Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Discrimination and Clustering for Multivariate Time Series

Yoshihide Kakizawa , Robert H. Shumway & Masanobu Taniguchi

To cite this article: Yoshihide Kakizawa , Robert H. Shumway & Masanobu Taniguchi (1998) Discrimination and Clustering for Multivariate Time Series, Journal of the American Statistical Association, 93:441, 328-340

To link to this article: https://doi.org/10.1080/01621459.1998.10474114

Published online: 17 Feb 2012.

Discrimination and Clustering for Multivariate Time Series

Yoshihide KAKIZAWA, Robert H. SHUMWAY, and Masanobu TANIGUCHI

Minimum discrimination information provides a useful generalization of likelihood methodology for classification and clustering of multivariate time series. Discrimination between different classes of multivariate time series that can be characterized by differing covariance or spectral structures is of importance in applications occurring in the analysis of geophysical and medical time series data. For discrimination between such multivariate series, Kullback–Leibler discrimination information and the Chernoff information measure are developed for the multivariate non-Gaussian case. Asymptotic error rates and limiting distributions are given for a generalized spectral disparity measure that includes the foregoing criteria as special cases. Applications to problems of clustering and classifying earthquakes and mining explosions are given.

KEY WORDS: Chernoff; Divergence; Kullback–Leibler; Minimum discrimination information; Robustness; Seismology; Spectral analysis.

1. INTRODUCTION

Classic problems in analyzing observed time series involve the grouping or clustering of such series into similar categories and the classification of new observed series known to belong to one of two or more categories. These two problems have, of course, been studied for conventional vector observations, and there exists a huge literature (see, e.g., Johnson and Wichern 1992; McLachlan 1992) devoted to discriminant and cluster analysis as applied to collections of multivariate normal vectors. Such methodology usually depends on isolating differences between the subpopulation means; the resulting linear functions are well adapted for computations and easily applied to clustering and discriminant analyses involving real data.

For cases where one may observe stationary vector time series, the similarities and differences between members of subpopulations may not always be characterized by the structures. Because a vector time series often involves thousands of correlated observations collected over time, the dimensionality of the time series case prohibits computations in the time domain using classical multivariate methods. The preponderance of such vector series in established disciplines such as seismology and in newly emerging studies involving data collected in functional magnetic resonance imaging studies makes discriminant and cluster analysis of vector time series of great interest. An early application of a parametric time domain version of a technique related to one of those considered in this article was to the problem of discriminating between levels of anesthesia that are insufficient or sufficient for deep surgery by Gersch, Martinelli, Yonemoto, Low, and MacEwan (1979). They developed a nearest-neighbor approach, based on the Kullback–Leibler information measure, for classifying bivariate electroencephalogram records. The approach measured distance using parametric autoregressive models for the autocovariance matrix and did not consider the approximations based on the spectrum that we are proposing in this article.

We mention first a specific example of classification techniques as they are applied in seismology, where the differences between realizations of vector time series often provide insight into similarities and differences between classes of events such as earthquakes, mining explosions, or nuclear explosions. Monitoring nuclear proliferation depends critically on being able to discriminate reliably between small nuclear explosions and the other categories. For large events, observed at teleseismic distances (10,000–15,000 km), the problem can be formulated in terms of discriminating between means, and this led to early procedures based either on features extracted from the waveforms (as in Booker and Mitronovas 1964) or on signal discrimination between mean vectors (as in Shumway and Unger 1974). Although such methods work well for large nuclear explosions, they are not well suited to either clustering or discriminating small events observed at regional distances (100–2,000 km), as is contemplated in current requirements for monitoring the proliferation of nuclear weapons.

More recent analyses have involved analyzing such regional data observed closer to the source. Because there is a paucity of such data, one can frequently look at mining explosions as surrogates that are expected to behave similarly to low-yield nuclear explosions. Figure 1 shows a typical earthquake and a mining explosion in Scandinavia, as measured by stations located in Scandinavia. The earthquake and explosion in Figure 1 are typical of those in the set taken as base data for this study, compiled by Blandford (1993). All events chosen were on or near land and were distributed uniformly over Scandinavia so as to minimize the possibility that discriminators might be keying on location or land-sea differences. A listing of the events, shown in Table 1, shows earthquakes (EQ) ranging in magnitude from 2.74 to 4.40 and explosions (EX) in the range 2.13 to 2.19. We added an event of uncertain origin that

Yoshihide Kakizawa is Lecturer, Faculty of Economics, Hokkaido University, Sapporo 060, Japan. Robert H. Shumway is Professor, Division of Statistics, University of California, Davis, CA 95616. Masanobu Taniguchi is Associate Professor, Department of Mathematical Science, Osaka University, Toyonaka 560, Japan. The authors are pleased to acknowledge the helpful and insightful comments of the reviewers and the associate editor. In particular, they acknowledge the reviewer who suggested the partitioning method for clustering. The article benefited also from a more detailed analysis of the method for choosing the regularization parameter α, in answer to a query from two of the reviewers.

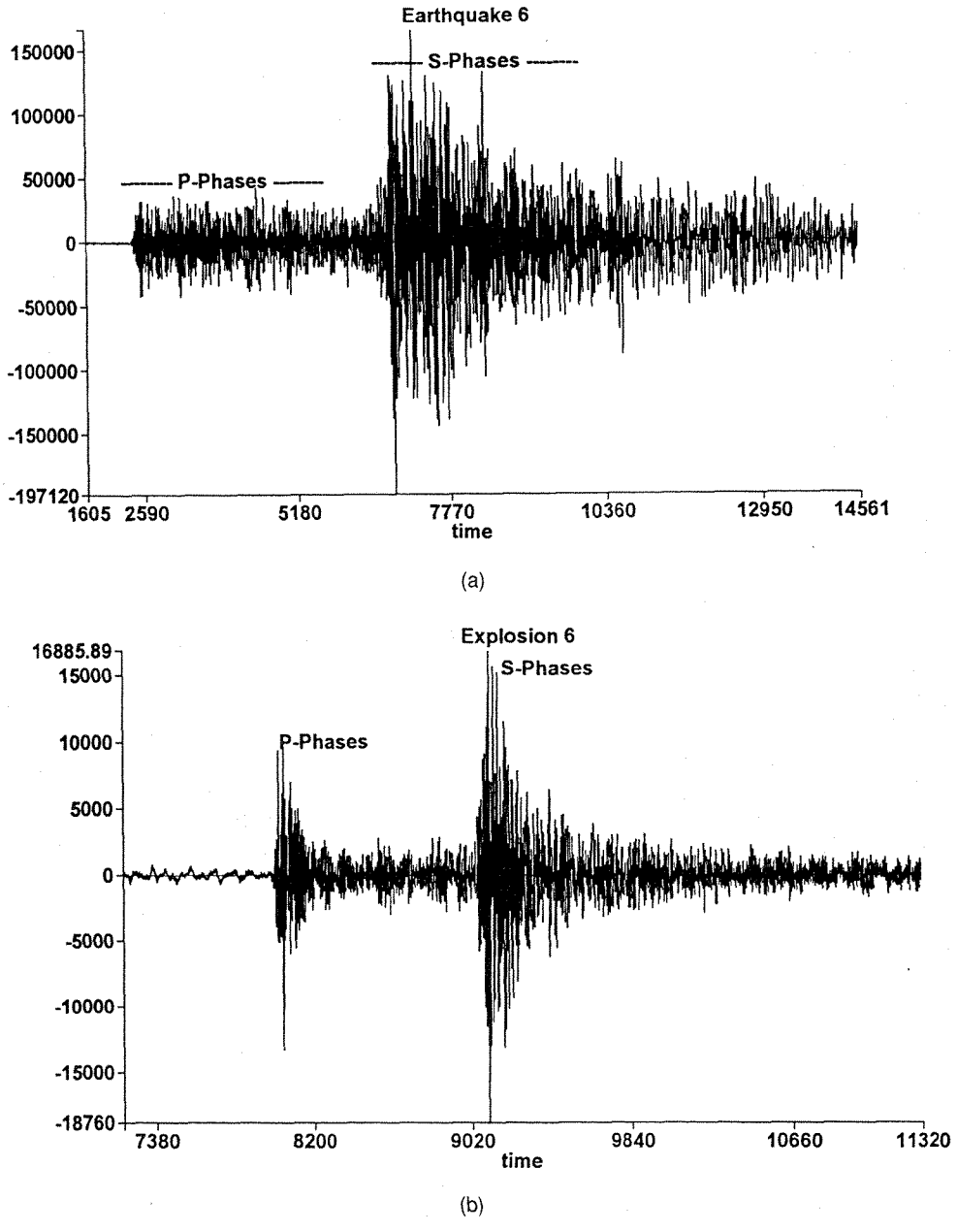

Figure 1. Typical Earthquake (a) and Explosion (b) Recorded at a Regional Array (40 Points/Second). Note the separate P and S arrivals, taken as the components of a two-dimensional vector time series.

