응용통계연구 제7권 2호
1994년 9월 pp. 225-237

A Generalized Likelihood Ratio Test in Outlier Detection

Jangsun Baek[1)]

Abstract

A generalized likelihood ratio test is developed to detect an outlier associated with monitoring nuclear proliferation. While the classical outlier detection methods consider continuous variables only, our approach allows both continuous and discrete variables or a mixture of continuous and discrete variables to be used. In addition, our method is free of the normality assumption, which is the key assumption in most of the classical methods. The proposed test is constructed by applying the bootstrap to a generalized likelihood ratio. We investigate the performance of the test by studying the power with simulations.

1. Introduction

Among the statistical issues associated with monitoring nuclear proliferation is the subject of outlier detection. As nuclear events are likely to occur in regions where no previous activity has occurred, one of the primary concerns is to detect the occurrence of unusual events.

Suppose the variables $\mathbf{V}=(V_1, V_2, \cdots, V_k)$ are used to characterize the occurrence of an event. The V_i will be referred to as "features." Various seismic measures on events, which include mine explosions and nuclear explosions, are used as features. Suppose further that a training sample $\{\mathbf{V}_i\}_{i=1}^{n}$ is available from past mine explosions and a new observation, $\mathbf{V}_{n+1}$, is obtained which must be classified as to whether or not it belongs to the same population as the training sample. If $\mathbf{V}_{n+1}$ is the feature measures on a nuclear explosion, it must be detected as an unusual event compared with the training sample. This problem is referred to as outlier detection.

There is vast literature on the outlier detection problem (Barnett and Lewis, 1978). Campbell (1980), Hampel et al. (1986), Rousseeuw (1985), Rousseeuw and van Zomeren (1990) and Hadi (1992) used robust estimators of the location and covariance matrix for the detection of outliers in multivariate data. Bacon-Shone and Fung (1987) developed a graphical method. Even though much research has been carried out on the continuous variable case, there is very little previous work on the mixed variables (discrete and

1) Department of Statistics, Chonnam National University, Kwangju, 500-757, KOREA.

continuous) case.

In Section 2 of the present paper, we develop a generalized likelihood ratio test to detect an outlier with discrete and continuous variables. The proposed test is constructed by applying the bootstrap (Efron, 1979, 1982) to a generalized likelihood ratio. Section 3 contains a simulation result to investigate the performance of the test by examining the power. In Section 4, we make some concluding remarks and discussion.

2. A bootstrap generalized likelihood ratio test

Suppose for a given event, the variables $\mathbf{V}=(\mathbf{Z}', \mathbf{X}')$, with $\mathbf{Z}=(Z_1, \cdots, Z_r)'$ and $\mathbf{X}=(X_1, \cdots, X_p)'$, are used to characterize the occurrence of the event, where $Z_1, \cdots, Z_r$ are discrete and $X_1, \cdots, X_p$ are continuous. The Z_i and X_i are "features." Suppose further that a training sample $\{\mathbf{V}_i=(\mathbf{Z}_i', \mathbf{X}_i')\}_{i=1}^{n}$ is available from the past events and that a new observation, $\mathbf{V}_{n+1}$, must be classified as to whether or not it belongs to the same population as the training sample.

More specifically, let the jth discrete variable Z_j have k_j categories, $j=1, \cdots, r$. Then the vector of discrete variables $\mathbf{Z}$ may be expressed as a multinomial random variable $\mathbf{Y}=(Y_1, \cdots, Y_k)'$, where $Y_m=0$ or 1, $m=1, \cdots, k$, $\sum_{m=1}^{k} Y_m=1$, and $k=\prod_{j=1}^{r} k_j$. Thus, each distinct pattern of $\mathbf{Z}$ defines a multinomial cell uniquely. Then, following Olkin and Tate (1961), for now, in order to be specific, it is assumed that $\mathbf{X}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}_m$ given $\mathbf{Z}$ corresponding to cell m of $\mathbf{Y}$ (*i.e.*, when $Y_m=1$ $(m=1, \cdots, k)$) and common covariance matrix $\boldsymbol{\Sigma}$ in all cells. Furthermore, it is assumed that the probability of obtaining an observation in cell m is p_m $(0 \le p_m \le 1, \sum_{m=1}^{k} p_m=1)$. Hence we consider the training sample $\{\mathbf{V}_i\}_{i=1}^{n}$ to be from the joint probability density function $f(\cdot\,; \mathbf{p}_1, \mathbf{U}_1, \boldsymbol{\Sigma})$, where

