A Bootstrap Generalized Likelihood Ratio Test in Discriminant Analysis

by

J. Baek, H.L. Gray, W.A. Woodward, J. Miller and M. Fisk

Technical Report No. SMU/DS/TR-280

A Bootstrap Generalized Likelihood Ratio Test in Discriminant Analysis

J. Baek, H. L. Gray, W. A. Woodward, J. Miller and M. Fisk

Abstract

A generalized likelihood ratio test is developed for classification in two populations when one needs to control one of the probabilities of misclassification. The proposed classification procedure is constructed by applying the parametric bootstrap to the generalized likelihood ratio. There are known methods for controlling this misclassification probability for the case where normal distributions with the same covariance matrix are assumed. Our approach, however, can be applied to not only this case but to the case of normal distributions with different covariance matrices and the case of a mixture of discrete and continuous variables.

The results given here do not depend on normality but can, in fact, be applied to any distribution for which the maximum likelihood estimates exist. We do, however, restrict our simulation of these results to the normal distribution if the variates are all continuous. Three cases are simulated: normal distributions with equal covariance matrix, normal distributions with unequal covariance matrices, and mixture of categorical and normal variables. An application to classifying seismic events is presented.

Keywords: Bootstrap, hypothesis testing, discriminant analysis, mixtures of continuous and discrete variables, mixed variables

This research was partially supported by ARPA Contracts No. F19628-90-C-0135 and No. F29601-91-K-DB25.

1. Introduction

One of the primary problems associated with monitoring worldwide nuclear proliferation is the problem of distinguishing seismically between small earthquakes and explosions. Although the statistical problem appears to be one of discriminant analysis, it is actually one of testing hypotheses since the political and physical environment will usually require one of the errors to be preassigned.

Classical approaches for discriminant analysis in two populations depend on the ratio of the probabilities or probability density functions. The classification rule based on the ratio is optimal in the sense that it minimizes the total probability of misclassification (Welch 1939). Under the assumptions of normality, equal covariances, and unknown parameters for the variables, Anderson (1951) derived a classification rule based on the linear discriminant function, which is known as Anderson's W statistic, by substituting estimates for the parameters in the ratio. When the covariance matrices are not equal, replacing each parameter by its estimate gives the classical quadratic discriminant function (Seber, 1984, p297; Anderson, 1984, p235).

Among other classification rules is a hypothesis-testing approach which is derived by obtaining the generalized likelihood ratio. This rule based on the assumption of normal distributions with equal covariance matrices, was proposed by Anderson (1958), studied by John (1960, 1963), and has become known as John's Z statistic. Krzanowski (1982) extended this approach to mixed discrete and continuous variables. For more discriminant procedures in the mixture case, see Knoke (1982), Krzanowski (1975, 1979, 1980), and Tu and Han (1982).

Most of these classical classification rules allocate the individual to be classified to one of the populations if the ratio is less than a cut-off point c, and to the other otherwise. The cut-off point c is usually based on the probabilities of drawing an observation from the individual populations and the costs of misclassification. Associated with these procedures are the resulting misclassification probabilities. When, as in the problem of interest here, it is important to fix one of these probabilities of misclassification, the statistician will need to determine the cut-off point to allow this probability of misclassification to be prespecified.

When this probability is prespecified the problem then becomes one of testing a hypothesis. However, because of the setting of this problem we shall continue to refer to it as a classification problem. When the p-dimensional characteristic variable $\mathbf{V} \sim N_p(\boldsymbol{\mu}^{(0)}, \Sigma)$ for a population π_0, $\mathbf{V} \sim N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ for another population π_1, and $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\mu}^{(1)}$, Σ are unknown, Anderson (1973) and Kanazawa (1979) obtained the asymptotically normal expansion of the distribution of statistics W and Z respectively, which are used to find the cut-off point for a fixed value of the particular misclassification probability. In other cases (for example $\Sigma^{(0)} \neq \Sigma^{(1)}$ or V not normal) the asymptotic distribution of the classification statistics is, in general, unknown so that no hypothesis test is available.

In this paper we determine a test of the classification hypothesis that satisfies the following requirements:

i) $\Sigma^{(0)}$ is not necessarily equal to $\Sigma^{(1)}$

ii) The p-dimensional discriminant variable may be a mixture of continuous and discrete variables

iii) The continuous variables need not be normally distributed.

Examples of continuous discriminants that are commonly used in the nuclear monitoring setting are ratios of amplitudes or spectra for different time windows and frequency bands of the observed seismogram. Earthquakes typically generate more shear energy than compressional energy, while explosions usually have much more compressional energy than shear. Since compressional waves propagate faster than shear elastic waves, this leads to larger relative amplitudes in different time windows for the two source types. Although explosive devices are expected to have more intrinsic high frequency content than earthquakes, explosions are usually shallower, in more anelastic materials than the deeper earthquakes, which tends to attenuate the high frequency content. As a result, spectral ratios of particular portions of the seismograms are useful discriminants in some regions of the world.

Some examples of categorical variables that are commonly used are presence of cepstral peaks, regional seismicity (high/low), location (off-shore/on-shore), depth (deep/shallow), and, in the context of associating mine blasts with a particular mine, day of the week.

The inability to treat a mixture of discrete and continuous variables rigorously in this setting has limited the application of many statistical classification methods in the past. This has led to rule-based approaches (Sereno and Wahl, 1993) which are somewhat ad hoc, artificial intelligence approaches (Baumgardt, et al, 1992), or inappropriate applications of linear discriminant functions or chi-squared tests. It is vital, however, for monitoring applications that these issues are all addressed with statistical rigor so that the error rates involved have meaning. The classification method proposed here satisfactorily addresses this problem by applying the bootstrap to the generalized likelihood ratio. Although this method is actually a test of hypothesis, it could just as well be used as a method for classification in the classical sense with the bootstrap being used to determine the probabilities of misclassification. For additional discussion of procedures for classifying seismic events see Shumway(1988).

In Section 2, we discuss the motivation for the proposed bootstrap likelihood ratio classification procedure, show how to construct the bootstrap likelihood ratio statistic, and explain how to determine the cut-off point for a desired misclassification probability. Section 3 is devoted to the application of the procedure to three cases. In *Example 1*, the bootstrap likelihood ratio statistic is shown to perform almost as well as the statistics W and Z which are specifically designed for *Example 1*, i.e. the case where two normal distributions with the same covariance matrix are considered. The bootstrap also performs quite well for both the normal case with different covariance matrices (*Example 2*) and the case of a mixture of continuous and discrete variates (*Example 3*), where, in either case, classical classification rules cannot control the probability of misclassification since their limiting distributions are unknown. In *Example 4* we apply the results developed here to some real seismic discriminant data and in Section 4 we present some concluding remarks.

2. Bootstrap Generalized Likelihood Ratio Test for Classification

2.1. Motivation

Let $\mathbf{V}' = (V_1, \ldots, V_p)$ be a p-dimensional random vector which is used to classify an individual into either population π_0 or population π_1. For $i = 0, 1$, let $f_i(\mathbf{v} \mid \theta^{(i)})$ be the probability or probability density function of $\mathbf{V}$ evaluated at $\mathbf{v}$, if $\mathbf{v}$ comes from population π_i, where $\theta^{(i)}$ is the set of unknown parameters. The components of $\mathbf{V}$ may be all discrete, all continuous, or mixture of discrete and continuous variables. In the mixed variables case, for example, let $\mathbf{V}' = (\mathrm{Y}, \mathrm{X})$ with $\mathbf{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_k)$ and $\mathbf{X} = \mathrm{X}_1, \ldots, \mathrm{X}_{p-k}$, where $\mathrm{Y}_1, \ldots, \mathrm{Y}_k$ are discrete and $\mathrm{X}_1, \ldots, \mathrm{X}_{p-k}$ are continuous. Suppose $\mathbf{Y}$ has the probability $f_{i,\mathbf{Y}}(\mathrm{Y}|\theta^{(i)}_{\mathrm{Y}})$ and the conditional probability density function of X given Y is $f_{i,\mathbf{X}|\mathbf{Y}}(\mathrm{X}|\theta^{(i)}_{\mathbf{X}|\mathbf{Y}}, \mathrm{Y})$. Then the joint probability density function of $\mathbf{V}$ in π_i is given by

$$f_i(\mathrm{v}|\theta^{(i)}) = f_{i,\mathbf{Y}}(\mathrm{y}|\theta^{(i)}_{\mathrm{Y}}) f_{i,\mathbf{X}|\mathbf{Y}}(\mathrm{x}|\theta^{(i)}_{\mathbf{X}|\mathbf{Y}}, \mathrm{y}), \tag{1}$$

where $\theta^{(i)} = \{\theta^{(i)}_{\mathbf{Y}}, \theta^{(i)}_{\mathbf{X}|\mathbf{Y}}\}$, $i = 0, 1$. See Olkin and Tate (1961) for the mixture of the multinomial and the multivariate normal distributions.

For any given classification rule, suppose that the region R_i is such that $\mathbf{v} \in R_i$ implies that $\mathbf{v}$ is classified as belonging to π_i. Further assume that $R_0 \cap R_1 = \emptyset$. The respective probabilities of misclassification are

$$P(1|0) = \int_{R_1} f_0(\mathbf{v} \mid \theta^{(0)})\, d\mathbf{v}$$

$$P(0|1) = \int_{R_0} f_1(\mathbf{v} \mid \theta^{(1)})\, d\mathbf{v},$$

where $d\mathbf{v} = dv_1 \ldots dv_p$. The classical classification rules obtain the optimal regions R_0 and R_1 based on $f_0(\mathbf{v} \mid \theta^{(0)}) \big/ f_1(\mathbf{v} \mid \theta^{(1)})$ according to their classification principles (such as minimization of the total probability of misclassification, minimization of the total cost of misclassification, maximization of the posterior probability, minimax classification, etc.). However under any one of these classification principles, neither $P(1|0)$ nor $P(0|1)$ is fixed in advance at a certain value, which here we desire.