was located in the Novaya Zemlya region of Russia (see, e.g., Ryall, Baumgardt, Fisk, and Riviere-Barbier 1996). All events except the Russian event occurred in the Scandinavian peninsula and were recorded by seismic arrays located in Norway by Norwegian and Arctic experimental seismic stations (NORESS, ARCESS) and in Finland by Finnish experimental seismic stations (FINESS).

Each plotted record, as shown in Figure 1, is actually composed of two phases: the *P group* of first arrivals and the *S group* of later arrivals. The arrivals are really different signals that take different paths through the crust and are separated by a fairly long time interval. Seismologists tend to identify the first segment of each of the phases as being of importance, and we continue that practice by separating the P and S phases into two separate signals. Hence we think of both phases together in Figure 1 as a bivariate vector series with 1,024 points in each vector. Extracting the two phases and plotting for two typical earthquakes and explosions, we obtain the four bivariate processes displayed in Figure 2. The P and S groups can often be broken down further into shorter phases, identified as P_n, P_g and S_n, L_g but we do not pursue this partition here, because the subphases are not clearly discernible in the mix. In general, the starting points of the initial P and later S phases can be identified directly from the records, as in Figure 1. If more accurate specifications are needed, one can use the known locations and velocities of the various components to pick an exact start time. Slight variations in the choice of the start time will have small negligible effects on the off-diagonal elements of the spectral matrices used as components of the discrimination methodology.

The simplest approaches to discriminating between the earthquake and explosion group have been based on either

Table 1. Seismic Events Used for Clustering and Discriminant Analysis

No.	Type	Date	Array	Magnitude	Latitude	Longitude
1	EQ	6/16/92	FINESS	3.22	65.5	22.9
2	EQ	8/24/91	ARCESS	3.18	65.7	32.1
3	EQ	9/23/91	NORESS	3.15	64.5	21.3
4	EQ	1/4/92	FINESS	3.60	67.8	15.1
5	EQ	2/19/92	ARCESS	3.26	59.2	10.9
6	EQ	4/13/92	NORESS	4.40	51.4	6.1
7	EQ	4/14/92	NORESS	3.38	59.5	5.9
8	EQ	5/18/92	NORESS	2.74	66.9	13.7
9	EX	3/23/91	ARCESS	2.85	69.2	34.3
10	EX	4/13/91	FINESS	2.60	61.8	30.7
11	EX	4/26/91	ARCESS	2.95	67.6	33.9
12	EX	8/3/91	ARCESS	2.13	67.6	30.6
13	EX	9/5/91	ARCESS	2.32	67.1	21.0
14	EX	12/10/91	FINESS	2.59	59.5	24.1
15	EX	12/29/91	ARCESS	2.96	69.4	30.8
16	EX	3/25/92	NORESS	2.94	64.7	30.8
17	NZ	12/31/92	NORESS	2.50	73.6	55.2

the relative amplitudes of the P and S phases or on relative power components in various frequency bands. Inspecting the bivariate series plotted in Figure 2 shows, for example, that the ratios of the S amplitudes to the P amplitudes for earthquakes tend to be somewhat higher than those for explosions. This suggests extracting the amplitudes of the P and S phases as bivariate *discrimination features*. A scatterplot of the logarithms of the amplitudes, shown in Figure 3, indicates that a rough separation might be possible using this variable as a discriminant. Considerable effort has also been expended on using various spectral ratios involving the P and S phases. Bennett and Murphy (1986) have noted that for western U.S. events, earthquake L_g (S) contained more high frequencies than the explosion L_g (Note that the opposite appears to be true for the Scandinavian events in Figure 2.) Taylor, Denny, Vergino, and Glaser (1989) used the similar spectral ratios for the P phase as well. Dysart and Pulli (1990) and Pulli (1996) have considered spectral ratios for Scandinavian events and noted that the ratio of the S spectra to P spectra is generally higher for earthquakes than for explosions. Richards, Kim, and Ekström (1993) noted that for eastern U.S. events, the ratios of P to S spectra are generally higher for explosions; that is, S to P spectra are again higher for earthquakes. Pulli (1996) has suggested retaining integrated power in selected frequency bands; the recommended bands are 2–5 Hz, 5–10 Hz, and 10–20 Hz, where Hz denotes Hertz measured in cycles per second. We show a scatterplot of the 2–5 Hz integrated power in the lower panel of Figure 3 for the 17 events in Table 1. Various frequency bands can be chosen and the ratios used for conventional discriminant analysis on the means as in the aforementioned references (see also Shumway 1996).

The fact that all of the foregoing discriminants involve the spectrum in some way or an amplitude that is proportional to the integrated spectrum suggests that all of the important information for discriminating between the two classes might be contained in the spectral matrices of the bivariate (P, S) series. This in turn suggests that the some measure of disparity or divergence between events in various classes might be formulated in terms of the spectral matrices. We may argue, as mentioned earlier, that it is differences in the covariance matrices, not the mean departures, that determine differences between event classes. Figure 2, for example, shows quite different waveforms for each of the earthquakes and explosions, suggesting that similarities within the two classes, if they exist, are stochastic