$$f(\mathbf{V}; \mathbf{p}_1, \mathbf{U}_1, \boldsymbol{\Sigma})=f_1(\mathbf{Y}; \mathbf{p}_1) f_2(\mathbf{X}; \mathbf{U}_1, \boldsymbol{\Sigma} | \mathbf{Y})$$

with

$$f_1(\mathbf{Y};\mathbf{p}_1)=\prod_{m=1}^{k}p_{1m}^{Ym}, \tag{1}$$

$$\begin{aligned} f_2(\mathbf{X};\mathbf{U}_1,\Sigma|Y_m=1,Y_j=0,j=1,\cdots,m-1,m+1,\cdots,k) \\ =(2\pi)^{-P/2}|\Sigma|^{-1/2}\exp\{-(1/2)(\mathbf{X}-\mu_{1m})'\Sigma^{-1}(\mathbf{X}-\mu_{1m})\}, \end{aligned} \tag{2}$$

$\mathbf{p}_1=(p_{11},\cdots,p_{1,k-1})'$, $p_{1k}=1-\sum_{j=1}^{k-1}p_{1j}$, $0\le p_{1m}\le 1$, $m=1,\cdots,k$, and $\mathbf{U}_1=(\mu_{11},\cdots,\mu_{1k})$. Similarly let the new individual $\mathbf{V}_{n+1}$ have the joint probability density $f(\cdot;\mathbf{p}_2,\mathbf{U}_2,\Sigma)$, where

$$f(\mathbf{V};\mathbf{p}_2,\mathbf{U}_2,\Sigma)=f_1(\mathbf{Y};\mathbf{p}_2)f_2(\mathbf{X};\mathbf{U}_2,\Sigma|\mathbf{Y}),$$

f_1 and f_2 are similarly defined as in (1) and (2) with $\mathbf{p}_2=(p_{21},\cdots,p_{2,k-1})'$ and $\mathbf{U}_2=(\mu_{21},\cdots,\mu_{2k})$ in place of $\mathbf{p}_1,\mathbf{U}_1$, respectively.

Now we employ a hypothesis-testing approach to classify $\mathbf{V}_{n+1}$. That is, the classification of $\mathbf{V}_{n+1}$ is accomplished by testing the hypothesis $H_0:\mathbf{p}_1=\mathbf{p}_2,\ \mathbf{U}_1=\mathbf{U}_2$ versus $H_1:\mathbf{p}_1\neq\mathbf{p}_2$ or $\mathbf{U}_1\neq\mathbf{U}_2$. We use the generalized likelihood ratio method to construct a test. Let $\Omega_0=\{\theta=(\mathbf{p}_1,\mathbf{p}_2,\mathbf{U}_1,\mathbf{U}_2,\Sigma)|p_{1j}=p_{2j}\in[0,1],\ j=1,\cdots,k-1,\ \mu_{1m}=\mu_{2m}\in R^p,\ m=1,\cdots,k,\ \Sigma$: positive definite $\}$, and let $\Omega=\{\theta=(\mathbf{p}_1,\mathbf{p}_2,\mathbf{U}_1,\mathbf{U}_2,\Sigma)|p_{ij}\in[0,1],\ \mu_{im}\in R^p,\ i=1,2,\ j=1,\cdots,k-1,\ m=1,\cdots,k,\ \Sigma$: positive definite $\}$. Furthermore, let n_m denote the number of members of the training sample whose discrete variables fall in cell m. Then, the likelihood of the training sample $\{\mathbf{V}_i\}_{i=1}^{n}$ is given by

$$\begin{aligned} &L(\theta;\mathbf{V}_1,\cdots,\mathbf{V}_n) \\ &=\{(2\pi)^p|\Sigma|\}^{-n/2}\left(\prod_{m=1}^{k}p_{1m}^{n_m}\right)\exp\{-(1/2)\sum_{i=1}^{n}(\mathbf{X}_i-\mu_i)'\Sigma^{-1}(\mathbf{X}_i-\mu_i)\}, \end{aligned}$$

where μ_i takes the value μ_{1m} if $\mathbf{Y}_i$ falls in the mth cell, $m=1,\cdots,k$. Now suppose that the discrete components of $\mathbf{V}_{n+1}$ place it into cell m. Then

$$\begin{aligned} L(\theta;\mathbf{V}_1,\cdots,\mathbf{V}_{n+1})= L(\theta;\mathbf{V}_1,\cdots,\mathbf{V}_n)(2\pi)^{-p/2}p_{2m}|\Sigma|^{-1/2} \\ \cdot\exp\{-(1/2)(\mathbf{X}_{n+1}-\mu_{2m})'\Sigma^{-1}(\mathbf{X}_{n+1}-\mu_{2m})\}. \end{aligned}$$