2.2. Bootstrapping the Log Likelihood Ratio Test Statistic

Suppose we have the training samples $\{\mathbf{v}_1^{(0)}, \mathbf{v}_2^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}\}$ of size N_0, and $\{\mathbf{v}_1^{(1)}, \mathbf{v}_2^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)}\}$ of size N_1 from π_0 and π_1, respectively. A new observation whose value is $\mathbf{v}$ must be classified as from either π_0 or π_1. Now we employ a hypothesis-testing approach to classify $\mathbf{v}$. That is, the classification of $\mathbf{v}$ is accomplished by testing the hypothesis

$$\mathrm{H}_0\!: \mathbf{v}, \mathbf{v}_1^{(0)}, \mathbf{v}_2^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)} \in \pi_0\,;\;\; \mathbf{v}_1^{(1)}, \mathbf{v}_2^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)} \in \pi_1$$

$$\mathrm{H}_1\colon\ \mathbf{v}_1^{(0)}, \mathbf{v}_2^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)} \in \pi_0\ ;\ \mathbf{v}, \mathbf{v}_1^{(1)}, \mathbf{v}_2^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)} \in \pi_1.$$

We use the generalized likelihood ratio method to construct a test. The likelihood of the two training samples is given by

$$L(\theta^{(0)}, \theta^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)}) = \prod_{j=1}^{N_0} f_0(\mathbf{v}_j^{(0)} \mid \theta^{(0)}) \prod_{j=1}^{N_1} f_1(\mathbf{v}_j^{(1)} \mid \theta^{(1)}). \quad (2)$$

Consider now the new individual $\mathbf{v}$ to be classified. If this individual is included with the training sample from π_i, then an extra multiplying factor

$$\mathrm{L}_i(\theta^{(i)} \mid \mathbf{v}) = f_i(\mathbf{v} \mid \theta^{(i)})$$

must be incorporated in (2). The generalized likelihood ratio is therefore either unity or given by

$$LR = \frac{\sup_{\{\theta^{(0)},\theta^{(1)} \mid \mathrm{H}_0\}}\{L_0(\theta^{(0)} \mid \mathbf{v})\ L(\theta^{(0)}, \theta^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})\}}{\sup_{\{\theta^{(0)},\theta^{(1)} \mid \mathrm{H}_1\}}\{L_1(\theta^{(1)} \mid \mathbf{v})\ L(\theta^{(0)}, \theta^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})\}}$$

$$= \frac{L_0(\hat\theta_0^{(0)} \mid \mathbf{v})\ L(\hat\theta_0^{(0)}, \hat\theta_0^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})}{L_1(\hat\theta_1^{(1)} \mid \mathbf{v})\ L(\hat\theta_1^{(0)}, \hat\theta_1^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})}\ , \quad (3)$$

where $\hat\theta_0^{(i)}$ is the Maximum Likelihood Estimator (MLE) of $\theta^{(i)}$ under H_0 and $\hat\theta_1^{(i)}$ is the MLE of $\theta^{(i)}$ under H_1, $i = 0, 1$. Now let $\lambda = \log(LR)$. It intuitively follows that small values of λ provide evidence against H_0 and thus the generalized likelihood ratio test is to reject H_0 if $\lambda \le \lambda_\alpha$, where λ_α is chosen to provide a size α test.

Let $P(\lambda \le \lambda_\alpha \mid \mathrm{H}_0)$ denote the size of the Type I error and $P(\lambda > \lambda_\alpha \mid \mathrm{H}_1)$ denote the size of the Type II error for a constant λ_α. Then $P(\lambda \le \lambda_\alpha \mid \mathrm{H}_0)$ is the probability of misclassification $P(1|0)$, and $P(\lambda > \lambda_\alpha \mid \mathrm{H}_1)$ is the probability of misclassification $P(0|1)$ when R_0 and R_1 are defined in terms of λ_α. Therefore we can construct a classification rule which can control one of the probabilities of misclassification by fixing the size of the test if we know the distribution of $\lambda(\mathrm{V}, \mathrm{V}_1^{(0)}, \ldots, \mathrm{V}_{N_0}^{(0)}, \mathrm{V}_1^{(1)}, \ldots, \mathrm{V}_{N_1}^{(1)})$. In most cases it is difficult to obtain the exact distribution of the test statistic λ. The distribution, however, can be approximated by employing the bootstrap method (Efron 1979, 1982).

Since the form of the probability density function is assumed known, the bootstrap samples can be obtained from the estimated density function. This is called the parametric bootstrap (Efron 1979), and we employ it in this study. We have examined the use of the nonparametric approach of resampling with replacement from the training samples, and for the training samples of size 25 or larger, this nonparametric bootstrapping yielded similar results to those reported here.

The likelihood ratio statistic for the test of the null hypothesis H_0 versus the alternative H_1 can be parametrically bootstrapped as follows. Given the training samples $\{\mathbf{v}_j^{(0)}\}_{j=1}^{N_1}$, $\{\mathbf{v}_j^{(1)}\}_{j=1}^{N_1}$, bootstrap samples $\{\mathbf{v}_j^{*(0)}\}_{j=1}^{N_0+1}$, $\{\mathbf{v}_j^{*(1)}\}_{j=1}^{N_1}$ are generated randomly from $f_0(\mathbf{v} \mid \hat{\boldsymbol{\theta}}_1{}^{(0)})$ and $f_1(\mathbf{v} \mid \hat{\boldsymbol{\theta}}_0{}^{(1)})$, respectively, where $\hat{\boldsymbol{\theta}}_1{}^{(0)}$ and $\hat{\boldsymbol{\theta}}_0{}^{(1)}$ are obtained from the original samples $\{\mathbf{v}_j^{(0)}\}_{j=1}^{\mathrm{N}_0}$ and $\{\mathbf{v}_j^{(1)}\}_{j=1}^{\mathrm{N}_1}$, respectively. The value of λ, to be denoted λ^*, is computed for the bootstrap samples by substituting $\mathbf{v}_{N_0+1}^{*(0)}$, $\{\mathbf{v}_1^{*(0)}, \ldots, \mathbf{v}_{N_0}^{*(0)}\}$, $\mathbf{v}_1^{*(1)}, \ldots, \mathbf{v}_{N_1}^{*(1)}\}$ for $\mathbf{v}$, $\{\mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}\}$, $\{\mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)}\}$ in (3), respectively. This process is repeated independently B times, and the replicated values of λ^*, $\{\lambda_j^*\}_{j=1}^{B}$, evaluated from the successive bootstrap samples, can be used to assess the true null distribution of λ. In particular, the αth empirical quantile of $\{\lambda_j^*\}_{j=1}^{B}$, denoted by λ_α^*, will essentially approach λ_α, the true critical value for the test of size α, for large N_0 and N_1 as B tends to infinity. (See Bickel and Freedman (1981) for some asymptotic theory on the quantile process for the bootstrap.). Thus we use λ_α^* as a critical value for the test of size α. Therefore, we allocate $\mathbf{v}$ to π_1 if $\lambda \leq \lambda_\alpha^*$, and allocate $\mathbf{v}$ to π_0, otherwise.

McLachlan (1987) showed the relationship between λ_α^* and the bootstrap replication size B for the specified test size α. In general, given a set of B order statistics from a population, the probability that a randomly selected member from the population is less than or equal to the jth order statistic is $j/(B+1)$. Thus, if $\alpha = j/(B+1)$, then λ^*_α is the jth smallest value of $\{\lambda^*_j\}_{j=1}^{B}$, i.e. if $\alpha = 0.05$ and $B = 299$ then λ^*_α is the 15th smallest value of $\{\lambda^*_j\}_{j=1}^{299}$.

3. Applications

The bootstrap generalized likelihood ratio test proposed here allows the p-dimensional characteristic variable $\mathbf{V}$ to be discrete, continuous, or a combination of discrete and continuous variables, and its probability or probability density function $f_i(\mathbf{V} \mid \theta^{(i)})$ for π_i is assumed to be known except for the value of the parameter $\theta^{(i)}$, $i =$ 0, 1. It can therefore be applied to the classification problem in each of these cases when one needs to control one of the probabilities of misclassification. As we will see, the bootstrap generalized likelihood ratio test essentially achieves the required probability of misclassification for even a moderate size sample. Throughout, we assume that we have random samples $\{\mathbf{v}_j^{(0)}\}_{j=1}^{N_0}$ from π_0, and $\{\mathbf{v}_j^{(1)}\}_{j=1}^{N_1}$ from π_1.

In the following four examples we consider four distinct scenarios. In the first example we consider the simple case where the observations are all normal with equal covariances. Of course this case is well established, but we consider it to demonstrate that very little is lost by using the bootstrap rather than the exact distribution. In *Example 2*, we continue to assume normality but drop the assumption of equal covariances. In this case the bootstrap is necessary in order to determine the proper critical point. However, it is not necessary to bootstrap the likelihood ratio, but instead one could bootstrap the quadratic discriminant function, Q. This example demonstrates that these two bootstrap approaches yield essentially the same result. In *Example 3* we consider a mixture of normal and binomial variates where, to our knowledge, no alternative to the method introduced here is available. Finally, in *Example 4* we consider a set of real data which is treated as a mixture of normal and multinomial data.

Example 1: Normal Distributions with Equal Covariance Matrix

Suppose that $f_i(\mathbf{v} \mid \theta^{(i)})$ is the density function for $N_p(\mu^{(i)}, \Sigma^{(i)})$ with $\Sigma^{(0)} = \Sigma^{(1)}$, $(= \Sigma)$, where $\theta^{(i)} = (\mu^{(i)}, \Sigma)$. Replacing the unknown parameters in $f_0(\mathbf{v} \mid (\mu^{(0)}, \Sigma))/f_2(\mathbf{v} \mid (\mu^{(1)}, \Sigma))$ by their estimates leads to the well-known Anderson's W statistic (A2). The likelihood ratio (A3) is characterized by John's Z statistic (A4). On the other hand, the log likelihood ratio statistic, λ, is given in (A4) and is obtained directly by taking the log of the expression (A3) and dividing it by a constant. The monotonic relationship between Z and λ is obvious. If the values of W, Z, and λ are greater than their cut-off points, then π_0 is favored for $\mathbf{v}$, and π_1 is preferred otherwise.

Now we want to choose the cut-off point so that one probability of misclassification is controlled. Let α be the desired $P(1|0)$. Anderson (1973) has obtained from the asymptotic normal distribution of W, the following approximate cut-off point W_α, which attains the desired probability α to within $O(N^{-2})$. For large N_0 and N_1,

$$W_\alpha = \frac{1}{2} D^2 + D\left[u_0 - \frac{1}{N_0}\left(\frac{p-1}{D} - \frac{1}{2} u_0\right) + \frac{1}{N}\left(\left(p - \frac{3}{4}\right) u_0 + \frac{1}{4} u_0^3\right)\right],$$

where $N = N_0 + N_1 - 2$, $D = \sqrt{(\bar{\mathbf{v}}^{(0)} - \bar{\mathbf{v}}^{(1)})' S^{-1} (\bar{\mathbf{v}}^{(0)} - \bar{\mathbf{v}}^{(1)})}$, u_0 is such that $\Phi(u_0) = \alpha$, and $\Phi(\cdot)$ is the cumulative $N(0, 1)$ density function. Kanazawa (1979) has obtained the asymptotic cut-off point Z_α for the Z statistic. For large N_0 and N_1,

$$Z_\alpha = \frac{1}{2} D^2 + D\,[u_0 + \frac{1}{2N_0 D}\left(u_0^2 + Du_0 - (p-1)\right)$$

$$- \frac{1}{2N_1 D}\left(u_0^2 + 2Du_0 + (p-1) + D^2\right) + \frac{1}{4N}\left(u_0^3 + (4p-3)u_0\right)],$$

where D and u_0 are the same as above.