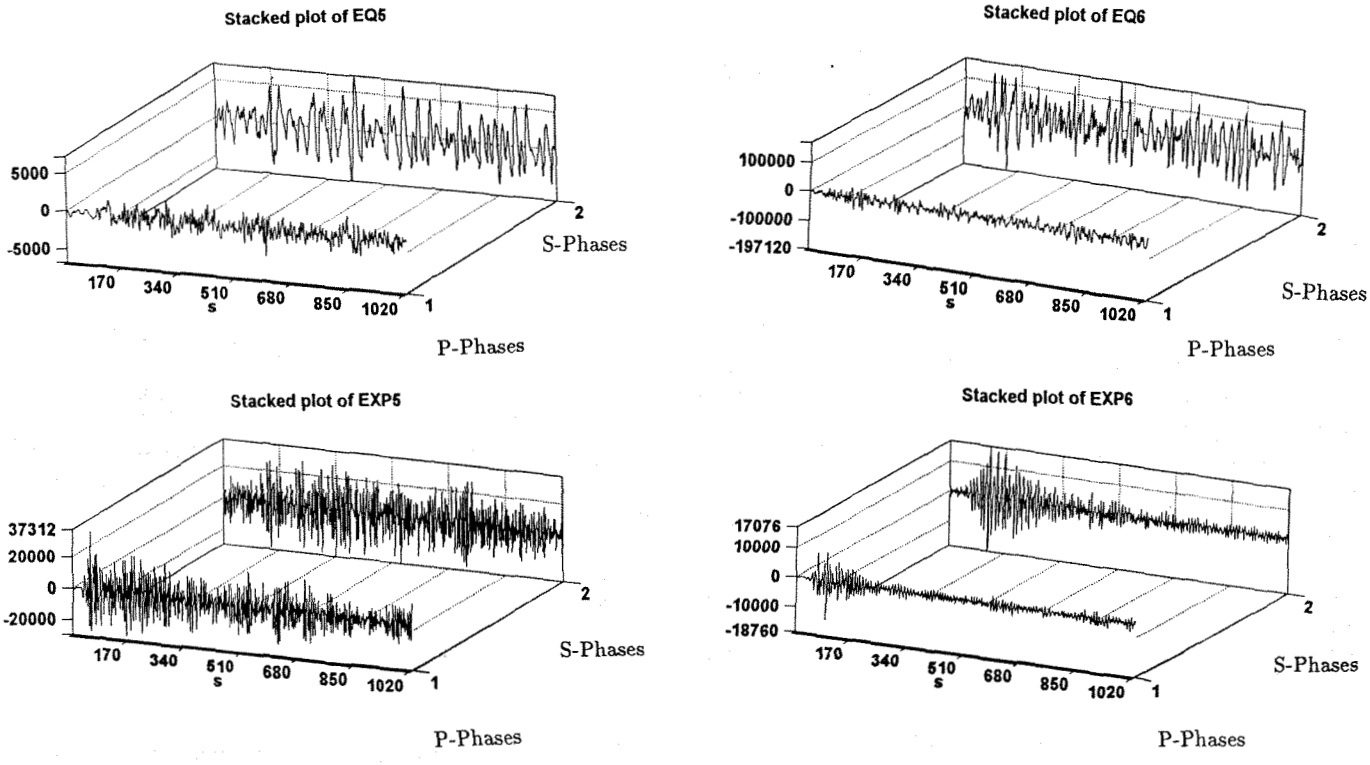

Figure 2. Two Earthquakes and Explosions, Shown as Bivariate Series Composed of Separate P and S Arrivals.

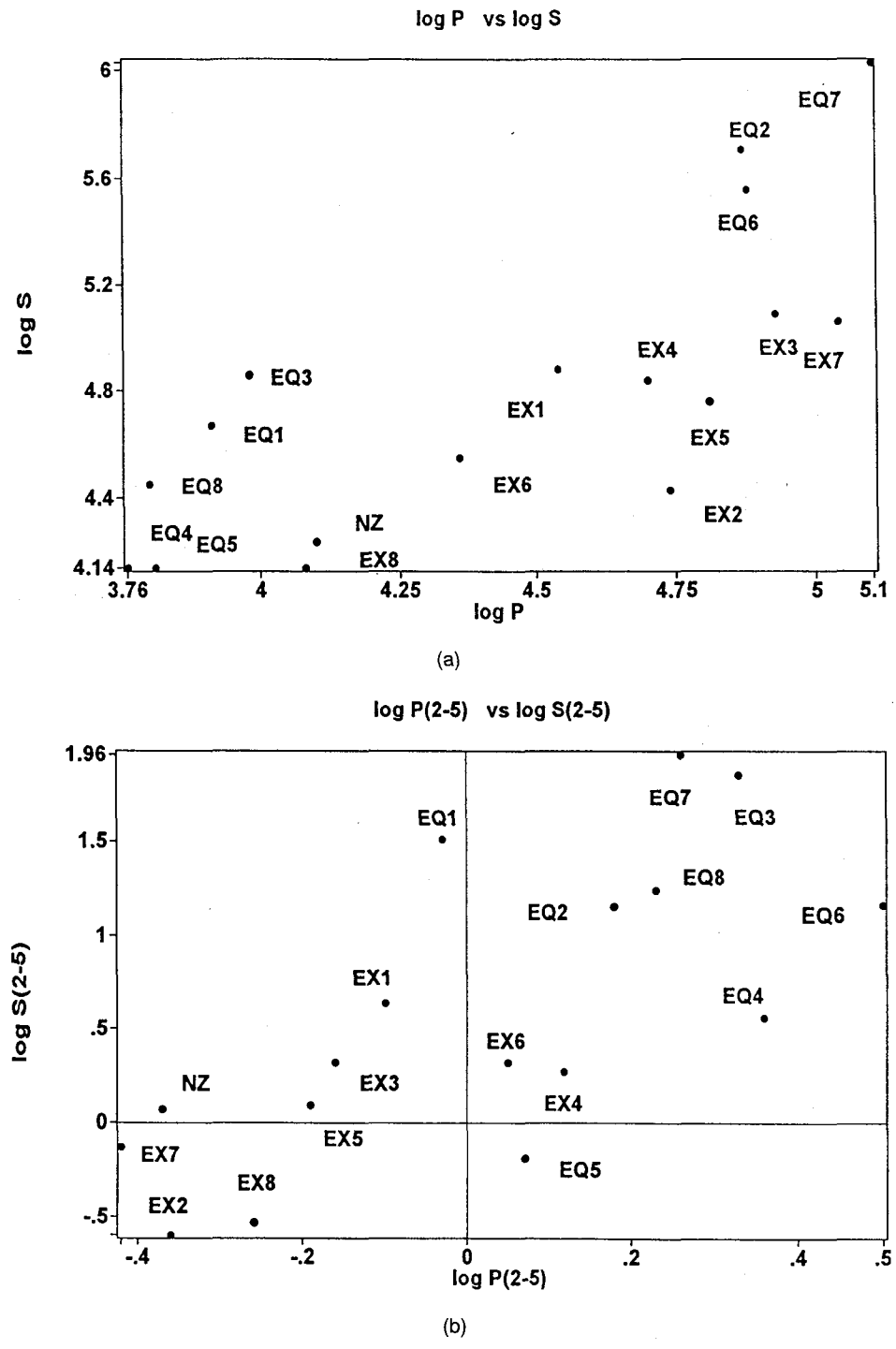

Figure 3. Features Commonly Extracted From Earthquake and Explosion Recordings: (a) Logarithms of Amplitudes and (b) Integrated Spectra (2–5 Hz).

in nature. Based on this, Alagón (1989), Dargahi-Noubary and Laycock (1981), and Shumway (1982) have applied approximations based on the Whittle (1954) likelihood for treating the problem of discriminating between univariate processes with different spectral densities. This latter approach might be termed *signal discrimination,* whereas the former approach based on particular spectral ratios or amplitudes is basically a *feature extraction* method. Shumway (1996) reviewed and compared the results using these two approaches. For the purpose of comparison, we include a very brief discussion of the amplitude and spectral feature extraction approach to discrimination in Section 4.

To summarize, the preceding remarks tend to suggest that similarities among and differences between multivariate stationary time series can be characterized in terms of the structure of the covariance or, equivalently, the spectral matrices. We follow this lead by developing various multivariate measures of disparity between two vector stationary processes. It is argued that the Kullback–Leibler (1951) and Chernoff (1952) (see also Renyi 1961) information measures form the natural building blocks for constructing measures appropriate for clustering and discriminating among multivariate time series. In particular, note that Parzen (1990) proposed the Chernoff information (see also

Renyi 1961) and that Zhang and Taniguchi (1994, 1995) have shown robustness of this measure to non-Gaussian departures and to peak contamination in certain univariate situations. Shumway (1996) compared univariate classification results for earthquakes and explosions using both the Kullback–Leibler and Chernoff information measures.

This article first develops the Kullback–Leibler and Chernoff information measures, referred to as KL and CH in the sequel, as special cases of a general *disparity measure* that depends on a simple function of the spectral matrices. A process is defined to be close to another process if the two sample spectral matrices are close in the sense of yielding a small value for one of the disparity measures. A process is close to one having a particular hypothetical covariance matrix if its sample spectral matrix is close to the hypothetical spectral matrix of the target class. In this way we define measures of disparity between processes and target classes and propose computationally feasible procedures for cluster and discriminant analysis. For series that are not previously members of well-defined groups, a cluster analysis can establish hierarchical or partitioned groupings and suggest classifications that tend to minimize various within-group disparities. We use the various disparity measures developed as elements of a distance matrix and then apply both hierarchical clustering and partitioning techniques (Johnson and Wichern 1992) Blashford, Alenderfer, and Morey 1982; to identify natural clusterings within the group of earthquakes and explosions of Table 1, augmented by one event of unknown origin.