The generalized likelihood ratio is therefore defined by

$$\lambda_1=\frac{\sup_{(\theta\in\Omega_0)}L(\theta;\mathbf{V}_1,\cdots,\mathbf{V}_{n+1})}{\sup_{(\theta\in\Omega)}L(\theta;\mathbf{V}_1,\cdots,\mathbf{V}_{n+1})} \tag{3}$$

$$=\frac{L(\hat{\theta}_0;\mathbf{V}_1,\cdots,\mathbf{V}_{n+1})}{L(\hat{\theta};\mathbf{V}_1,\cdots,\mathbf{V}_{n+1})}, \tag{4}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}_0$ is the MLE of $\boldsymbol{\theta}$ under the restriction that H_0 is true. It intuitively follows that small values of λ_1 provide evidence against H_0, and thus the generalized likelihood ratio test is to reject H_0 if $\lambda_1 \leq \lambda_1(\alpha)$, where $\lambda_1(\alpha)$ is chosen to provide a size α test.

For the case in which all feature variables are continuous ($\mathbf{V} = \mathbf{X}'$) , and in fact normal, Caroni and Prescott (1992) showed that the hypothesis-testing approach with the likelihood ratio statistic can be used successfully for outlier detection. That is, the likelihood ratio statistic for testing $H_0^* : \mathbf{X}_i \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $i=1, \cdots, n+1$, against $H_1^* : \mathbf{X}_i \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $i=1, \cdots, n$, and $\mathbf{X}_{n+1} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, leads to Wilks's statistic (Wilks, 1963), $W = |\mathbf{A}_n| / |\mathbf{A}_{n+1}|$, where $\mathbf{A}_n = \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)'$ and $\mathbf{A}_{n+1} = \sum_{i=1}^{n+1} (\mathbf{X}_i - \overline{\mathbf{X}}_{n+1})(\mathbf{X}_i - \overline{\mathbf{X}}_{n+1})'$, which is commonly used for multivariate outlier detection. It is also easily shown that W is essentially Hotelling's T^2 statistic (Anderson 1984), where $T^2 = n/(n+1)(\mathbf{X}_{n+1} - \overline{\mathbf{X}}_n)' \mathbf{S}_n^{-1} (\mathbf{X}_{n+1} - \overline{\mathbf{X}}_n)$ with $\mathbf{S}_n = \mathbf{A}_n/(n-1)$. This test statistic is generally used for testing the equality of two normal population means when the covariance matrices are assumed equal but unknown. Thus, in the continuous case, the outlier detection problem, when based on location, is a special case of testing the equality of two population means.

Note that λ_1 can be determined in the same manner regardless of the assumed pdf's for $\mathbf{Z}$ and $\mathbf{X}$. In any case the distribution of the test statistic λ_1 under H_0, will most likely be intractable due to the nature of the observations whose elements are discrete and continuous variables. For example, since the regularity conditions do not hold for $-2\log \lambda_1$ to have its usual asymptotic chi-square null distribution, it is not easy to determine the critical value $\lambda_1(\alpha)$. This difficulty is, however, overcome using the bootstrap method (Efron, 1979, 1982).

Given the training sample $\{\mathbf{V}_i\}_{i=1}^{n}$ and $\mathbf{V}_{n+1}$, the bootstrap method proceeds in three steps: (i) Draw B bootstrap samples. Each bootstrap sample $\{\mathbf{V}_{ij}^*\}_{i=1}^{n+1}$ is a random sample of size $n+1$ drawn with replacement from the actual sample $\{\mathbf{V}_i\}_{i=1}^{n+1}$; (ii) for each bootstrap sample $\{\mathbf{V}_{ij}^*\}_{i=1}^{n+1}$, evaluate the statistic λ_1, say $\lambda_{1j}^* = \lambda_1(\mathbf{V}_{1j}^*, \mathbf{V}_{2j}^*, \cdots, \mathbf{V}_{n+1,j}^*)$, $j=1, \cdots, B$; and (iii) calculate the αth empirical quantile of $\{\lambda_{1j}^*\}_{j=1}^{B}$, $\lambda_1^*(\alpha)$. As $B \rightarrow \infty$, $\lambda_1^*(\alpha)$ will essentially, for large n, approach $\lambda_1(\alpha)$, the true critical value for the test of size α. For most situations B in the range 50 to 200 is quite adequate (Efron and Tibshirani, 1986).