Instead of deriving the limiting distribution, the cut-off point λ_α^* of the bootstrap log likelihood ratio statistic λ is obtained by the parametric bootstrap procedure described in Section 2.2. Using the MLEs of $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\mu}^{(1)}$, and Σ from the training samples $\{\mathbf{v}_j^{(i)}\}_{j=1}^{N_i}$, $i = 0, 1$, bootstrap samples $\{\mathbf{v}_j^{*(0)}\}_{j=1}^{N_0+1}\}$, $\{\mathbf{v}_j^{*(1)}\}_{j=1}^{N_1}$ are generated from a $\mathrm{N}_p(\overline{\mathbf{v}}^{(0)}, \mathrm{A}/(N_0 + N_1))$ and a $\mathrm{N}_p(\overline{\mathbf{v}}^{(1)}, \mathrm{A}/(N_0 + N_1))$, respectively. We compute the value of the log likelihood ratio statistic, λ^* corresponding to (A4), for the bootstrap samples by replacing $\mathbf{v}$, $\overline{\mathbf{v}}^{(0)}$, $\overline{\mathbf{v}}^{(1)}$, S by $\mathbf{v}_{N_0+1}^{*(0)}$, $\overline{\mathbf{v}}^{*(0)}$, $\overline{\mathbf{v}}^{*(1)}$, S^*, respectively, where $\overline{\mathbf{v}}^{*(i)} = \sum_{j=1}^{N_i} \mathbf{v}_j^{*(i)}/N_i$, i=0, 1, and S^* is calculated according to (A1) for the bootstrap samples. This process is repeated independently B times. Then λ_α^* is the αth empirical quantile of $\{\lambda_j^*\}_{j=1}^{B}$, where $\{\lambda_j^*\}_{j=1}^{B}$ are the values of λ^* evaluated from the successive bootstrap samples.

For given α, let $P_W(1|0)$, $P_Z(1|0)$ and $P_\lambda(1|0)$ be the probabilities that the new individual is misclassified into π_1 by the statistics W, Z and λ using the cut-off points W_α, Z_α, λ_α^*, respectively. Then $P_W(1|0) = P(W \le W_\alpha|\pi_0)$, $P_Z(1|0) = P(Z \le Z_\alpha|\pi_0)$, and $P_\lambda(1|0) = P(\lambda \le \lambda_\alpha^* \mid \pi_0)$. We will examine how close $P_W(1|0)$, $P_Z(1|0)$ and $P_\lambda(1|0)$ are to the desired misclassification probability, $\alpha = P(1|0)$, for the normal distributions with equal covariance matrix by Monte Carlo method. We generate two sets of random samples $\left\{\mathbf{v}_i, \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}\right\}_{i=1}^{M}$, $\left\{\{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\right\}_{i=1}^{M}$ from $\mathrm{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma)$ and $\mathrm{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma)$, respectively, where

$$\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \boldsymbol{\mu}^{(1)} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \ \text{and} \sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

For each $i = 1, 2, \ldots, M$, we obtain the values of the statistics W, Z, λ, say W_i, Z_i, λ_i, using $\{\mathbf{v}_i$, $\{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}$, $\{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}$, and compare them to their corresponding critical values $W_{i\alpha}$, $Z_{i\alpha}$, $\lambda^*_{i\alpha}$ for a fixed α. $B = 499$ bootstrap samples are used for $\lambda^*_{i\alpha}$. Then $P_W(1|0)$, $P_Z(1|0)$ and $P_\lambda(1|0)$ are estimated by the proportion of times that the value of the statistic is less than or equal to its critical value among M trials. Since $\hat{P}_W(1|0)$ is the usual estimate of a proportion, its standard deviation (s.d.) is estimated by $\sqrt{\hat{P}_W(1|0)(1-\hat{P}_W(1|0))/M}$. The standard deviation estimates of $\hat{P}_Z(1|0)$ and $\hat{P}_\lambda(1|0)$ are obtained similarly. The first portion of Table 1 shows the estimates of the probability of misclassification with their standard deviations (s.d.) for the different sample sizes with $\alpha = 0.05$, $M = 10{,}000$. The results for $\hat{P}_W(1|0)$ and $\hat{P}_Z(1|0)$ are identical when $N_0 = N_1 = 25$ since $Z = (N_0/(N_0+1))\,W$ for $N_0 = N_1$. Although for the sample sizes considered, the bootstrap estimate does not attain the same precision as the W or Z statistic's estimate, it is clearly competitive.

Table 1. The estimates of the probability of misclassification, $P(1|0) = 0.05$, and the estimates of the power, $P(1|1)$

$\hat{P}_W(1\|0)$	$\hat{P}_Z(1\|0)$	$\hat{P}_\lambda(1\|0)$
	$N_0 = N_1 = 25$	
0.054 (0.002)	0.054 (0.002)	0.061 (0.002)
	$N_0 = 30,\ N_1 = 45$	
0.055 (0.002)	0.055 (0.002)	0.060 (0.002)
$\hat{P}_W(1\|1)$	$\hat{P}_Z(1\|1)$	$\hat{P}_\lambda(1\|1)$
	$N_0 = 30,\ N_1 = 45$	
0.726 (0.004)	0.725 (0.004)	0.736 (0.004)

Now we compare the powers, $P(1|1)$, for W, Z and λ. Random samples $\{\{v_{ij}^{(0)}\}_{j=1}^{N_0}\}_{i=1}^{M}$, $\{v_i,\{v_{ij}^{(1)}\}_{j=1}^{N_1}\}_{i=1}^{M}$ are generated from $N_2(\mu^{(0)}, \Sigma)$ and $N_2(\mu^{(1)}, \Sigma)$, respectively with the same parameters as above. The power estimates for W, Z and λ, $\hat{P}_W(1|1)$, $\hat{P}_Z(1|1)$ and $\hat{P}_\lambda(1|1)$, are obtained in the same way as for $\hat{P}_W(1|0)$, $\hat{P}_Z(1|0)$ and $\hat{P}_\lambda(1|0)$, respectively. For $\alpha = 0.05$, $N_0 = 30$, $N_1 = 45$, $M = 10{,}000$ and $B = 499$, the power estimates are similar to each other with the bootstrap being slightly better (undoubtedly, due to the slightly larger critical region) as shown in the second portion of Table 1.

Example 2: Normal Distributions with Unequal Covariance Matrices

Let π_0 and π_1 be $N_p(\mu^{(0)}, \Sigma^{(0)})$ and $N_p(\mu^{(1)}, \Sigma^{(1)})$ with $\mu^{(0)} \neq \mu^{(1)}$ and $\Sigma^{(0)} \neq \Sigma^{(1)}$. When the parameters are unknown, a classical classification rule known as the quadratic discriminant function is obtained by taking the log after substituting estimates, $\overline{v}^{(0)}$, $\overline{v}^{(1)}$, $S^{(0)}$ and $S^{(1)}$ of $\mu^{(0)}$, $\mu^{(1)}$, $\Sigma^{(0)}$, and $\Sigma^{(1)}$ into the ratio of the two multivariate normal probability density functions, $f_0(v\,|\mu^{(0)}, \Sigma^{(0)})\,/f_1(v\mid \mu^{(1)}, \Sigma^{(1)})$. The quadratic discriminant function Q is given in (A5), and v is classified to π_0 if $Q > 0$ and to π_1 otherwise. The probabilities of misclassification of Q are difficult to control since even its limiting distribution is unknown.

Following the hypothesis-testing approach of (2), the MLEs of $\mu^{(0)}$, $\mu^{(1)}$, $\Sigma^{(0)}$, $\Sigma^{(1)}$ under H_0 and H_1 are given in the Appendix. The log likelihood ratio statistic, λ, is given in (A6), and to evaluate the cut-off point λ_α^*, for the desired probability of misclassification, $P(1|0) = \alpha$, we generate bootstrap samples $\{v_j^{*(0)}\}_{j=1}^{N_0+1}$, $\{v_j^{*(1)}\}_{j=1}^{N_1}$ from a $N_p(\overline{v}^{(0)}, A^{(0)}/N_1)$ and a $N_p(\overline{v}^{(1)}, A^{(1)}/N_2)$, respectively. Following the same bootstrap procedure as in *Example 1*, the αth empirical quantile λ_α^* is obtained from the values of the log likelihood ratio statistic λ for the successive bootstrap samples. The bootstrap generalized likelihood ratio classification rule with misclassification probability $P(1|0) = \alpha$ is, therefore, to assign $\mathbf{v}$ to π_1 if $\lambda(\mathbf{v}) \leq \lambda^*_\alpha$, and to π_0, otherwise.

Consider two bivariate normal distributions $N_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$, $N_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$, where

$$\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma^{(0)} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \text{ and } \Sigma^{(1)} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}.$$

Suppose we apply the Q statistic for classification using the usual classification rule, *i.e.* $\mathbf{v}$ is classified to π_0 if $Q > 0$ and to π_1 otherwise. The probability of misclassification of interest, i.e. $P_Q(1|0)$ is $P(Q \leq 0 \,|\pi_1)$. In order to determine the probability of this classification error we conduct a simulation. We generate $\{\mathbf{v}_i, \{\mathbf{v}^{(0)}_{ij}\}_{j=1}^{N_0}\}_{i=1}^{M}$, and $\{\mathbf{v}^{(1)}_{ij}\}_{j=1}^{N_1}\}_{i=1}^{M}$ from $N_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $N_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$, respectively. We obtain the Q statistics for $\{\mathbf{v}_i, \{\mathbf{v}^{(0)}_{ij}\}_{j=1}^{N_0}\}, \{\mathbf{v}^{(1)}_{ij}\}_{j=1}^{N_1}$, $i = 1, 2, \ldots, M$, and denote these $Q_1, Q_2, \ldots, Q_M$. Then $P_Q(1|0)$ is estimated by $\hat{P}_Q(1|0)$ which is the proportion of Q_i values that are less than or equal to zero. $\hat{P}_Q(1|0)$ (with its standard deviation) is 0.274 (0.004) for $N_0 = 100$, $N_1 = 150$ and $M = 10{,}000$. When it is important to keep the probability of misclassification $P_Q(1|0)$ small, an error this large may be unacceptable, resulting in the need for the method we are describing.

Now we consider the log likelihood ratio statistic λ. First, we would like to know how well the parametric bootstrap procedure approximates the true null distribution of λ. Since the true null distribution of λ is not known, we generate samples $\{\mathbf{v}_i, \{\mathbf{v}^{(0)}_{ij}\}_{j=1}^{N_0}\}_{i=1}^{M}$ from $N_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\{\{\mathbf{v}^{(1)}_{ij}\}_{j=1}^{N_1}\}_{i=1}^{M}$ from $N_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$ with $M = 100{,}000$. Applying $\{\mathbf{v}_i, \{\mathbf{v}^{(0)}_{ij}\}_{j=1}^{N_0}, \{\mathbf{v}^{(1)}_{ij}\}_{j=1}^{N_1}\}_{i=1}^{M}$ to (A6), we can obtain $\{\lambda_i\}_{i=1}^{M}$. The true null cumulative distribution function (cdf) of λ is approximated by the empirical cdf using $\{\lambda_i\}_{i=1}^{M}$ for $(N_0, N_1) = (10, 15)$, $(N_0, N_1) = (30, 45)$, and $(N_0, N_1) = (100, 150)$. The true critical value λ_α is approximated by -1.900, -1.504, -1.353 respectively. These are the αth quantiles of $\{\lambda_i\}_{i=1}^{M}$ where $\alpha = 0.05$ for $(N_0, N_1) = (10, 15)$, $(N_0, N_1) = (30,$

45), and $(N_0, N_1) = (100, 150)$, respectively. In this simulation, $B = 299$ is used for the bootstrap replication size because of computer-time constraints. Our investigation indicates that the results using $B = 299$ and $B = 499$ are similar.