When the elements of the population belong to well-defined alternative populations, we are interested in classifying a new realization of unknown origin into one of the defined populations. Rules for discriminant analysis are proposed that exploit the disparity between the sample spectral matrices and the group-average spectral matrices, as measured by the Kullback–Leibler and Chernoff information criteria. The resulting methodology is applied to the earthquake and explosion populations and to the new event of unknown origin. A theoretical section follows that establishes asymptotic behavior of a general disparity criterion that includes both the KL and CH criteria as special cases. Convergence in probability of the sample disparity and a central limit theorem are established. Under contiguous alternatives, the misclassification rates are shown to converge to normal integrals, and a condition for non-Gaussian robustness is proposed.

2. MEASURES OF DISPARITY

We suppose first that we have a collection of zero-mean m-dimensional vector stationary time series $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_m(t))'$, $t = 1, 2, \ldots, T$. The probability density functions of the $mT \times 1$ vector $\mathbf{x} = (\mathbf{X}(1)', \mathbf{X}(2)', \ldots, \mathbf{X}(T)')'$ are denoted by $p(\mathbf{x})$ and $q(\mathbf{x})$, where the two density functions typically correspond to two different hypotheses about the observed vector $\mathbf{x}$. In the stationary case, we use the $\mathbf{f}(\lambda)$ and $\mathbf{g}(\lambda)$ for the spectral density matrices corresponding to the $m \times m$ matrices of autocovariance functions $\mathbf{R}_p(s-t)$ and $\mathbf{R}_q(s-t)$. Although the theory developed later transcends the usual normality assumption, it is convenient to make this assumption temporarily in motivating the measures of disparity between the densities $p(\cdot)$ and $q(\cdot)$.

One classical measure of disparity between two multivariate densities is the Kullback–Leibler (KL) discrimination information, defined (Kullback and Leibler 1951; Kullback (1978) by

$$I(p;q) = E_p\left\{\log\frac{p(\mathbf{x})}{q(\mathbf{x})}\right\}, \tag{1}$$

where E_p denotes the expectation under the density $p(\cdot)$. The KL discrimination information takes the form

$$I(p;q) = \frac{1}{2}\left(\mathrm{tr}\{\mathbf{R}_p\mathbf{R}_q^{-1}\} - \log\frac{|\mathbf{R}_p|}{|\mathbf{R}_q|} - mT\right) \tag{2}$$

when $p(\mathbf{x})$ and $q(\mathbf{x})$ correspond to competing zero-mean multivariate normal distributions. The $mT \times mT$ covariance matrices $\mathbf{R}_p$ and $\mathbf{R}_q$ contain the $m \times m$ matrices $\mathbf{R}_p(s-t)$, $\mathbf{R}_q(s-t)$, $s, t = 1, \ldots, T$ as blocks. A symmetric measure of disparity, the *J divergence,* is defined as

$$J(p;q) = I(p;q) + I(q;p) \tag{3}$$

(see Kullback 1978). This has all the properties of a distance except the triangle inequality property and hence is called a *quasi-distance.* The J divergence will be particularly useful for hierarchical cluster analysis.

Parzen (1990) proposed using the Chernoff (CH) information measure (Chernoff 1952; Renyi 1961)

$$B_\alpha(p;q) = -\log E_p\left\{\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right)^\alpha\right\} \tag{4}$$

as a measure of disparity between the two densities, where the measure is indexed by α, $0 < \alpha < 1$. For $\alpha = .5$, the Chernoff information measure is the symmetric divergence measure proposed by Bhattacharya (1943), and we retain the notation $B_\alpha(\cdot\,;\cdot)$ because of that connection. For two normal random vectors differing only in the covariance structure, the foregoing measure takes the value

$$B_\alpha(p;q) = \frac{1}{2}\left(\log\frac{|\alpha\mathbf{R}_p + (1-\alpha)\mathbf{R}_q|}{|\mathbf{R}_q|} - \alpha\log\frac{|\mathbf{R}_p|}{|\mathbf{R}_q|}\right). \tag{5}$$

It is of interest to note the antisymmetry property $B_\alpha(p;q) = B_{1-\alpha}(q;p)$ and that $B_\alpha(p;q)$, scaled by $\alpha(1-\alpha)$ converges to $\mathbf{I}(p;q)$ for $\alpha \to 0$ and to $\mathbf{I}(q;p)$ for $\alpha \to 1$. Hence the Chernoff measure tends to behave like the two Kullback–Leibler measures for values of the parameter α that are near the boundaries 0 and 1. Again, we may define a quasi-distance by letting

$$JB_\alpha(p;q) = B_\alpha(p;q) + B_\alpha(q;p). \tag{6}$$

It should be recognized that the information measures (2) and (5) both involve $mT \times mT$ matrices that will be very large and unwieldy for observed data such as are given in Figure 2, where $mT = 2{,}048$. It is usual in such cases to use spectral approximations based on the limiting values of (2) and (5), expressed in terms of the spectral matrices $\mathbf{f}(\lambda)$ and $\mathbf{g}(\lambda)$ of the process $\mathbf{X}(t)$. For example, Kazakos

and Papantoni-Kazakos (1980) (see also Hannan 1970 and Pinsker 1964, Shumway and Unger 1974 for special cases) have derived the appropriate limiting versions of (2) and (5), say $\lim_{T\to\infty} T^{-1}\mathbf{I}(p;q)$ and $\lim_{T\to\infty} T^{-1}B_\alpha(p;q)$, as

$$I(\mathbf{f};\mathbf{g}) = \frac{1}{2}\int_{-\pi}^{\pi}\left(\text{tr}\{\mathbf{f}\mathbf{g}^{-1}\} - \log\frac{|\mathbf{f}|}{|\mathbf{g}|} - p\right)\frac{d\lambda}{2\pi} \tag{7}$$

and

$$B_\alpha(\mathbf{f};\mathbf{g}) = \frac{1}{2}\int_{-\pi}^{\pi}\left(\log\frac{|\alpha\mathbf{f}+(1-\alpha)\mathbf{g}|}{|\mathbf{g}|} - \alpha\log\frac{|\mathbf{f}|}{|\mathbf{g}|}\right)\frac{d\lambda}{2\pi}, \tag{8}$$

where we suppress the argument λ in $\mathbf{f}(\lambda)$ and $\mathbf{g}(\lambda)$ within integrals when there can be no confusion. Note here that the spectral matrices $\mathbf{f}(\lambda)$ and $\mathbf{g}(\lambda)$ correspond to the multivariate densities $p(\mathbf{x})$ and $q(\mathbf{x})$. The advantage of the limiting forms is that the evaluation problem is reduced to inverting $m \times m$ matrices and then approximating the integrals (7) and (8) by sums. The fact that the limits of (2) and (5) are (7) and (8) implies that for large T, no information is lost in going from the covariance matrix representation to a representation in terms of the spectral matrices. This reduces the dimension of the calculation from one involving $mT \times mT$ matrices to one involving $m \times m$ matrices.

Both forms (7) and (8) are functions of the matrix product $\mathbf{f}(\lambda)\mathbf{g}^{-1}(\lambda)$ and can be seen to be special cases of a more general *disparity measure,* which can be written as

$$D_H(\mathbf{f};\mathbf{g}) = \frac{1}{2}\int_{-\pi}^{\pi} H(\mathbf{f}\mathbf{g}^{-1})\frac{d\lambda}{2\pi} \tag{9}$$

for some matrix-valued function $H(\cdot)$. To ensure that general measures of the form $D_H(\mathbf{f};\mathbf{g})$ in (9) have the quasi-distance property, we require $D_H(\mathbf{f};\mathbf{g}) \geq 0$ with equality if and only if $\mathbf{f} = \mathbf{g}$ almost everywhere. The function $H(\mathbf{Z})$ must have a unique minimum at $\mathbf{Z} = \mathbf{I}_m$, the identity matrix. Certainly, there are many possible choices of $H(\mathbf{Z})$ such that $D_H(\cdot\,,\cdot)$ satisfies the quasi-distance property, but we consider only the two corresponding to (7) and (8):

$$H_{\mathbf{I}}(\mathbf{Z}) = \text{tr}\{\mathbf{Z}\} - \log|\mathbf{Z}| - p \tag{10}$$

and

$$H_B(\mathbf{Z}) = \log|\alpha\mathbf{Z} + (1-\alpha)\mathbf{I}_p| - \alpha\log|\mathbf{Z}|. \tag{11}$$