Since the form of the probability density function is assumed known, the bootstrap samples can be obtained from the estimated density function. That is, the bootstrap samples at stage (i) could be drawn from $f(\mathbf{V}; \hat{\boldsymbol{\theta}}_0)$, not from the original sample $\{\mathbf{V}_i\}_{i=1}^{n+1}$, where $\hat{\boldsymbol{\theta}}_0$ is the MLE of $\boldsymbol{\theta}$ under H_0. This is called the parametric bootstrap, and we employ it in this research.

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 sample $\{\mathbf{V}_i\}_{i=1}^{n}$ and $\mathbf{V}_{n+1}$, a bootstrap sample $\{\mathbf{V}_i^*\}_{i=1}^{n+1}$ is generated randomly from $f(\mathbf{V}; \hat{\boldsymbol{\theta}}_0)$, where $\hat{\boldsymbol{\theta}}_0$ is the MLE of $\boldsymbol{\theta}$, under the null hypothesis, from the original sample $\{\mathbf{V}_i\}_{i=1}^{n+1}$. The value of λ_1, to be denoted λ_1^*, is computed for the bootstrap sample using (4). This process is repeated independently B times, and the replicated values of λ_1^*, $\{\lambda_{1j}^*\}_{j=1}^{B}$, evaluated from the successive bootstrap samples, can be used to assess the true null distribution of λ_1. In particular, the αth empirical quantile of $\{\lambda_{1j}^*\}_{j=1}^{B}$, denoted by $\lambda_1^*(\alpha)$, approximates the true α th quantile $\lambda_1(\alpha)$. That is, $\lambda_1(\alpha) \approx \lambda_1^*(\alpha)$. Thus we use $\lambda_1^*(\alpha)$ as a critical value for the test of size α. Therefore, we reject H_0 if $\lambda_1 \leq \lambda_1^*(\alpha)$, where $\lambda_1^*(\alpha)$ is obtained as discussed above.

McLachlan (1987) showed the relationship between $\lambda_1^*(\alpha)$ and the bootstrap replication size B for the specified test size α. If the bootstrap and true null distribution of λ_1 were the same, the original and subsequent bootstrap values of λ_1 can be treated as the realizations of a random sample of size $B+1$, and the probability that a specified member is smaller than or equal to the jth smallest member of the others is $j/(B+1)$. That is, $\alpha = j/(B+1)$. For example, for $\alpha = 0.05$, we need $B = 199$ with $j = 10$. Therefore $\lambda_1^*(\alpha)$ is the 10th smallest value of $\{\lambda_{1j}^*\}_{j=1}^{200}$ for $B = 199$ and $\alpha = 0.05$.

3. Simulations

We investigate the performance of the test by examining the power with simulations. We consider a simple situation in which we have a discrete variable from a Bernoulli(p) distribution, and an independent continuous variable distributed $N(\mu,\sigma^2)$. Let $\{\mathbf{V}_i = (Y_i, X_i)\}_{i=1}^{n}$ be a training sample, where $Y_i \sim$ Bernoulli(p_1) and $X_i \sim N(\mu_1,\sigma^2)$, $i=1,\cdots,n$. Let $\mathbf{V}_{n+1} = (Y_{n+1}, X_{n+1})$ be a new observation where $Y_{n+1} \sim$ Bernoulli(p_2) and

$X_{n+1} \sim N(\mu_2, \sigma^2)$. Then the goal is to test $H_0 : p_1 = p_2,\ \mu_1 = \mu_2$ versus $H_1 : p_1 \neq p_2$ or $\mu_1 \neq \mu_2$.