For a set of random samples $\{\mathbf{v}, \{\mathbf{v}_j^{(0)}\}_{j=1}^{N_0}\}, \{\mathbf{v}_j^{(1)}\}_{j=1}^{N_1}\}$ under H_0 with $(N_0, N_1) = (10, 15)$, $(N_0, N_1) = (30, 45)$, and $(N_0, N_1) = (100, 150)$, the empirical null distribution of the bootstrap log likelihood statistic using $\{\lambda_j^*\}_{j=1}^{B}$ with $B = 299$ is also plotted around the true null cdf in Figure 1. Inspection of this figure shows that the bootstrap null distribution approximates the true null distribution of the log likelihood ratio statistic quite well as the sample sizes increase and does surprisingly well for small samples.

Even though the null distribution of the Q statistic is unknown, the cut-off point, Q_α, for misclassification probability, $P(1|0) = \alpha$, can be approximated by the same parametric bootstrap procedure as for λ. That is, we evaluate the Q statistic for B successive bootstrap samples and call them $Q_1^*, Q_2^*, \ldots, Q_B^*$. Then Q_α is approximated by Q_α^*, the αth empirical quantile of $\{Q_j^*\}_{j=1}^{B}$. Therefore, one can allocate $\mathbf{v}$ to π_1 if $Q \leq Q_\alpha^*$, and allocate $\mathbf{v}$ to π_0, otherwise.

With the same simulation data used to get $\hat{P}_Q(1|0) = 0.274$ above, $\hat{P}_{QB}(1|0)$ (s.d.), the estimate of a fixed $P(1|0) = 0.05$ by the parametrically bootstrapped Q statistic QB, is 0.050 (0.002) for $B = 499$. $\hat{P}_\lambda(1|0)$ (s.d.) of the bootstrapped λ, i.e. λ^*, is 0.049 (0.002) for the same bootstrap samples as for QB. Both bootstrap estimates are close to the true fixed misclassification probability $P(1|0) = 0.05$.

To further compare the two tests we now investigate their respective powers, $P(1|1)$, for different parameter values. Consider bivariate normal distributions $\mathrm{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\mathrm{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$. Let ρ_0 and ρ_1 be the correlation coefficient for $\mathrm{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\mathrm{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$ respectively. We assume that $\rho_0 = 0.5$, $\rho_1 = -0.5$ and that both distributions have the same marginal variances, $\sigma_1^2 = 1$ and $\sigma_2^2 = 1$. That is,

$$\sum^{(0)} = \begin{pmatrix} \sigma_1^2 & \rho_0\sigma_1\sigma_2 \\ \rho_0\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

$$\sum^{(1)} = \begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix},$$

For $\boldsymbol{\mu}^{(0)} = (0,\ 0)'$, we examine the power, $P(1|1)$ of the bootstrap Q statistic and the bootstrap λ statistic at $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(0)} + \Delta(\sigma_1,\ \sigma_2)'$, $\Delta = 1,\ 2,\ 3$, for small samples ($N_0 = 10$, $N_1 = 15$) and for large samples ($N_0 = 100$, $N_1 = 150$). For each $\Delta = 1,\ 2,\ 3$ under H_1, we randomly generate $\{\{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}\}_{i=1}^{M}$ from $\mathrm{N}_2(\boldsymbol{\mu}^{(0)},\ \Sigma^{(0)})$ and $\{\mathbf{v}_i,\ \{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}_{i=1}^{M}$ from $N_2(\boldsymbol{\mu}^{(1)},\ \Sigma^{(1)})$ with $N_0 = 10$, $N_1 = 15$ and $M = 10{,}000$. For each $i = 1, \ldots, M$ and for $\alpha = 0.05$, $\{\mathbf{v}_i,\ \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0},\ \{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}$ is used for the parametric bootstrap to obtain the cut-off points Q_α^* and λ_α^* for QB and λ, respectively. The bootstrap replication size B used here is 499. Then the power estimate $\hat{P}_{QB}(1|1)$ for QB is the proportion of times that the Q statistic value is less than or equal to Q_α^* out of M trials. The power estimate $\hat{P}_\lambda(1|1)$ for the bootstrap λ is obtained similarly. $\hat{P}_{QB}(1|1)$ and $\hat{P}_\lambda(1|1)$ are listed along with those for large samples ($N_0 = 100$, $N_1 = 150$) in Table 2.

Table 2. Power comparison between the bootstrap λ and the bootstrap Q (QB) with $B = 499$. Entry is power estimate with its standard deviation.

	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
		$N_0 = 10,\ N_1 = 15$	
$\hat{P}_\lambda(1\|1)$	0.310 (0.0046)	0.815 (0.0039)	0.992 (0.0009)
$\hat{P}_{QB}(1\|1)$	0.302 (0.0046)	0.795 (0.0040)	0.990 (0.0010)
		$N_0 = 100,\ N_1 = 150$	
$\hat{P}_\lambda(1\|1)$	0.375 (0.0048)	0.884 (0.0032)	0.999 (0.0003)
$\hat{P}_{QB}(1\|1)$	0.376 (0.0048)	0.884 (0.0032)	0.999 (0.0003)

In this simulation, the bootstrap λ has slightly higher power than the bootstrap Q for small samples, but there is little difference for large samples.

Example 3: Mixture of Categorical and Continuous Variables

In this final example we consider a mixture of continuous and discrete variates. Of the discriminant functions in the previous sections, only the λ statistic applies in this case. Suppose the variable V is a mixture of discrete and continuous variables. Let $\mathrm{V}' = (\boldsymbol{Z}, \mathrm{X})$ with $\boldsymbol{Z} = (Z_1, \ldots, Z_r)$ and $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_q)$ where $Z_1, \ldots, Z_r$ are discrete and $\mathrm{X}_1, \ldots, \mathrm{X}_q$ are continuous, r and q are positive integers. Suppose further the jth discrete variable Z_j has k_j categories, $j = 1, \ldots, r$. Then the vector of discrete variables $\boldsymbol{Z}$ may be expressed as a multinomial random variable $Y' = (Y_1, \ldots, Y_k)$, where $Y_m = 0$ or 1, $m = 1, \ldots, k$, $\sum_{m=1}^{k} Y_m = 1$, and $k = \prod_{j=1}^{r} k_j$. Thus, each distinct pattern of $\boldsymbol{Z}$ defines a multinomial cell of $\boldsymbol{Y}$ uniquely. It is assumed that the probability of obtaining an observation in cell m for π_i is $p_m^{(i)}$, ($0 \leq p_m^{(i)} \leq 1$, $\sum_{m=1}^{k} p_m^{(i)} = 1$), $i = 0, 1$. Then the joint probability density function of V in π_i is given by (1), where $\theta_{\mathrm{Y}}^{(i)\prime} = (p_1^{(i)}, \ldots, p_{k-1}^{(i)})$ and $\theta_{X|Y}^{(i)}$ is the set of parameters of X given Y.

For the population π_i, the conditional pdf of X given Y, $f_{i,\mathrm{X}|\mathrm{Y}}(\mathrm{X} \mid \mathrm{Y})$, may be of any proper type depending on the relationship between X and Y. Following Olkin and Tate (1961), for this example we assume that X has a conditional multivariate normal distribution with mean $\mu_m^{(i)}$ given Y belonging to cell m and common covariance matrix $\Sigma^{(i)}$ in all cells. If Y belongs to cell m, i.e., if $\mathrm{Y} = (\mathrm{Y}_1, \cdots, \mathrm{Y}_{m-1}, \mathrm{Y}_m, \mathrm{Y}_{m+1}, \cdots, Y_k) = (0, \cdots, 0, 1, 0, \cdots, 0)$, then $f_{i,\mathrm{Y}}(\mathrm{Y} \mid \theta_{\mathrm{Y}}^{(i)})$ and $f_{i,\mathrm{X}|\mathrm{Y}}(\mathrm{X} \mid \theta_{\mathrm{X}|\mathrm{Y}}^{(i)}, \mathrm{Y})$ of (1) are given as follows:

$$f_{i,\mathrm{Y}}(\mathrm{Y} \mid \theta_{\mathrm{Y}}^{(i)}) = p_m^{(i)}$$

$$f_{i,\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\theta^{(i)}_{\mathbf{X}|\mathbf{Y}}, \mathbf{Y}) = (2\pi)^{-q/2} \mid \Sigma^{(i)}|^{-1/2} exp\{ -(1/2)(\mathbf{x}-\mu^{(i)}_m)'(\Sigma^{(i)})^{-1}(\mathbf{x}-\mu^{(i)}_m)\}.$$

Let the jth member of the training sample, $\{\mathbf{v}^{(i)}_1, \mathbf{v}^{(i)}_2, \ldots, \mathbf{v}^{(i)}_{N_i}\}$ from π_i be denoted by $\{\mathbf{v}^{(i)\prime}_j = (\mathbf{y}^{(i)}_j, \mathbf{x}^{(i)}_j)\}$, where $\mathbf{y}^{(i)}_j$ is the vector of binary variables obtained from the discrete components $\mathbf{z}$ of $\mathbf{v}^{(i)}_j$, and $\mathbf{x}^{(i)}_j$ is the vector of continuous variables. Let $n^{(i)}_m$ denote the number of individuals of the training sample from π_i that fall in cell m defined by $\mathbf{Y}$. Then $N_i = \sum^{k}_{m=1} n^{(i)}_m$, $i = 0, 1$. The likelihood of the two training samples is given by

$$L = \prod_{i=0}^{1} [\Big\{\prod_{m=1}^{k} (p^{(i)}_m)^{n^{(i)}_m} \Big\} \Big\{(2\pi)^q \, |\Sigma^{(i)}|\Big\}^{\frac{-N_i}{2}}$$

$$\cdot \; exp\Big\{-\frac{1}{2} \sum_{j=1}^{N_i} (\mathbf{x}^{(i)}_j - \mu_{ij})' \, (\Sigma^{(i)})^{-1} \, (\mathbf{x}^{(i)}_j - \mu_{ij})\Big\}] \, , \qquad (4)$$

where μ_{ij} takes the value $\mu^{(i)}_m$ if $\mathbf{y}^{(i)}_j$ falls in the mth cell, $m = 1, \ldots, k$.

Consider now the new individual $\mathbf{v}$ to be classified, and suppose that the discrete components place it into cell l. If this individual is included with the training sample from π_i, then an extra multiplying factor

$$L^{(i)}_l = (2\pi)^{-q/2} \, |\Sigma^{(i)}|^{-1/2} p^{(i)}_l \; exp\Big\{-\tfrac{1}{2}(\mathbf{x}-\mu^{(i)}_l)' \, (\Sigma^{(i)})^{-1}(\mathbf{x}-\mu^{(i)}_l)\Big\}$$

must be incorporated in (4) to construct the generalized likelihood ratio test statistic of (3). $\mathbf{x}^{(i)}_j$ must belong to one of k subgroups corresponding to the conditional distributions depending on the value of $\mathbf{y}^{(i)}_j$ for $j = 1, \ldots, N_i$, $i = 0, 1$. Let $\mathbf{x}^{(i)}_{sm}$ be the sth member of mth subgroup of the continuous variable measurements whose discrete covariates fall in the mth cell. Then any element of $\{\mathbf{x}^{(i)}_j\}^{N_i}_{j=1}$ belongs to one of k subgroups $\{\{\mathbf{x}^{(i)}_{sm}\}^{n^{(i)}_m}_{s=1}\}^{k}_{m=1}$ where, of course, some of the $n^{(i)}_m$ could be zero. Hence we can rewrite the exponent of (4) as

$$-\frac{1}{2}\sum_{m=1}^{k} \left(\sum_{s=1}^{n_m^{(i)}}(\mathbf{x}_{sm}^{(i)} - \mu_m^{(i)})' \, (\Sigma^{(i)})^{-1}(\mathbf{x}_{sm}^{(i)} - \mu_m^{(i)})\right).$$

The MLEs under H_0 and H_1 are given in the appendix, and the log likelihood ratio statistic is given in (A7).