Note that another possible choice is the quadratic function

$$H_Q(\mathbf{Z}) = \frac{1}{2}\,\text{tr}(\mathbf{Z} - \mathbf{I}_p)^2. \tag{12}$$

Generally, $D_H(\cdot\,,\cdot)$ is not symmetric but can easily be made so by defining

$$\tilde{H}(\mathbf{Z}) = H(\mathbf{Z}) + H(\mathbf{Z}^{-1}), \tag{13}$$

and we obtain the symmetric J divergence in (3) and the Chernoff quasi-distance in (6). The general form (9) and hence the various cases defined by the Kullback–Leibler discrimination information (7) and the Chernoff information divergence (8) can be approximated in the usual fashion by sums over frequencies of the form $\lambda_s = 2\pi s/T$, $s = 1, 2, \ldots, T$, say

$$D_H(\mathbf{f};\mathbf{g}) \approx \frac{1}{2}\,T^{-1}\sum_s H(\mathbf{f}(\lambda_s)\mathbf{g}^{-1}(\lambda_s)). \tag{14}$$

In applications, the sum can be taken over a subset of frequencies to keep out higher frequencies where the determinant of the spectral matrix $\mathbf{g}(\lambda)$ is near 0. This corresponds to the weighted measure

$$D_{\phi H}(\mathbf{f};\mathbf{g}) = \frac{1}{2}\int_{-\pi}^{\pi}\phi(\lambda)H(\mathbf{f}(\lambda)\mathbf{g}^{-1}(\lambda))\frac{d\lambda}{2\pi}, \tag{15}$$

where $\phi(\cdot)$ is any nonnegative symmetric weight function.

3. CLUSTER ANALYSIS

The measures of disparity between spectral matrices developed in the previous section can be used as quasi-distance measures for clustering multivariate vector series. For example, let $\mathbf{f}_T(\lambda_s)$ and $\mathbf{g}_T(\lambda_s)$ be spectral matrix estimators for two different vector series, computed here as the usual nonparametric estimator. Then consider the approximations for the distance measures computed from the J divergence,

$$J(\mathbf{f}_T;\mathbf{g}_T) = \frac{1}{2}\,T^{-1}\sum_s(\text{tr}\{\mathbf{f}_T\mathbf{g}_T^{-1}\} + \text{tr}\{\mathbf{g}_T\mathbf{f}_T^{-1}\} - 2p), \tag{16}$$

and the symmetric Chernoff information divergence,

$$JB_\alpha(\mathbf{f}_T;\mathbf{g}_T) = \frac{1}{2}\,T^{-1}\sum_s\left(\log\frac{|\alpha\mathbf{f}_T+(1-\alpha)\mathbf{g}_T|}{|\mathbf{g}_T|} + \log\frac{|\alpha\mathbf{g}_T+(1-\alpha)\mathbf{f}_T|}{|\mathbf{f}_T|}\right). \tag{17}$$

These two distances come from applying (7) and (8) under the approximation (14).

Therefore, it is natural to propose using the two quasi-distances (16) and (17) to cluster similar time series. Using (16) or (17) on a sample of time series produces a quasi-distance matrix that can be used as input for one of the hierarchical clustering procedures (see Johnson and Wichern 1992). In general, one clusters first the two elements closest in the sense of (16) or (17). Then these two items become a cluster, and one can compute distances between nonclustered items as before. The distance between nonclustered items and a current cluster is defined as the average of the distances to elements in the cluster. Again, we combine the objects that are closest together. One may also compute the distance between nonclustered and clustered items as the distance between the nonclustered item and the closest element of the cluster. The clusters identified at each distance value can be displayed in a tree structure, or *dendrogram.*

Alternatively, one may think of clustering as a partitioning of the sample into a prespecified number of groups. MacQueen (1967) has proposed *k-means clustering* using

the Mahalonobis distance between an observation and the group mean vectors with a reassignment procedure that puts vectors into their closest affinity group. To see how this procedure could be applied in the present context, consider a preliminary partition of the sample into k groups and define the disparity between the spectral matrix of any particular sample vector series, say $\mathbf{f}_T(\lambda)$, and the spectrum of group j, say $\mathbf{f}_j(\lambda)$, in terms of the discrete approximations to (7) and (8) [see also (14)]—namely,

$$I(\mathbf{f}_T;\mathbf{f}_j) = \frac{1}{2}\,T^{-1}\sum_s \left(\operatorname{tr}\{\mathbf{f}_T\mathbf{f}_j^{-1}\} - \log\frac{|\mathbf{f}_T|}{|\mathbf{f}_j|} - p\right) \quad (18)$$

and

$$B_\alpha(\mathbf{f}_T;\mathbf{f}_j) = \frac{1}{2}T^{-1}\sum_s \left(\log\frac{|\alpha\mathbf{f}_T + (1-\alpha)\mathbf{f}_j|}{|\mathbf{f}_j|} - \alpha\log\frac{|\mathbf{f}_T|}{|\mathbf{f}_j|}\right). \quad (19)$$

We may estimate the group spectral matrix using the average sample spectral matrix for the jth group, say

$$\bar{\mathbf{f}}_j(\lambda) = \frac{1}{n_j}\sum_{l=1}^{n_j}\mathbf{f}_T^{(l)}(\lambda), \quad (20)$$

where $\mathbf{f}_T^{(l)}$ are the nonparametric estimators for the spectral matrix of lth realization in group j for $l = 1, 2, \ldots, n_j$. With this as a proviso, make either an arbitrary or systematic partition of the observed series into k groups and compute the disparity between a sample value and each of the group means using either $I(\mathbf{f}_T;\bar{\mathbf{f}}_j)$ or $B_\alpha(\mathbf{f}_T;\bar{\mathbf{f}}_j)$. Assign the series to the group for which its disparity is minimized and recalculate the centroid of the group if necessary. Continue this procedure until there are no more reassignments. The final partition determines the best cluster configuration.

In the case of seismic data like those collected in Table 1, we are motivated toward cluster analysis by the knowledge that often collections of events are not well identified and cannot be reliably categorized into finer partitions composed of say, earthquakes, mining explosions, nuclear explosions, or chemical explosions. Yet the data of Table 1 provide a reasonable test bed for proposed clustering procedures, because we can be fairly sure of the correct partitions in advance. With the foregoing general principles in mind, we applied both hierarchical and partitioned clustering to the events in Table 1.

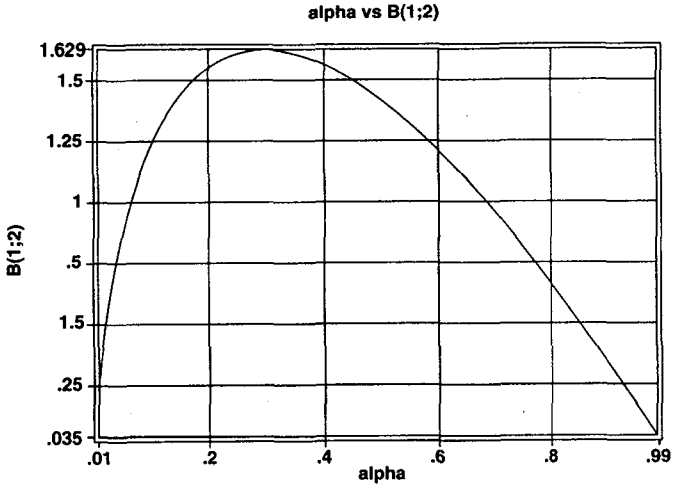

Figure 4. Plot of the Chernoff Information Measure for the Disparity of the Earthquake and Explosion Populations Showing a Maximum for α = .30.