We examine how well the bootstrap estimate of the critical value, $\lambda_1^*(\alpha)$ approximates the true critical value, $\lambda_1(\alpha)$, and the power of the method by comparing the power using the test statistic λ_1 with the power associated with the bootstrap likelihood ratio test based on $\lambda_1^*(\alpha)$. We consider $p_1 = p_2$ and $\mu_2 = \mu_1 + \Delta\sigma$, where $\Delta \in \{0,1,2,3\}$. In the case $\Delta = 0$, the power is the significance level. We set $n = 100$, $p_1 = p_2 = 0.5$, $\mu_1 = 0$ and $\sigma = 1$ in the simulation. For the likelihood ratio λ_1, the power of the test is defined by $P(\lambda_1 \leq \lambda_1(\alpha) \mid \Delta)$. Though the true critical value $\lambda_1(\alpha)$ is unknown, since we know the parameters we can get a very close estimate of $\lambda_1(\alpha)$ using a Monte Carlo procedure and estimate the power. The procedure is described as follows:

For given n, Δ and very large positive integers K, M,

For i = 1 to K Do:

Generate a random sample $\{\mathbf{V}_{i1}, \cdots, \mathbf{V}_{i,n+1}\}$ under H_0.

Calculate λ_{1i} using (4) with $\{\mathbf{V}_{ij}\}_{j=1}^{n+1}$.

End For Loop.

$\lambda_1(\alpha) \approx \alpha$th quantile of $\{\lambda_{1i}\}_{i=1}^{K}$.

For i = 1 to M Do:

Generate a random sample $\{\mathbf{V}_{i1}, \cdots, \mathbf{V}_{in}\}$ under H_0.

Generate $\mathbf{V}_{i,n+1}$ randomly under (H_1, Δ).

Calculate λ_{1i} using (4) with $\{\mathbf{V}_{ij}\}_{j=1}^{n+1}$.

End For Loop.

$P(\lambda_1 \leq \lambda_1(\alpha) \mid \Delta) \approx \sum_{i=1}^{M} I(\lambda_{1i} \leq \lambda_1(\alpha))/M$, where $I(\cdot)$ is the indicator function.

Now the power of the bootstrap likelihood ratio test is estimated as follows:

For given n, Δ and very large positive integers B, M,

For i = 1 to M Do:

Generate a random sample $\{\mathbf{V}_{i1}, \cdots, \mathbf{V}_{in}\}$ under H_0.

Generate $\mathbf{V}_{i,n+1}$ randomly under (H_1, Δ).

Calculate $\hat{\boldsymbol{\theta}}_{0i}$ and $\hat{\boldsymbol{\theta}}_i$ in (4) with $\{\mathbf{V}_{ij}\}_{j=1}^{n+1}$.

Calculate λ_{1i} using (4).

For j = 1 to B Do:

Generate bootstrap sample $\{\mathbf{V}_{ij1}^*, \cdots, \mathbf{V}_{ij,n+1}^*\}$ using $\hat{\boldsymbol{\theta}}_{0i}$.

Calculate λ_{1ij}^* using (4) with $\{\mathbf{V}_{ijk}^*\}_{k=1}^{n+1}$.

End For Loop.

$\lambda_{1i}^*(\alpha) \approx \alpha$th quantile of $\{\lambda_{1ij}^*\}_{j=1}^{B}$.

End For Loop.

$P(\lambda_1 \le \lambda_1^*(\alpha) | \Delta) \approx \sum_{i=1}^{M} I(\lambda_{1i} \le \lambda_{1i}^*(\alpha))/M$, where $I(\cdot)$ is the indicator function.

Table 1 shows $P(\lambda_1 \le \lambda_1(\alpha) | \Delta)$ and $P(\lambda_1 \le \lambda_1^*(\alpha) | \Delta)$ for M=1000 and 10000 with K= 100000, B=199, and α=0.05. As M increases, the power of the bootstrap likelihood ratio test λ_1 essentially converges to the power of λ_1 for large n. Note also that the estimate of Type I error, $P(\lambda_1 \le \lambda_1^*(\alpha) \mid \Delta = 0)$, is very close to its true value α=0.05 for M=10000.

Now set $p_2 = p_1 + 0.2\Delta_1$ and $\mu_2 = \mu_1 + \Delta_2\sigma$, where $\Delta_1 \in \{0,1,2\}$ and $\Delta_2 \in \{0, 0.25i, i=1,2,\cdots,12\}$. For p_1=0.5, μ_1=0 and σ=1, $P(\lambda_1 \le \lambda_1^*(\alpha) \mid \Delta_1, \Delta_2)$ are calculated and plotted in Figure 1 with n=100, B=199, M=1000, and α=0.05. As the separation between μ_1 and μ_2 increases, the power of the bootstrap likelihood ratio test increases. On the other hand, the test is not sensitive to separations between p_1 and p_2 when p_1 is around 0.5.

