistical methods.* New York: McGraw-Hill.

Murnane, R. J., & Raizen, S. A. (Eds.) (1988). *Improving indicators of the quality of science and math education in Grades K–12.* Washington, DC: National Academy Press.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23,* 297–321.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Park, B. U., & Turlach, B. A. (1992). Practical performance of several data driven bandwidth selectors. *Computational Statistics, 7,* 251–270.

Perlman, M. D. (1980). Unbiasedness of likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *Annals of Statistics, 8,* 247–263.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist., 27,* 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics, 9,* 65–78.

Sain, S. R., Baggerly, K. A., & Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association, 89,* 807–817.

Sam, E. (1999). *Nonparametric curve estimation.* New York: Springer.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990, May). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan,* 692–697.

Shavelson, R., McDonnell, L., Oakes, J., Carey, N., & Picus, L. (1987). *Indicator systems for monitoring math and science education.* Santa Monica, CA: RAND.

Shaycoft, M. F. (1967). *The high school years: Growth in cognitive skills.* Pittsburgh, PA: American Institutes for Research.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* New York: Chapman & Hall.

Smith, M. S. (1988 March). Educational indicators. *Phi Delta Kappan,* pp. 187–191.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics, 10,* 1040–1053.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics, 12,* 1285–1297.

Tatsuoka, M. M. (1971). *Multivariate analysis.* New York: Wiley.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F., & Gelpke, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society A, 144,* 145–174.

Whittle, P. (1958). On the smoothing of probability density functions. *Journal of the Royal Statistical Society B, 20,* 334–343.

Authors

MIAO-HSIANG LIN is a Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 11529 ROC; miao@stat.sinica.edu.tw. Her areas of specialization are educational measurement and testing, psychometrics, and applied statistics.

SU-YUN HUANG is a Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 11529 ROC; syhuang@stat.sinica.edu.tw. Her area of specialization is mathematical statistics.

YUAN-CHIN CHANG is an Associate Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 11529 ROC; ycchang@stat.sinica.edu.tw. His area of specialization is mathematical statistics.

Manuscript received March 2003
Revision received September 2003
Accepted September 2003