For the Chernoff (CH) measure given by (17) or (19), one must choose a value for the parameter α. The argument that we make for this is based on choosing a value that is not too far from that implied by conventional likelihood or the minimum discrimination information principle of Kullback (1978). Note that in the remarks following (5), we mentioned that $B_\alpha(\mathbf{f}_T;\mathbf{f}_j)$ behaves like $I(\mathbf{f}_T;\mathbf{f}_j)$ for α tending to 0. Hence for small values of α, the Chernoff measure can be regarded as a correction to the Kullback–Leibler (KL) value that will be negligible for small values of α. When $\mathbf{f}_T(\lambda)$ is the periodogram, the KL measure is shown to be the same as that obtained using the Whittle approximation to the log-likelihood, as is mentioned in the next section. Thus to choose from among small values of α, we propose maximizing the disparity between group populations by searching $B_\alpha(\mathbf{f}_j;\mathbf{f}_k)$ over the parameter α. For example, the partition into eight earthquakes and eight explosions yields the function shown in Figure 4, and we note that the Chernoff information measure is maximized at $\alpha = .30$. This means that the disparity between the two groups is largest for this value, so that group differences are enhanced by this correction.

We begin by applying the hierarchical clustering procedure using the disparities (16) and (17) with $\alpha = .3$. Note that the sampling rate was 40 points per second, leading to a folding frequency of 20 Hz. The bandwidth chosen for all spectral estimators was 2 Hz and, we cut the summations off at about 8 Hz because the spectra were essentially 0 after that point. Dendrograms formed using the two procedures are shown in Figure 5, in which the earthquakes are identified with the numbers 1–8 and the known explosions appear as 9–16. The event of unknown origin at Novaya-Zemyla is 17. One can identify in both the KL and CH results, two to three clusters, with earthquakes 4 and 5 (EQ4, EQ5) clustering with the eighth explosion (EX8) in the three-cluster configuration. These events are widely separated (northern and southern Norway and Finland), so there is no apparent geographical reason for the clustering. The unknown event at Novaya-Zemyla, 17, merges with the explosion group quite early. The other two clusters are basically earthquakes excepting EQ4 and EQ5 and explosions without EX8. For the KL disparity, two clusters contain the eight earthquakes and EX8 in one cluster and the rest of the explosions plus the Novaya-Zemyla event in the other. The Chernoff disparity measure puts EQ4 and EQ5 in the explosion group.

To apply partitioned spectral clustering using (18) and (19), we considered beginning with either a random partition or with the partitions suggested by the hierarchical procedure just described. The average disparity between the sample members and their corresponding group means is presented for each configuration so that they can be compared. In Table 2 we give the method used for reassignment as either KL using (18) or CH using (19) in parentheses.

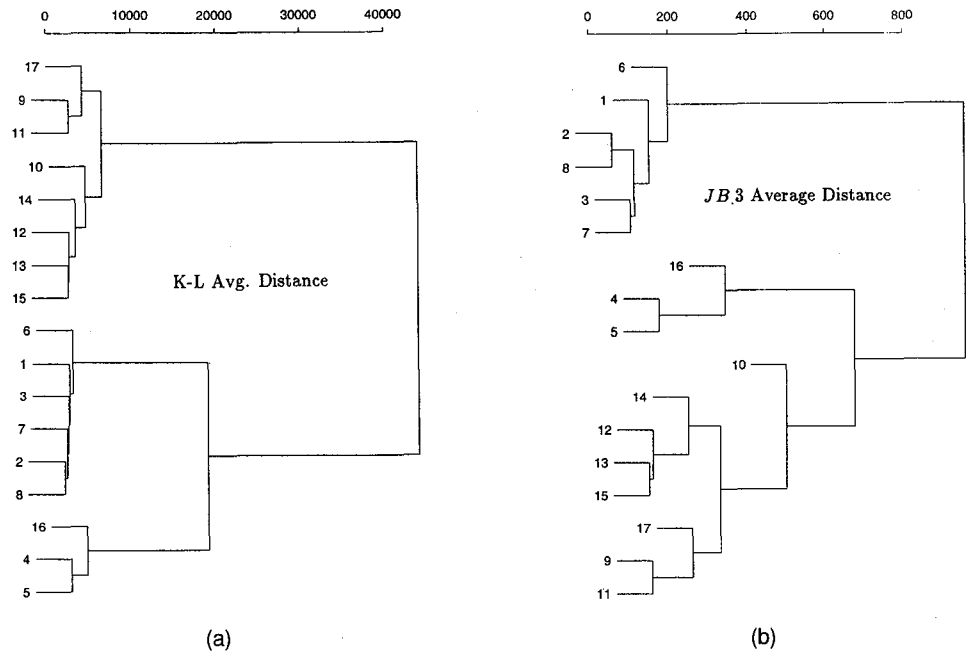

Figure 5. Results of Hierarchical Clustering Using Kullback–Leibler (a) and Chernoff (b) Disparity Measures.

In the two-group configuration, beginning with a random partition results in six reassignments on the first pass and one reassignment on the second pass. The final configuration assigns EQ4 and EQ5, along with Novaya-Zemyla, to the explosion group. Starting with either the KL- or the CH-suggested configurations gives a final partition with the earthquakes in one group and the explosions plus the Novaya-Zemyla event in the other group. Note also the smaller average disparities (177.7 and 41.5) for this starting configuration. The best three-group partition again puts EQ4, EQ5, and EX8 in the third cluster if one chooses the partition with the smallest average disparity. A competitor is the random starting configuration in combination with reassignment by the Chernoff disparity. With this procedure, EQ4 moves back into the earthquake group.

To summarize, there appears to be good natural hierarchical clustering into the two main groups—earthquakes and mining explosions—with the cluster containing the two earthquakes (events EQ4 and EQ5) and the last explosion (EQ8) appearing in the explosion group according to the KL divergence and in the earthquake group according to the average CH divergence. In the partitioned clusters, beginning with either KL or CH, the earthquakes all appear in one group and the explosions plus the Novaya-Zemyla event all appear in the other. Based on the two-group clustering results, there seems to be a reasonable basis for expecting to be able to discriminate between earthquakes and mining explosions. We explore this conjecture in the next section.

4. DISCRIMINANT ANALYSIS: APPLICATIONS

The disparity measure developed in Section 3 can also be adapted to the problem of discriminating between processes with unequal spectral matrices. Suppose that we wish to investigate the problem of classifying a realization $\mathbf{X}_T = (\mathbf{X}(1)', \mathbf{X}(2)', \ldots, \mathbf{X}(T)')'$ into one of several known categories Π_j, $j = 1, \ldots, N$, where Π_j is described by spectral density matrix $\mathbf{f}_j(\lambda)$. It is generally accepted that a reasonable approach to classification can be developed by

Table 2. Partitioned Spectral Clustering Results

Beginning configuration	*Cluster 1*	*Cluster 2*	*Cluster 3*	*K–L*	*Chernoff*
		Two Groups			
Random (KL or CH)	EQ 123678	EX 12345678 EQ 45 NZ		200.8	47.5
K-L or CH (KL or CH)	EQ 12345678	EX 12345678 NZ		177.7	41.5
		Three Groups			
Random (KL)	EQ 123678	EX 1234567 EQ 4 NZ	EX 8 EQ 5	153.5	36.2
Random (CH)	EQ 12345678	EX 1234567 NZ	EX 8 EQ 5	137.8	31.1
KL or CH (KL or CH)	EQ 123678	EX 1234567 NZ	EX 8 EQ 45	136.2	30.6

appealing to likelihood arguments using the *periodogram matrix*

$$I_T(\lambda) = \frac{1}{2\pi} T^{-1} \left\{ \sum_{t=1}^{T} \mathbf{X}(t) \exp\{-i\lambda t\} \right\} \times \left\{ \sum_{t=1}^{T} \mathbf{X}(t) \exp\{-i\lambda t\} \right\}^{*}, \quad (21)$$

where * denotes complex conjugate transpose. If the process is Gaussian, then the spectral approximation (Whittle 1954) to the log-likelihood of $\mathbf{X}_T$ is given by

$$L(\mathbf{f}_j) = -\frac{1}{2} T^{-1} \sum_{s=0}^{T-1} [\log |\mathbf{f}_j(\lambda_s)| + \operatorname{tr}\{\mathbf{I}_T(\lambda_s)\mathbf{f}_j^{-1}(\lambda_s)\}] \quad (22)$$

or its integral version,

$$L(\mathbf{f}_j) = -\frac{1}{2} \int_{-\pi}^{\pi} [\log |\mathbf{f}_j(\lambda)| + \operatorname{tr}\{\mathbf{I}_T(\lambda)\mathbf{f}_j^{-1}(\lambda)\}] \frac{d\lambda}{2\pi}, \quad (23)$$

and the likelihood-based discriminant rule is to classify $\mathbf{X}_T$ into Π_j if and only if $L(\mathbf{f}_j) > L(\mathbf{f}_k)$ for $k \neq j$ (see Alagön 1989, Dargahi-Noubary and Laycock 1981, and Shumway 1982). Even if the process is non-Gaussian, Zhang and Taniguchi (1994) adopted $L(\mathbf{f}_j)$ as the criterion and discussed its non-Gaussian robustness.