For the combinations of large p_1 and small p_2, which represent the real situation better, we examine the power of the test λ_1. We choose p_1=0.8, 0.9, 0.95 and p_2=0.05, 0.1. For μ_1=0, σ=1, and each combination of p_1 and p_2, $P(\lambda_1 \le \lambda_1^*(\alpha) \mid \Delta_2)$ is obtained and plotted in Figure 2 with $\Delta_2 \in \{0, 0.25i, i=1,2,\cdots,12\}$, n = 100, B = 199, M = 1000, and α = 0.05. That is, the power estimates for the test λ_1, $P(\lambda_1 \le \lambda_1^*(\alpha) \mid \Delta_2)$ are plotted in Figure 2 (a) and (b) for p_2 = 0.05, and for p_2 = 0.1 with different values of p_1(p_1 = 0.8, 0.9, 0.95), respectively. The plot shows that the larger difference between p_1 and p_2 produces the better power curve for p_1 = 0.8, 0.9, 0.95 and p_2 = 0.05, 0.1.

Table 1. Simulated power of the test of size $\alpha = 0.05$ for $n = 100$ using 199 bootstrap replications ($p_1=p_2=0.5$, $\mu_1=0$, $\sigma=1$)

	$\Delta=0$	$\Delta=1$	$\Delta=2$	$\Delta=3$
		M=1000		
$P(\lambda_1 \le \lambda_1(\alpha) \mid \Delta)$	0.054	0.162	0.495	0.832
$P(\lambda_1 \le \lambda_1^*(\alpha) \mid \Delta)$	0.056	0.157	0.491	0.826
		M=10000		
$P(\lambda_1 \le \lambda_1(\alpha) \mid \Delta)$	0.0497	0.1624	0.4986	0.8451
$P(\lambda_1 \le \lambda_1^*(\alpha) \mid \Delta)$	0.0494	0.1619	0.4940	0.8373

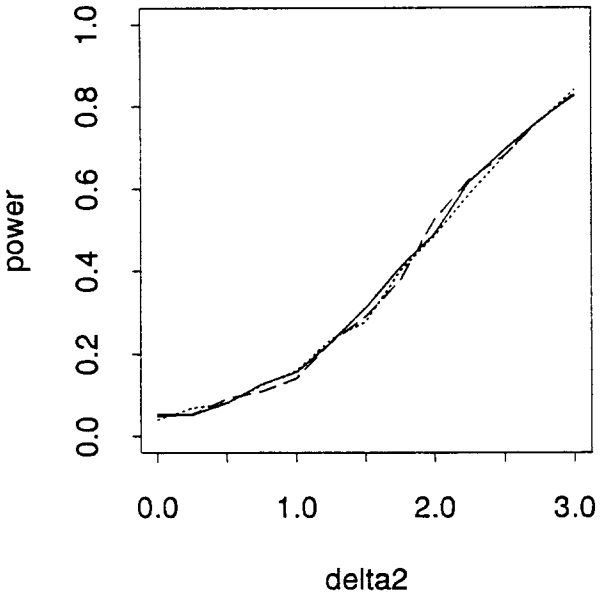
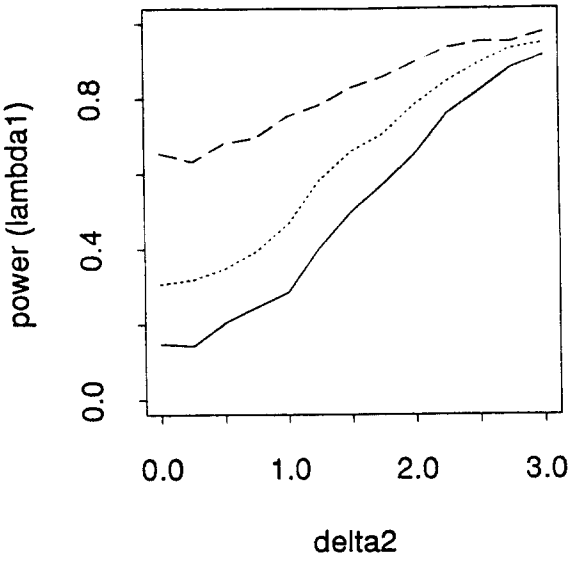
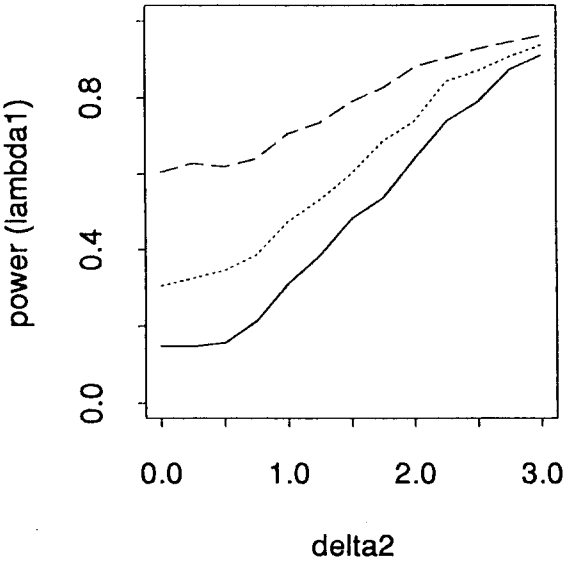