Krzanowski (1982) considered a similar likelihood ratio statistic when $\Sigma^{(0)} = \Sigma^{(1)}$, $(= \Sigma)$, and $\mu_m^{(i)}$ and Σ are estimated by a second-order regression model of $\mathbf{X}$ on $\mathbf{Y}$. Then he allocated a new individual to π_0 if his likelihood ratio statistic is greater than or equal to 1 and to π_1 otherwise. He did not consider the problem when it is desired to control one of the misclassification errors.

We investigate the performance of the bootstrap log likelihood ratio test by examining the power with a simulation. We consider a simple situation in which we have a discrete variable from a Bernoulli(p) distribution and an independent continuous variable distributed $N(\mu, \sigma^2)$. For $i = 0, 1$, let $\{\mathbf{v}_j^{(i)} = (z_j^{(i)}, x_j^{(i)})'\}_{j=1}^{N_i}$ be a random sample from π_i, where $z_j^{(i)} \sim$ Bernoulli(p_i) and $x_j^{(i)} \sim N(\mu_i, \sigma_i^2)$. Let $\mathbf{v} = (z, x)'$ be a new observation to be classified where $z \sim$ Bernoulli(p_1) and $x \sim N(\mu_1, \sigma_1^2)$.

We examine the power of the bootstrap λ, $\mathrm{P}_\lambda(1|1)$, for different parameter values. We set $p_0 = 0.1$, $\mu_0 = 0$, $\sigma_0 = 0.5$, and $\sigma_1 = 1$. For $p_1 = 0.9$, 0.7, and 0.5, the estimate of $P_\lambda(1|1)$ is obtained for $\mu_1 = 0.5 + \Delta\sigma_1$ where $\Delta = \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$. The power estimate, $\hat{P}_\lambda(1|1)$, is the proportion of times that the λ statistic value is less than or equal to λ_α^* out of 2000 trials, where λ_α^* is the bootstrap cut-off point at α significance level. With $N_0 = N_1 = 50$, $B = 299$, and $\alpha = 0.05$, these power estimates are plotted in Figure 2. As the separation between μ_0 and μ_1 increases, the power of the bootstrap likelihood ratio test increases. The plot also shows that the larger differences between p_0 and p_1 produces the better power curves. Simulations were also performed to verify the significance level of the test. The results were good and essentially the same as Table 1.

Example 4: A Real Data Example

Unfortunately no suitable unclassified data with categorical variables comparing nuclear explosions to earthquakes are available for this paper. However, there is a considerable amount of mining explosion data available as training data. Therefore, to illustrate the method developed here, we have applied the bootstrap generalized likelihood ratio test to observations at the ARCESS seismic array in Norway which consist of mining blasts from two separate mines (HB6 and HD9) located in the Kola Peninsula of the former Soviet Union. (For other applications of the bootstrap generalized likelihood ratio test to seismic event identification, see Fisk and Gray (1993); Fisk et al., (1993).) Fifteen blasts were observed from mine HB6 and sixteen blasts were observed from mine HD9.

The variables used here are day-of-the-week (DOW), slowness (inverse group velocity measured in seconds/degree) of Pn (SLOW), and rectilinearity of Pn (RECT). Pn is typically the first prominent portion of the seismogram to arrive for signals observed at regional distances (<2000 km). These data are part of a data set established by Sereno and Patnaik (1992) as a testbed for seismic signal identification problems. Other features are also available in this data set, but most have many missing data values, a problem we are currently addressing.

A histogram plot of DOW is plotted in Figure 3 for the two sets of mining blasts. Note that the HD9 blasts occur predominantly on day 5, while the HB6 blasts occur more uniformly throughout the week. Dot plots of the continuous variables are shown in Figure 4. SLOW exhibits relatively good separation, while there is considerable overlap for RECT.

In order to assess the value of the discrete variable we considered cases in which DOW is either included or excluded. Since the day on which an event occurred has no influence on the seismogram, we treated the continuous variables as independent of DOW. Furthermore, we assumed unequal covariance matrices since the variances for SLOW are significantly different. Setting the significance level at 0.01 and 0.05, we estimated the power using the bootstrap with and without DOW. Table 3 gives the results using both continuous variables, while Table 4 gives the results using only RECT, with and without DOW. Since SLOW is such a strong discriminator, Table 4 better demonstrates the power that may be gained by making use of an available discrete feature.

Table 3. Bootstrap estimates of power using both SLOW and RECT.

Significance	DOW excluded	DOW included
0.01	0.962	0.982
0.05	0.980	0.986

Table 4. Bootstrap estimates of power using RECT.

Significance	DOW excluded	DOW included
0.01	0.266	0.377
0.05	0.529	0.736

The power was estimated in these tables using a parametric bootstrap approach. Specifically, given the training samples of size $N_0 = 15$ and $N_1 = 16$ available from the two mines, $\pi_0 = \text{HB6}$ and $\pi_1 = \text{HD9}$, ML estimates of the associated parameters are obtained. For these data, the bootstrap is used to estimate the α-level critical value by simulating $B = 499$ replications. Each replication consists of training samples of sizes

N_0 and N_1 from the models fit to π_0 and π_1 along with an observation to be classified which is generated according to the model for π_0. As in the previous examples, the α-level critical value, λ_α^*, was obtained from the likelihood ratio statistics calculated from these replicates. The power is then estimated by again simulating B bootstrap replications, where each replicate consists of training samples of sizes N_0 and N_1 from the models fit to π_0 and π_1 along with an observation to be classified which this time is generated according to the model for π_1. The power is estimated as the proportion of the resulting B likelihood ratio statistics that are less than or equal to λ_α^*. A cross-validation procedure was also considered, and it gave results similar to those shown here. Efron (1983) has suggested an alternative bootstrap approach to remove the bias from the cross-validation estimate.

4. Concluding Remarks

When one needs to classify an individual with one of the misclassification probabilities under control but does not know the exact or limiting distribution of the statistic for classification, the bootstrap likelihood ratio method is shown to be useful. The statistic used for classification is derived from the likelihood ratio, and its limiting distribution furnishing the discriminant cut-off point is approximated successfully by the parametric bootstrap.

The bootstrap likelihood ratio statistic is shown to compete well with the statistics W and Z whose limiting distributions are known, for moderate sample sizes when two multivariate normal distributions with equal covariance matrices are considered. It also performs quite well for both the multivariate normal case with unequal covariance matrices and the case of a mixture of binary and normal variates, where classical classification rules cannot control the probability of misclassification. Moreover, the methodology considered here can be applied to any non-normal discrete or continuous variable, and to any mixture of continuous and discrete variables, whenever the MLEs exist. It should be noted that the precision of the test depends on the sample sizes N_0 and N_1, and the bootstrap replication size B. Small sample sizes may result in MLEs for the parametric bootstrap which are not close to the true parameter values. Adequate sample sizes for different dimensions of the classification variable may need to be studied. Finally, it should be noted that the method applied here could be applied to any test of hypothesis based on the generalized likelihood ratio. Actually, the approach considered here of calculating λ based on normal likelihoods and finding λ^*_α, should be a sensible approach for continuous, unimodal distributions. The robustness of this procedure is the topic of current research.

Appendix: Formulas Related to Examples

Example 1

$\mu^{(i)}$ is estimated by $\overline{\mathbf{v}}^{(i)} = \sum_{j=1}^{N_i} \mathbf{v}_j^{(i)}/N_i$ and Σ is estimated by

$$\mathrm{S} = \frac{(N_0-1)\mathrm{S}^{(0)} + (N_1-1)\mathrm{S}^{(1)}}{N_0+N_1-2}, \tag{A1}$$

where $\mathrm{S}^{(i)} = \sum_{j=1}^{N_i}(\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})\ (\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})'/(N_i-1)$, $i = 0, 1$. Anderson's W statistic is given by

$$W = \left[\mathbf{v} - \tfrac{1}{2}\left(\overline{\mathbf{v}}^{(0)} + \overline{\mathbf{v}}^{(1)}\right)\right]' \mathrm{S}^{-1}\left(\overline{\mathbf{v}}^{(0)} - \overline{\mathbf{v}}^{(1)}\right). \tag{A2}$$

Under the null hypothesis H_0, the MLEs of $\mu^{(0)}$, $\mu^{(1)}$, and Σ are

$$\hat{\mu}_0^{(0)} = (N_0\overline{\mathbf{v}}^{(0)} + \mathbf{v})/(N_0+1),$$

$$\hat{\mu}_0^{(1)} = \overline{\mathbf{v}}^{(1)},$$

$$\hat{\Sigma}_0 = \frac{1}{N_0 + N_1 + 1}\left[\mathrm{A} + \frac{N_0}{N_0 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)'\right],$$

where $\mathrm{A} = \sum_{i=0}^{1} \sum_{j=1}^{N_i} (\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})(\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})' = (N_0 + N_1 - 2)\mathrm{S}$. Under the alternative hypothesis H_1, the MLEs of the parameters are

$$\hat{\mu}_1^{(0)} = \overline{\mathbf{v}}^{(0)}$$

$$\hat{\mu}_1^{(1)} = (N_1\overline{\mathbf{v}}^{(1)} + \mathbf{v})/(N_1 + 1),$$

$$\hat{\Sigma}_0 = \frac{1}{N_0 + N_1 + 1}\left[\mathrm{A} + \frac{N_1}{N_1 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)'\right].$$

In this case the likelihood ratio given in (3), with $\hat{\theta}_0^{(0)} = (\hat{\mu}_0^{(0)}, \hat{\Sigma}_0)$, $\hat{\theta}_0^{(1)} = (\hat{\mu}_0^{(1)}, \hat{\Sigma}_0)$, $\hat{\theta}_1^{(0)} = (\hat{\mu}_1^{(0)}, \hat{\Sigma}_1)$, and $\hat{\theta}_1^{(1)} = (\hat{\mu}_1^{(1)}, \hat{\Sigma}_1)$ is, therefore,

$$\left[\frac{N + \frac{N_1}{N_1+1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})'\mathrm{S}^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})}{N + \frac{N_0}{N_0+1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})'\mathrm{S}^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})}\right]^{-(N_0+N_1+1)/2}, \tag{A3}$$

where $N = N_0 + N_1 - 2$. The likelihood ratio (A3) is characterized by John's Z statistic,

$$Z = \frac{1}{2}\left[\frac{N_1}{N_1 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)'\mathrm{S}^{-1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right) - \frac{N_0}{N_0 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)'\mathrm{S}^{-1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)\right].$$