In this article we consider a more general approach based on the disparity measure (9) or its computed summation approximation (14). That is, we measure the disparity between the sample spectrum of $\mathbf{X}_T$ and population Π_j by $D_H(\mathbf{f}_T; \mathbf{f}_j)$, where we use $H_{\mathbf{I}}(\mathbf{Z})$ in (10) or $H_B(\mathbf{Z})$ in (11) depending on whether we are classifying using the KL information, leading to the approximation (18), or the CH measure, which is approximated in (19). Now $D_H(\mathbf{f}_T; \mathbf{f}_j)$ can be viewed as the disparity between the observation $\mathbf{X}_T$ and Π_j, so the proposed rule is to classify $\mathbf{X}_T$ into Π_j if and only if $D_H(\mathbf{f}_T; \mathbf{f}_k) > D_H(\mathbf{f}_T; \mathbf{f}_j)$ for $k \neq j$. For $H_{\mathbf{I}}(\mathbf{Z})$ given by (10), this is consistent with the principle of minimum discrimination information as developed by Kullback (1978), who suggested minimizing the discrimination information between the sample and the population. In the case of two populations, we assign the realization to Π_1 or Π_2 according to whether $D_H > 0$ or $D_H \leq 0$, where D_H is the discriminant function defined by

$$D_H = D_H(\mathbf{f}_T; \mathbf{f}_2) - D_H(\mathbf{f}_T; \mathbf{f}_1). \quad (24)$$

For $H(\mathbf{Z})$ defined by (10), it is clear that following the rule, with $\mathbf{f}_T$ replaced by the periodogram matrix $\mathbf{I}_T$, is equivalent to the likelihood-based rule (Yuan and Rao 1992). In this section we investigate the performance of the KL discrimination information (18) where the rule, based on (24), reduces to assigning the observation to Π_1 or Π_2 according to whether $\mathbf{I}(\mathbf{f}_T; \mathbf{f}_2) - \mathbf{I}(\mathbf{f}_T; \mathbf{f}_1) > 0$ or ≤ 0. We also look at a robust version, the CH information measure (19), as suggested by Zhang and Taniguchi (1995), who studied the univariate time series case. The rule there is similar—classify into Π_1 of Π_2 according to the value of $B_\alpha(\mathbf{f}_2; \mathbf{f}_T) - B_\alpha(\mathbf{f}_1; \mathbf{f}_T)$. Note from the remarks following (5), that $B_\alpha(\mathbf{f}_j; \mathbf{f}_T) = B_{1-\alpha}(\mathbf{f}_T \mathbf{f}_j)$, so that here we can use the rule based on (24), namely $B_\alpha(\mathbf{f}_T; \mathbf{f}_2) - B_\alpha(\mathbf{f}_T; \mathbf{f}_1)$.

It is clear how to apply the foregoing methodology given that we know the two spectral matrices $\mathbf{f}_1(\lambda)$ and $\mathbf{f}_2(\lambda)$. We follow the usual approach in discriminant analysis, taking the average of the spectra of the realizations in group j, $j = 1, 2$ and using them as a *plug-in statistics* for $\mathbf{I}(\mathbf{f}_T; \bar{\mathbf{f}}_j)$ and $B_\alpha(\mathbf{f}_T; \bar{\mathbf{f}}_j)$ in the foregoing expressions. To obtain a reasonable value for the classification error rates, we use the use a holdout procedure; that is, replace $\bar{\mathbf{f}}_j(\lambda)$ by

$$\bar{\mathbf{f}}_j(\lambda) = \frac{1}{n_j - 1} \sum_{l \neq k} f_{T_j}^{(l)}(\lambda) \quad (25)$$

when classifying the kth observation in the training set.

As an example, we consider applying the foregoing procedure to the sample of eight earthquakes, eight explosions, and the event of unknown origin from the Novaya-Zemyla region. In this case we have the same argument for choosing the value $\alpha = .3$ as the maximizer of the disparity between the two group means. We present the holdout scores in Table 3 for three values of α.

It seems clear that the second column, corresponding to the choice $\alpha = .3$ that maximized the CH information, produces discriminant scores whose distributional properties behave nicely in both groups. They are symmetrically distributed on either side of the decision point 0 with means .58 and −.48 for the earthquake and explosion groups; the standard deviations for the two groups are .34 and .20. There are no misclassifications with this method. The KL approach generates discriminant values with different standard deviations (9.34 and 1.25) and there are two misclassified explosions. Although the learning populations are extremely small, the tendency of the CH measure to improve the distributional characteristics of the discriminant scores is clear. The classification results are summarized in Table 4; we see that there are no misclassified events using the CH measure with $\alpha = .3$. Very rough predicted error probabilities (Type I + Type II) were computed using the normal approximation and the sample means and variances of the hold-out discriminant scores.

Table 3. Discriminant Scores in Multivariate Spectral Discriminant Analysis

Event	KL	CH(α = .3)	CH(α = .5)	CH(α = .7)
EQ1	8.51	.54	.41	.23
EQ2	.81	.50	.40	.23
EQ3	30.80	1.04	.85	.55
EQ4	2.73	.10	−.02	−.12
EQ5	7.69	.11	−.12	−.35
EQ6	21.50	.79	.65	.41
EQ7	20.31	.85	.68	.42
EQ8	15.65	.70	.60	.40
EX1	.29	−.25	−.33	−.45
EX2	−2.55	−.75	−1.20	−1.59
EX3	−1.82	−.51	−.77	−.91
EX4	−1.89	−.55	−.87	−1.10
EX5	−1.16	−.45	−.77	−1.04
EX6	−2.12	−.61	−1.03	−1.23
EX7	−2.10	−.59	−.91	−1.11
EX8	.93	−.21	−.50	−.73

Table 4. Signal Discrimination and Feature Extraction Approaches

Criterion	Total error probability	Holdout errors
Kullback–Leibler	.21	2
Chernoff (α = .3)	.05	0
Chernoff (α = .5)	.10	3
Chernoff (α = .7)	.24	4
Log amplitudes	.08	1
Log spectra (2–5 Hz)	.20	2
Log spectra (5–10 Hz)	.50	4
Log spectra (10–20 Hz)	.54	4
Log amplitudes and spectra (2–5 Hz)	.06	1

It is natural to ask whether the feature extraction approach is competitive in this particular situation. In Table 4 we summarize the results of a very limited study of the features mentioned at the beginning of this article. Ordinary linear discriminant analysis was performed on the logarithms of the P and S amplitudes, as well as on the integrated log spectra in three frequency bands. Seismologists look at potential discriminators in pairs, composed of P and S components, and we followed this convention by adding first the most significant bivariate log amplitudes and then the next most significant bivariate log spectra in the 2–5 Hz range. Significance was determined by the ordinary equality of bivariate equal means test with vector variables already added as covariates. The two best discriminators were the logarithms of P and S amplitudes combined with spectra; the combination of these four variables had one misclassified observation. A very rough summary of the estimated overall error (Type I + Type II) is given for all methods in Table 4. We used the sample Mahalanobis distance for the features and the sample means and variances of the holdout observations in the multivariate spectral approach for estimating the Type I and Type II errors. Other methods such as the bootstrap, jackknife, and cross-validation have been discussed by McLachlan (1992). Note that the Chernoff measure has the best performance with no holdout errors and an overall estimated error rate of only .05.