Figure 1. Power curves of the bootstrap liklehood ratio test λ_1 where $p_2 = p_1 + 0.2\Delta_1$ and $\mu_2 = \mu_1 + \Delta_2\sigma$, $p_1=0.5$, $\mu_1=0$, and $\sigma=1$. Solid line: $\Delta_1=0$, dotted line: $\Delta_1=1$, and broken line : $\Delta_1=2$. Delta2 denotes Δ_2.



(a) test λ_1: $p_2=0.05$, $\mu_2 = \mu_1 + \Delta_2\sigma$ where $\mu_1=0$ and $\sigma=1$.



(b) test λ_1: $p_2=0.1$, $\mu_2 = \mu_1 + \Delta_2\sigma$ where $\mu_1=0$ and $\sigma=1$.

Figure 2. Power curves of the bootstrap liklehood ratio test λ_1. Solid line: $p_1=0.8$, dotted line: $p_1=0.9$, and broken line: $p_1=0.95$. Delta2 denotes Δ_2.

4. Concluding remarks and discussion

This research implies that for outlier detection in the mixed variables (discrete and continuous) case, the bootstrap likelihood ratio method may be a useful tool. Moreover, although we have assumed normality for clarity, the methodology considered here can be applied to any mixture of continuous and discrete variables for which the likelihood ratio is defined. We can always construct the bootstrap likelihood ratio test whenever the MLEs exist in (3) even if the continuous distribution is not normal. Therefore the critical value $(\lambda_1(\alpha))$ for the test, controlling the type I error (α) of declaring an outlier when there is not, is easily obtained by the parametric bootstrap.

It should be noted that the power of the test depends on the training sample size n and the bootstrap replication size B. Small sample size may force the test to be less powerful, and a large value for B dramatically increases the computing time though it increases the accuracy. Thus the sensitivity of the test against n and B may be studied further.

We have assumed that the covariance of the continuous variable of the new observation $\mathbf{V}_{n+1}$ is the same as that of the population of the training sample. In many situations in practice, however, it may be more realistic that they are different from each other. If they are not equal, we cannot obtain the covariance estimate of $\mathbf{V}_{n+1}$ in $\hat{\boldsymbol{\theta}}$, the MLE of $\boldsymbol{\theta}$ in the denominator of (3) since it is not reasonable to estimate the covariance of $\mathbf{V}_{n+1}$ with the only one observation. We may, however, be able to estimate the covariance of the continuous features of unusual event from the past nuclear explosions data. Let $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ be the covariance matrix of $\mathbf{X}$ given $\mathbf{Y}$ from the training sample and from the new observation, respectively. When the covariances are not equal $(\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2)$, we actually test $H_0^{**} : \mathbf{p}_1 = \mathbf{p}_2,\ \mathbf{U}_1 = \mathbf{U}_2,\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ against $H_1^{**} : \mathbf{p}_1 \neq \mathbf{p}_2$ or $\mathbf{U}_1 \neq \mathbf{U}_2$ or $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. Thus it seems reasonable to modify the denominator of the likelihood ratio, λ_1 and we may define the new ratio, λ_2, by