Thus

$$\lambda = \log\left\{N + \frac{N_1}{N_1+1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})'\,\mathrm{S}^{-1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)\right\}$$

$$-\log\left\{N + \frac{N_0}{N_0+1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})'\,\mathrm{S}^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})\right\}. \tag{A4}$$

Example 2

The quadratic discriminant function is given by

$$Q = \frac{1}{2}\log\left(\frac{|\mathrm{S}^{(1)}|}{|\mathrm{S}^{(0)}|}\right) + \frac{1}{2}\left[(\mathbf{v} - \overline{\mathbf{v}}^{(1)})'(\mathrm{S}^{(1)})^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})\right.$$

$$\left. - (\mathbf{v} - \overline{\mathbf{v}}^{(0)})'(\mathrm{S}^{(0)})^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})\right]. \tag{A5}$$

The MLEs of $\mu^{(0)}$, $\mu^{(1)}$, $\Sigma^{(0)}$, $\Sigma^{(1)}$ under H_0 are

$$\hat{\mu}_0^{(0)} = (N_0\overline{\mathbf{v}}^{(0)} + \mathbf{v})/(N_0 + 1),$$

$$\hat{\mu}_0^{(1)} = \overline{\mathbf{v}}^{(1)},$$

$$\hat{\Sigma}_0^{(0)} = \frac{1}{N_0+1}\left[\mathrm{A}^{(0)} + \frac{N_0}{N_0+1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)'\right],$$

$$\hat{\Sigma}_0^{(1)} = \frac{1}{N_1}\mathrm{A}^{(1)},$$

where $\mathrm{A}^{(i)} = \sum_{j=1}^{N_i}(\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})(\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})'$, $i = 0, 1$. Under the alternative hypothesis H_1, the MLEs are

$$\hat{\mu}_1^{(0)} = \overline{\mathbf{v}}^{(0)},$$

$$\hat{\mu}_1^{(1)} = (N_1\bar{\mathbf{v}}^{(1)} + \mathbf{v})/(N_1 + 1),$$

$$\hat{\Sigma}_1^{(0)} = \frac{1}{N_0}\mathbf{A}^{(0)},$$

$$\hat{\Sigma}_1^{(1)} = \frac{1}{N_1 + 1}\left[\mathbf{A}^{(1)} + \frac{N_1}{N_1 + 1}\left(\mathbf{v} - \bar{\mathbf{v}}^{(1)}\right)\left(\mathbf{v} - \bar{\mathbf{v}}^{(1)}\right)'\right].$$

The log likelihood ratio statistic is given by

$$\lambda = \tfrac{1}{2}\log\left(\frac{|\mathbf{S}^{(1)}|}{|\mathbf{S}^{(0)}|}\right) + \tfrac{1}{2}\left[(N_1 + 1)\log\left\{(N_1 - 1) + \frac{N_1}{N_1+1}(\mathbf{v} - \bar{\mathbf{v}}^{(1)})'(\mathbf{S}^{(1)})^{-1}(\mathbf{v} - \bar{\mathbf{v}}^{(1)})\right\}\right.$$

$$\left. - (N_0 + 1)\log\left\{(N_0 - 1) + \frac{N_0}{N_0+1}(\mathbf{v} - \bar{\mathbf{v}}^{(0)})'(\mathbf{S}^{(0)})^{-1}(\mathbf{v} - \bar{\mathbf{v}}^{(0)})\right\}\right] + C(N_0, N_1) \qquad \text{(A6)}$$

where $\mathbf{S}^{(i)} = \mathbf{A}^{(i)}/(N_i - 1)$, $i = 0, 1$, and

$$C(N_0, N_1) = \log\left[\frac{(N_0-1)^{(N_0+1-p)/2}\ (N_0+1)^{(N_0+1)p/2}\ N_1^{N_1 p/2}}{(N_1-1)^{(N_1+1-p)/2}\ N_0^{N_0 p/2}\ (N_1+1)^{(N_1+1)p/2}}\right].$$

Example 3

We consider the log likelihood ratio statistic under the scenario discussed in *Example 3*, i.e. the new individual $\mathbf{v}$ to be classified has discrete components that place it into cell l. The likelihood functions on the numerator and denominator of (3) are given by

$$L_l^{(i)}L = \{(2\pi)^q\}^{-(N_0+N_1+1)/2}|\Sigma^{(0)}|^{-N_0/2}\ |\Sigma^{(1)}|^{-N_1/2}\ |\Sigma^{(i)}|^{-1/2}$$

$$\cdot\left\{\prod_{h=0}^{1}\prod_{m=1}^{k}(p_m^{(h)})^{n_m^{(h)}}\right\}\left(p_l^{(i)}\right)$$

$$\cdot \; exp\left[-\frac{1}{2}\left\{\sum_{h=0}^{1}\sum_{m=1}^{k}\left(\sum_{s=1}^{n_m^{(h)}}(\mathbf{x}_{sm}^{(h)} - \mu_m^{(h)})'(\Sigma^{(h)})^{-1}(\mathbf{x}_{sm}^{(h)} - \mu_m^{(h)})\right)\right.\right.$$

$$\left.\left. + (\mathbf{x}-\mu_l^{(i)})'(\Sigma^{(i)})^{-1}(\mathbf{x}-\mu_l^{(i)})\right\}\right], \quad i = 0, 1.$$

Under H_0 the MLEs of $p_m^{(i)}$, $\mu_m^{(i)}$, $\Sigma^{(i)}$ are

$$\hat{p}_{m0}^{(0)} = n_m^{(0)}/(N_0 + 1), \quad m = 1, \ldots, l-1, l+1, \ldots, k,$$

$$\hat{p}_{l0}^{(1)} = \left(n_l^{(0)} + 1\right)/(N_0 + 1),$$

$$\hat{\mu}_{m0}^{(0)} = \overline{\mathbf{x}}_m^{(0)}, \quad m = 1, \ldots, l-1, l+1, \ldots, k,$$

$$\hat{\mu}_{l0}^{(0)} = (n_l^{(0)}\,\overline{\mathbf{x}}_l^{(0)} + \mathbf{x})/(n_l^{(0)} + 1),$$

$$\hat{\Sigma}_1^{(0)} = \frac{1}{N_0+1}\left[\mathbf{A}^{(0)} + \frac{N_l^{(0)}}{N_l^{(0)} + 1}(\mathbf{x} - \overline{\mathbf{x}}_l^{(0)})(\mathbf{x} - \overline{\mathbf{x}}_l^{(0)})'\right].$$

$$\hat{p}_{m0}^{(1)} = n_m^{(1)}/N_1, \quad m = 1, \ldots, k,$$

$$\hat{\mu}_{m0}^{(1)} = \overline{\mathbf{x}}_m^{(1)}, \quad m = 1, \ldots, k,$$

$$\hat{\Sigma}_0^{(1)} = \frac{1}{N_1}\,\mathbf{A}^{(1)},$$

where $\overline{\mathbf{x}}_m^{(i)} = \sum_{s=1}^{n_m^{(i)}}\mathbf{x}_{sm}^{(i)}/n_m^{(i)}$, $\mathbf{A}_m^{(i)} = \sum_{s=1}^{n_m^{(i)}}(\mathbf{x}_{sm}^{(i)} - \overline{\mathbf{x}}_m^{(i)})(\mathbf{x}_{sm}^{(i)} - \overline{\mathbf{x}}_m^{(i)})'$, $m = 1, \ldots, k$, and $\mathbf{A}^{(i)} = \sum_{m=1}^{k}\mathbf{A}_m^{(i)}$. Under the alternative hypothesis H_1 the MLEs are

$$\hat{p}_{m1}^{(0)} = n_m^{(0)}/N_0, \;\; m = 1, \ldots, k,$$

$$\hat{\mu}_{m1}^{(0)} = \overline{\mathbf{x}}_m^{(0)}, \;\; m = 1, \ldots, k,$$

$$\hat{\Sigma}_1^{(0)} = \frac{1}{N_0} \; \mathrm{A}^{(0)} ,$$

$$\hat{p}_{m1}^{(1)} = n_m^{(1)}/(N_1 + 1), \;\; m = 1, \ldots, l-1, l+1, \ldots, k,$$

$$\hat{p}_{l1}^{(1)} = (n_l^{(1)} + 1)/(N_1 + 1),$$

$$\hat{\mu}_{m1}^{(1)} = \overline{\mathbf{x}}_m^{(1)}, m = 1, \ldots, l-1, l+1, \ldots, k,$$

$$\hat{\mu}_{l1}^{(1)} = (n_l^{(1)} \overline{\mathrm{x}}_l^{(1)} + \mathrm{x})/(n_l^{(1)} + 1) ,$$

$$\hat{\Sigma}_1^{(1)} = \frac{1}{N_1+1}\left[\mathrm{A}^{(1)} + \frac{N_l^{(1)}}{N_l^{(1)} + 1} (\mathbf{x} - \overline{\mathbf{x}}_l^{(1)})(\mathbf{x} - \overline{\mathbf{x}}_l^{(1)})'\right].$$

Since the exponential term of $L_l^{(i)}L$ after replacing the parameters by their MLEs, is $exp\{-(1/2)q(N_0 + N_1 + 1)\}$ for $i = 0, 1$, the log likelihood ratio statistic is given by

$$\lambda = log\left\{ \left\{\prod_{i=0}^{1} \prod_{m=1}^{k} \left(\frac{\hat{p}_{m0}^{(i)}}{\hat{p}_{m1}^{(i)}}\right)^{n_m^{(i)}}\right\} \left(\frac{\hat{p}_{l0}^{(0)}}{\hat{p}_{l1}^{(1)}}\right) \left\{ \prod_{i=0}^{1} \left(\frac{|\hat{\Sigma}_0^{(i)}|}{|\hat{\Sigma}_1^{(i)}|}\right)^{-N_i/2}\right\} \left(\frac{|\hat{\Sigma}_0^{(0)}|}{|\hat{\Sigma}_1^{(1)}|}\right)^{-1/2}\right\}. \qquad \text{(A7)}$$

References

Anderson, T. W. (1951). "Classification by Multivariate Analysis," *Psychometrika*, 16, 631-650.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis,* New York: John Wiley & Sons.

Anderson, T. W. (1973), "An Asymptotic Expansion of the Distribution of the Standardized Classification Statistic *W*," *Annals of Statistics*, 1, 964-972.

Anderson, T. W. (1984), "An Introduction to Multivariate Statistical Analysis" 2nd Ed., John Wiley & Sons, New York.

Baumgardt, D.R., J. Carney, M. Maxson and S. Carter (1992), "Evaluation of Regional Discriminants using the Intelligent Seismic Event Identification System, Semi-Annual Technical Report SAS-TR-93-38, ENSCO, Inc., Springfield, VA.

Bickel, P. J. and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap", *Annals of Statistics,* 6, 1196-1217.

Efron, B., (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics,* 7, 1-26.

Efron, B., (1982), "The Jackknife, the Bootstrap and Other Resampling Plans," *SIAM,* Philadelphia.

Efron, B., (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316-331.

Fisk, M.D. and H.L. Gray (1993), "Event Identification Analysis of the Novaya Zemlya Event on 31 December 1992 Using Outlier and Classification Likelihood Ratio Tests," MRC-R-1449, Mission Research Corp., Santa Barbara, CA.