$$\lambda_2 = \frac{\sup_{(\boldsymbol{\theta}\in\Omega_0)} L(\boldsymbol{\theta}; \mathbf{V}_1, \cdots, \mathbf{V}_{n+1})}{\sup_{(\boldsymbol{\theta}\in\Omega_0)} L(\boldsymbol{\theta}; \mathbf{V}_1, \cdots, \mathbf{V}_n) f(\mathbf{V}_{n+1}; \hat{\mathbf{p}}_2, \hat{\mathbf{U}}_2, \hat{\boldsymbol{\Sigma}}_2)} \tag{5}$$

$$= \frac{L(\hat{\boldsymbol{\theta}}_0; \mathbf{V}_1, \cdots, \mathbf{V}_{n+1})}{L(\hat{\boldsymbol{\theta}}_n; \mathbf{V}_1, \cdots, \mathbf{V}_n) f(\mathbf{V}_{n+1}; \hat{\mathbf{p}}_2, \hat{\mathbf{U}}_2, \hat{\boldsymbol{\Sigma}}_2)},$$

where $\hat{\boldsymbol{\theta}}_n$ is the $\boldsymbol{\theta} \in \Omega_0$ which attains $\sup L(\boldsymbol{\theta}; \mathbf{V}_1, \cdots, \mathbf{V}_n)$, and $\hat{\mathbf{p}}_2$, $\hat{\mathbf{U}}_2$, $\hat{\boldsymbol{\Sigma}}_2$, are the

parameter estimates obtained from the past nuclear explosion events. It is noted that $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$ from the training sample under the null hypothesis H_0 while $\hat{\boldsymbol{\theta}}_0$ is the estimate defined in (4). From (5) it is clear that small values of λ_2 provide evidence against H_0, and thus $\mathbf{V}_{n+1}$ can be classified as an outlier by rejecting H_0 if $\lambda_2 \leq \lambda_2(\alpha)$, where $\lambda_2(\alpha)$ is the αth percentile of the distribution of λ_2 under H_0. The parametric bootstrap procedure to obtain the critical value $\lambda_2(\alpha)$ can be carried out similarly as for λ_1.

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis,* John Wiley & Sons, New York.

[2] Bacon-Shone, J. and Fung, W.K. (1987). A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data, *Applied Statistics, Vol.* 36, 153-162.

[3] Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data,* Wiley, Chicester, England.

[4] Campbell, N.A. (1980). Robust Procedures in Multivariate Analysis: I, Robust Covariance Estimation, *Applied Statistics, Vol.* 29, 231-237.

[5] Caroni, C. and Prescott, P. (1992). Sequential Application of Wilks's Multivariate Outlier Test, *Applied Statistics, Vol.* 41, 355-364.

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

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

[8] Efron, B. and Tibshirani, R. (1986). Bootstrap Method for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, *Statistical Science,* Vol. 1, 54-77.

[9] Hadi, A.S. (1992). Identifying Multiple Outliers in Multivariate Data, *Journal of the Royal Statistical Society, Series* B, Vol. 54, 761-771.

[10] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* Wiley, New York.

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

[12] Olkin, I. and Tate, R.F. (1961). Multivariate Correlation Models with Mixed Discrete and Continuous Variables, *The Annals of Mathematical Statistics,* Vol. 22, 92-96.

[13] Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point, *In Mathematical Statistics and Applications (eds W. Grossmann, G. Pflug, I. Vincze and W. Wertz), Vol.* B, 283-297, Reidel, Dordrecht.

[14] Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points (with comments), *Journal of the American Statistical Association, Vol.* 85, 633-651.

[15] Wilks, S.S. (1963). Multivariate Statistical Outliers, *Sankhya, Vol.* 25, 407-426.

이상점 탐지를 위한 일반화 우도비 검정

백장선[2)]

요약

본 연구에서는 핵확산 감시와 관련된 이상점 탐지를 위한 일반화 우도비 검정 방법이 개발되었다. 고전적인 이상점 탐지방법들이 연속형 변수만을 고려한 반면, 본 연구에서 제안된 방법은 연속형 변수, 이산형 변수, 혹은 이산형과 연속형이 혼합된 변수들에 모두 적용될 수 있다. 더우기 대부분의 고전적인 방법들에 있어서 주로 이용된 정규분포 가정을 필요로 하지 않는다. 본 연구에서 제안된 방법은 일반화 우도비에 붓스트랩 방법을 적용하여 구성되었다. 모의실험을 통하여 검정력을 고찰함으로써 제안된 검정방법의 성능을 연구하였다.

2) (500-757) 광주직할시 북구 용봉동 300, 전남대학교 통계학과.