Fisk, M.D. H.L. Gray, and G.D. McCartor (1993), "Applications of Generalized Likelihood Ratio Tests to Seismic Event Identification," PL-TR-93-2221, Phillips Laboratory, Hanscom AFB, MA.

John, S., (1960), "On Some Classification Problems, " *Sankhyá*, Series A 22, 301-308.

John, S., (1963), "On Classification by the Statistics R and Z," *Annals of the Institute of Statistical Mathematics*, 14, 237-246.

Kanazawa, M., (1979), "The Asymptotic Cut-off Point and Comparison of Error Probabilities in Covariate Discriminant Analysis," *Journal of the Japan Statistical Society,* 9, 7-17.

Knoke, J. D., (1982), "Discriminant Analysis with Discrete and Continuous Variables," *Biometrics*, 38, 191-200.

Krzanowski, W. J. ,(1975), "Discrimination and Classification using Both Binary and Continuous Variables," *Journal of the American Statistical Association*, 70, 782-790.

Krzanowski, W. J., (1979), "Some Linear Transformations for Mixtures of Binary and Continuous Variables, with Particular Reference to Linear Discriminant Analysis", *Biometrika*, 66, 33-39.

Krzanowski, W. J., (1980), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," *Biometrics*, 36, 493-499.

Krzanowski, W. J., (1982), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis-Testing Approach," *Biometrics*, 38, 991-1002.

McLachlan, G. J., (1987), "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *Applied Statistics*, 36 (3), 318-324.

Olkin, I. and Tate, R. F., (1961) "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics*, 32, 448-65.

Seber, G. A. F., (1984), *Multivariate Observations*, New York: John Wiley & Sons.

Sereno, T.J. and G.B. Patnaik, (1992), "Data to Test and Evaluate the Performance of Neural Network Architectures for Seismic Signal Discrimination," PL-TR-92-2110(I), Phillips Laboratory, Hanscom AFB, MA.

Sereno, T.J. and D. Wahl, (1993), "A Fuzzy-logic Approach to Regional Seismic Event Identification: Application to the Novaya Zemlya Event on 31 December 1992," SAIC-93/1156, Science Applications International Corp., San Diego, CA.

Shumway, R.H. (1988), *Applied Statistical Time Series Analysis*, Englewood Cliffs, New Jersey: Prentice Hall.

Tu, C. and Han, C., (1982), "Discriminant Analysis Based on Binary and Continuous Variables," *Journal of the American Statistical Association*, 77, 447-454.

Welch, B. L., (1939), "Note on Discriminant Functions," *Biometrika*, 31, 218-220.

Figure Captions

Figure 1. Plots of distribution functions. Solid line: true null distribution of λ, broken line: empirical null distribution of the bootstrap λ. (a) $N_0 = 100$, $N_1 = 150$.

Figure 2. Power curves of the bootstrap λ with mixed binary and continuous variables. $p_1 = 0.1$. Delta denotes Δ.

Figure 3. Histogram of Day of Week (DOW) for mines HD6 and HD9.

Figure 4. Dot Plots of Slowness (SLOW) and Rectilinearity (RECT) for mines HD6 and HD9.

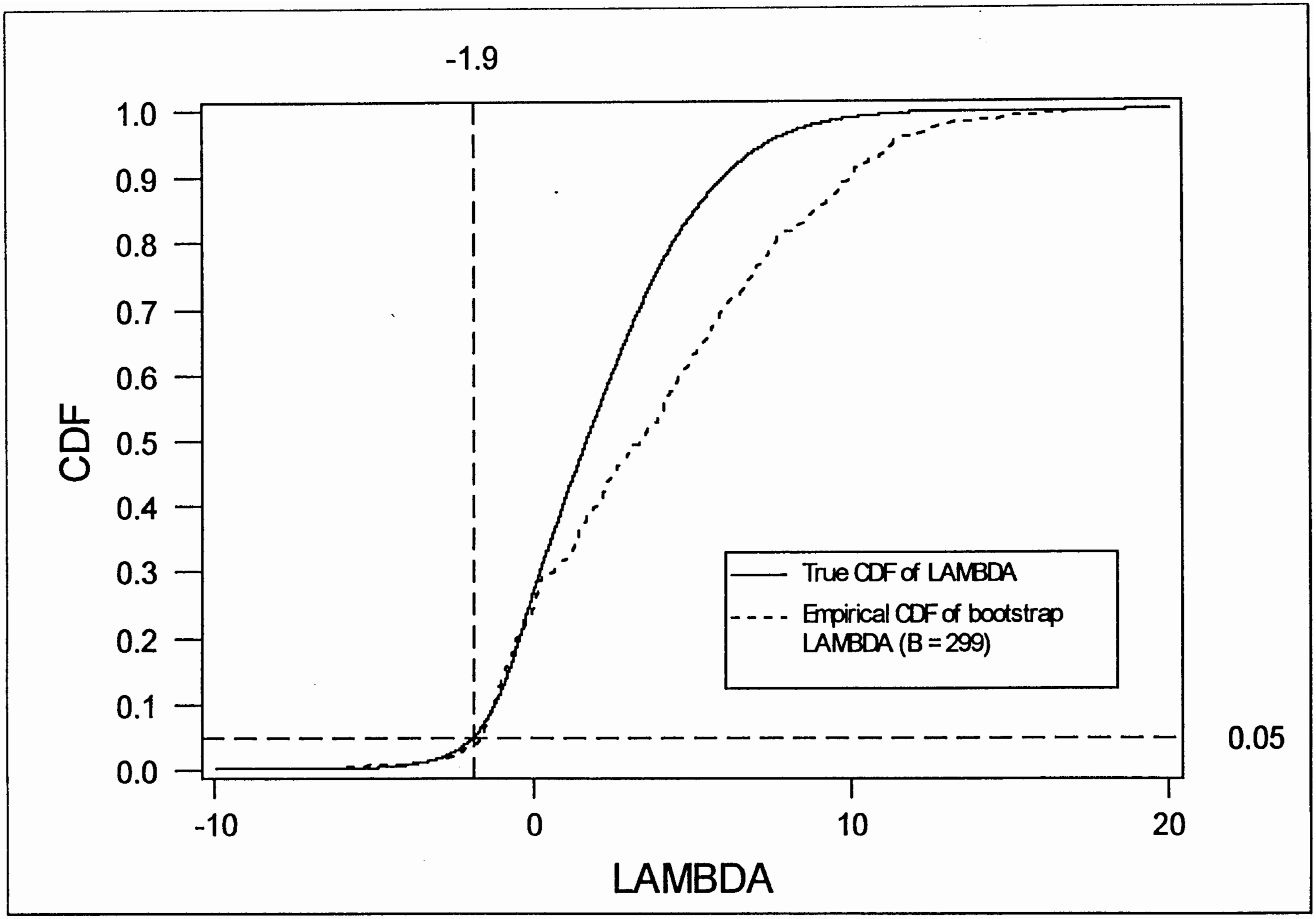

Figure 1(a). Plots of distribution functions. Solid line: true null distribution of λ. Broken line: empirical null distribution of λ. $N_0 = 10$, $N_1 = 15$.

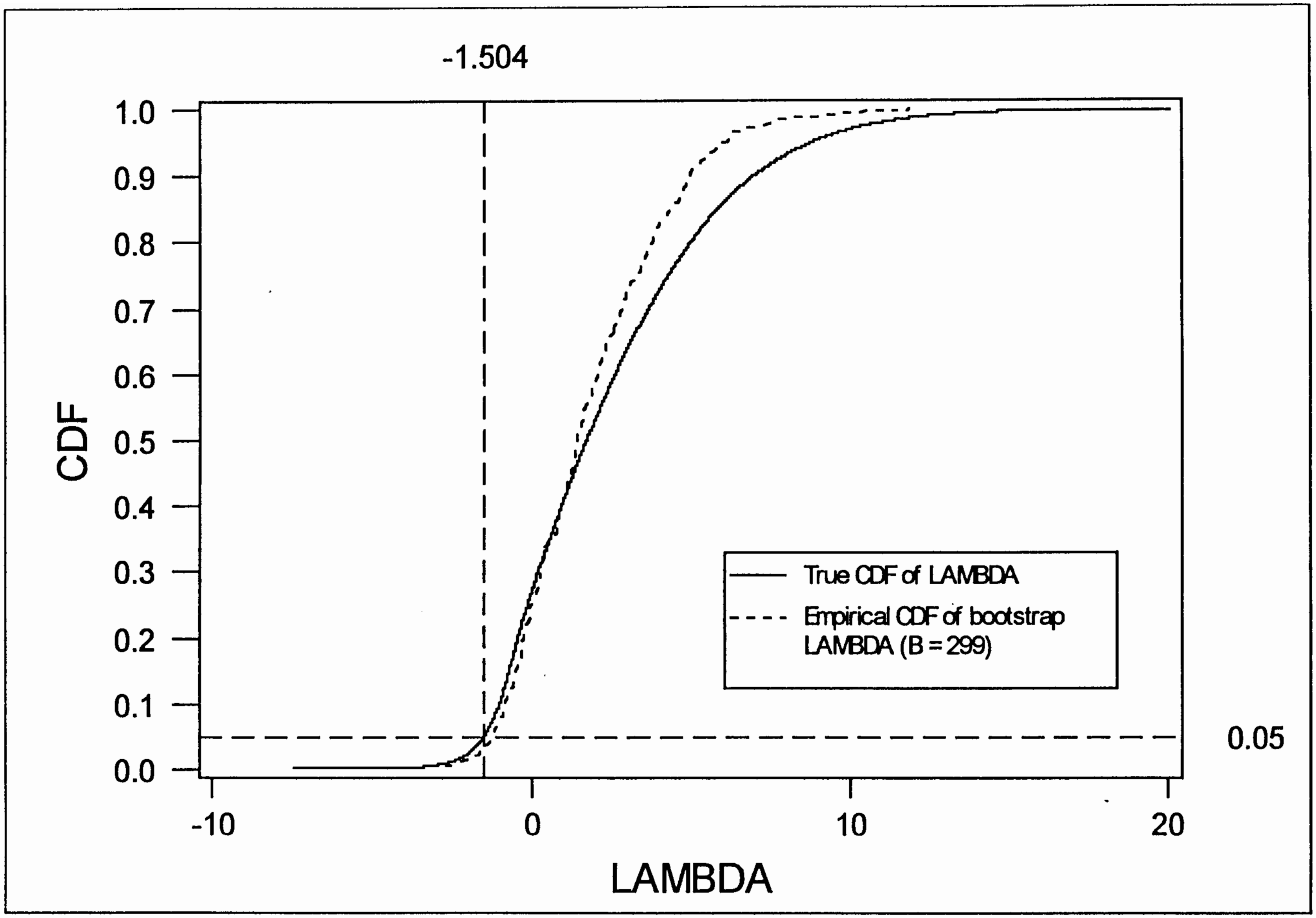

Figure 1(b). Plots of distribution functions. Solid line: true null distribution of λ. Broken line: empirical null distribution of λ. $N_0 = 30$, $N_1 = 45$.

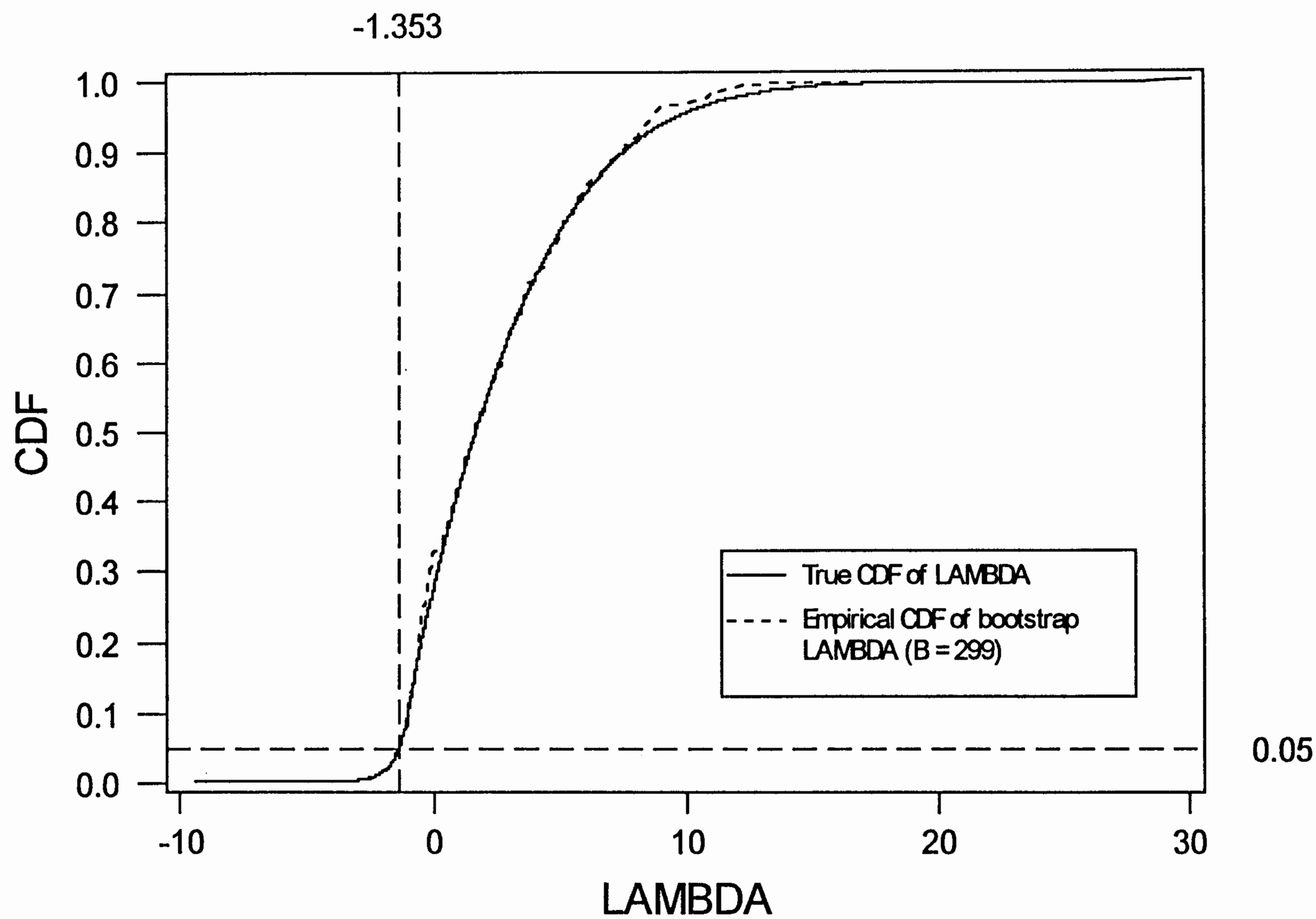

Figure 1(c). Plots of distribution functions. Solid line: true null distribution of λ. Broken line: empirical null distribution of λ. $N_0 = 100$, $N_1 = 150$.

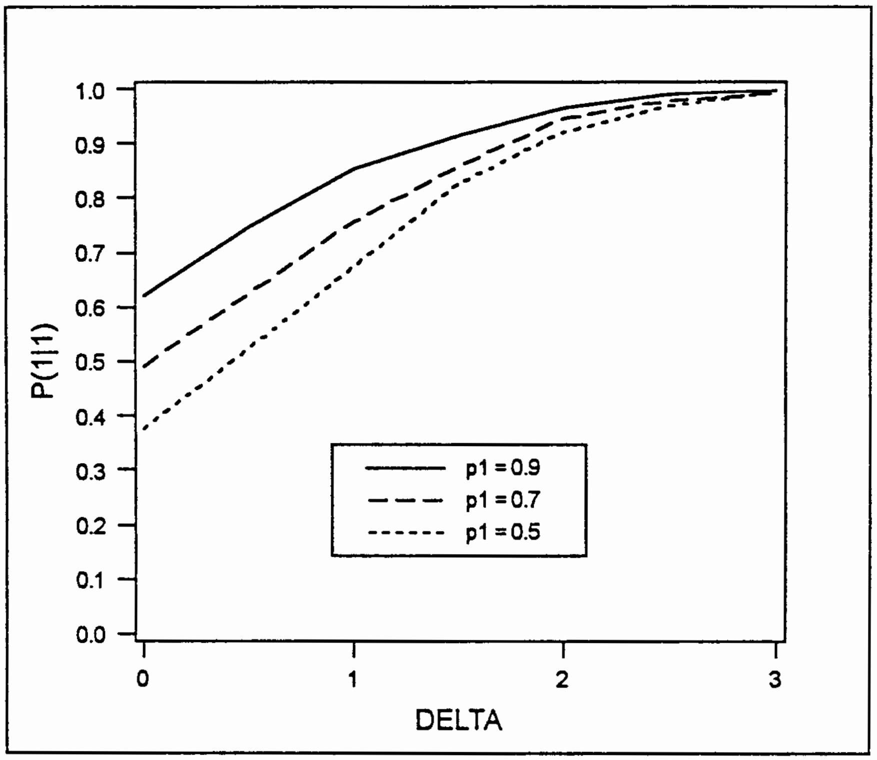

Figure 2. Power curves of bootstrap λ with mixed binary and continuous variables. $P_0 = 0.1$. DELTA denotes Δ.

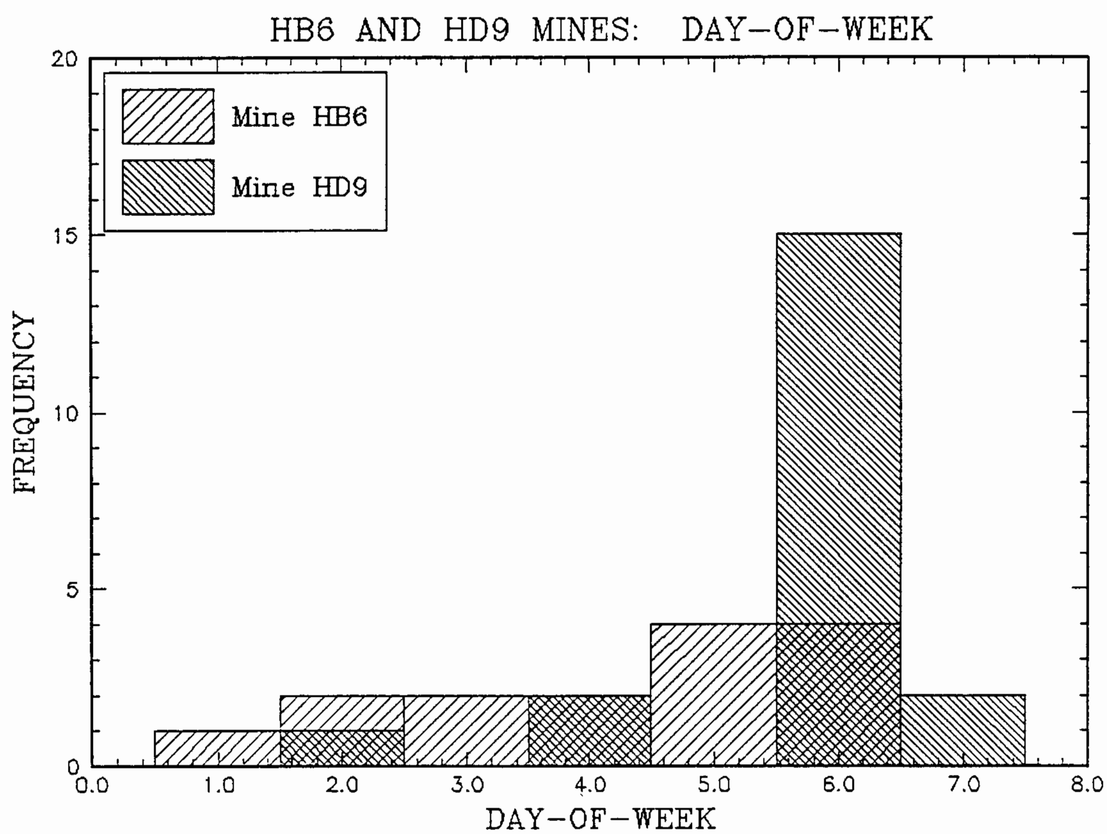

Figure 3. Histogram plot of the categorical variable day-of-week.

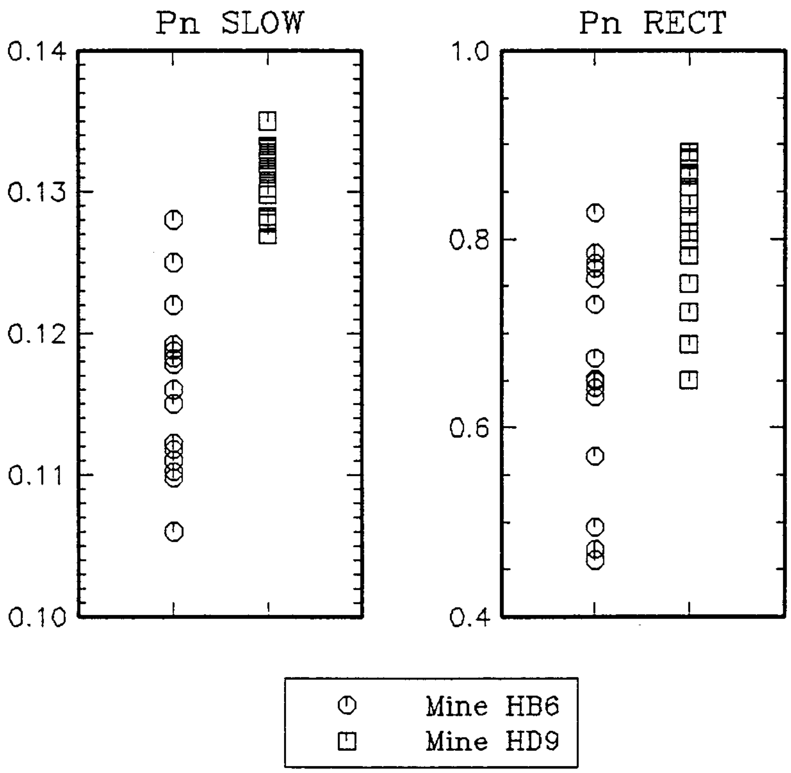

Figure 4. Dot plots of continuous variables used to classify mining blasts.