The Novaya-Zemlya event, NZ, was also classified using the average spectral matrices of the eight earthquakes and explosions and gave the values $-.49$ for KL and $-.31$ ($\alpha = .3$) for CH, putting it clearly in the explosion group. Although the NZ event falls within the explosion group, it is close enough to the boundary to create some doubt. The Russians have asserted that no mine blasting or nuclear testing occurred in the area, so the event remains as somewhat of a mystery. The fact that it was relatively far removed geographically from the test set may have introduced biases into the procedure.

5. DISCRIMINANT ANALYSIS: THEORY

In this section we examine the asymptotic properties of discriminant functions that have the general form (24), where $\mathbf{f}_T$ is the spectral density estimator based on the observation to be classified and $\mathbf{f}_j$ is the hypothetical spectral of the category Π_j that is assumed to be an m-variate linear process of the form $\mathbf{X}(t) = \sum_{s=0}^{\infty} \mathbf{G}^{(j)}(s)\mathbf{U}(t-s)$, where $\mathbf{G}^{(j)}(s)$'s are $m \times m$ matrices satisfying the usual regularity conditions (see, e.g., Hannan 1970) and $\mathbf{U}(t)$'s are iid $m \times 1$ vectors with zero means, positive definite covariance matrices $E[\mathbf{U}(t)\mathbf{U}(t)'] = \Omega$, and finite fourth-order cumulants κ_{abcd} for $a, b, c, d = 1, \ldots, m$. Note that the spectral density matrix implied by this representation is $\mathbf{f}_j(\lambda) = (2\pi)^{-1}A_j(\lambda)\Omega A_j^*(\lambda)$, where $A_j(\lambda) = \sum_{s=0}^{\infty} \mathbf{G}^{(j)}(s)\exp\{i\lambda s\}$. For theoretical arguments, we use the integral form (9) and the kernel estimator defined as

$$\mathbf{f}_T(\lambda) = \int_{-\pi}^{\pi} W_T(\lambda - \mu(\mathbf{I}_T(\mu)\, d\mu. \tag{26}$$

The use of $\mathbf{f}_T(\lambda)$ instead of the periodogram $\mathbf{I}_T(\lambda)$ is essential, because $D_H(\mathbf{I}_T; \mathbf{g})$ will not converge in probability to $D_H(\mathbf{f}_j; \mathbf{g})$ under Π_j. The periodogram is an asymptotically unbiased but inconsistent estimator for the spectral density. An exception is the case of the KL discriminant function based on $H_{\mathbf{I}}(\mathbf{Z})$, where the criterion is a linear function of $\mathbf{I}_T(\lambda)$. Here we do not give the definition of the weight function $W_T(\cdot)$, because this is independent of our theoretical results. (For the required regularity conditions, see Hannan 1970 and Taniguchi, Puri, and Kondo 1996.)

We consider two questions relating to the performance of the discriminant function (24). First, we derive the asymptotic distribution of the general discriminant score and expressions for the error rates

$$P_{D_H}(2|1) = \Pr(D_H \le 0|\Pi_1) \tag{27}$$

and

$$P_{D_H}(1|2) = \Pr(D_H > 0|\Pi_2). \tag{28}$$

Then, assuming a contiguous alternative for the spectral matrix, we give a condition for non-Gaussian robustness. Robustness in this context means that the non-Gaussianity of the innovation $\mathbf{U}(t)$ has no effect on the asymptotic error rates (see, e.g., Zhang and Taniguchi 1994). The results in this section may also be extended to the plug-in version of D_H, obtained by replacing $\mathbf{f}_1$ and $\mathbf{f}_2$ with the estimators $\bar{\mathbf{f}}_1$ and $\bar{\mathbf{f}}_2$ based on the training sample (Kakizawa 1996b).

We make the following assumption:

A1. $\mathbf{H}(\mathbf{Z})$, in Section 2, is a scalar holomorphic function of a matrix $\mathbf{Z}$ and has a unique minimum zero at $\mathbf{Z} = \mathbf{I}_m$, the $m \times m$ identity matrix.

To establish the asymptotic normality of D_H, we need an additional condition:

A2. The $m \times m$ matrix $\mathbf{Q}_{j,k}(\lambda) = \mathbf{f}_k(\lambda)^{-1}[\mathbf{H}^{(1)}(\mathbf{f}_j(\lambda)\mathbf{f}_k(\lambda)^{-1})]'$ satisfies $\mathbf{Q}_{j,k}(\lambda)^* = \mathbf{Q}_{j,k}(\lambda)$ and $\mathbf{Q}_{j,k}(-\lambda) = \mathbf{Q}_{j,k}(\lambda)'$, where $H^{(1)}(\cdot)$ is the first derivative of $H(\mathbf{Z})$ at $\mathbf{Z} = \cdot$ whose (a,b)th element is $\partial H(\mathbf{Z})/\partial \mathbf{Z}_{ab}$. This is the condition for using lemma A.3.3 of Hosoya and Taniguchi (1982). A1 implies that $H^{(1)}(\mathbf{I}_m)$ is a zero matrix and that the Hessian matrix of $H(\mathbf{Z})$ at $\mathbf{Z} = \mathbf{I}_m$ is positive definite.

In what follows, set $(j,k) = (1,2)$ or $(2,1)$.

Theorem 1 (Kakizawa 1996b). Suppose that $\mathbf{f}_1(\lambda) \neq \mathbf{f}_2(\lambda)$ on a set of positive Lebesgue measure. Then under

Π_j, $D_H \xrightarrow{p} (-1)^{j+1} D_H(\mathbf{f}_j; \mathbf{f}_k)$ and

$$\sqrt{T}\{D_H + (-1)^j D_H(\mathbf{f}_j; \mathbf{f}_k)\} \xrightarrow{d} \mathrm{N}(0, V_H^2(j,k)), \quad (29)$$

where $D_H(\mathbf{f}_j; \mathbf{f}_k)$ is the integral disparity (9) and

$$V_H^2(j,k) = \frac{1}{2}\int_{-\pi}^{\pi} \mathrm{tr}[\mathbf{Q}_{j,k}\mathbf{f}_j]^2 \frac{d\lambda}{2\pi} + \frac{1}{16\pi^2}\sum_{a,b,c,d} \kappa_{abcd}\gamma_{ab}(j,k)\gamma_{cd}(j,k), \quad (30)$$

with the $m \times m$ matrix $\mathbf{\Gamma}_H(j,k) = \{\gamma_{ab}(j,k)\}$ defined by

$$\mathbf{\Gamma}_H(j,k) = \int_{-\pi}^{\pi} A_j^* \mathbf{Q}_{j,k} A_j \frac{d\lambda}{2\pi}. \quad (31)$$

In view of Theorem 1, the limiting forms of the error rates (27) and (28) are $\lim_{T\to\infty} P_{D_H}(k|j) = 0$ for $(j,k) = (1,2),(2,1)$. This shows that the discriminant D_H is consistent. From (30), one might also approximate them as the normal integrals

$$P_{D_H}(j|k) \approx \Phi\left(-\sqrt{T}\,\frac{D_H(\mathbf{f}_j; \mathbf{f}_k)}{V_H(j,k)}\right), \quad (32)$$

which depend on the fourth-order cumulants unless (31) is a zero matrix. To look at robustness, we assume that the hypothetical m-variate linear process is generated as

$$\mathbf{X}(t) = \sum_{s=0}^{\infty} \mathbf{G}(s)\mathbf{U}(t-s) \quad (33)$$

under Π_1 and as

$$\mathbf{X}(t) = \sum_{s=0}^{\infty} \{\mathbf{G}(s) + T^{-1/2}\mathbf{H}(s)\}\mathbf{U}(t-s) \quad (34)$$

under Π_2. That is, the spectral densities associated with Π_1 and Π_2 are

$$\mathbf{f}_1(\lambda) = (2\pi)^{-1}\mathbf{A}(\lambda)\mathbf{\Omega}\mathbf{A}^*(\lambda) \quad (35)$$

and

$$\mathbf{f}_2(\lambda) = (2\pi)^{-1} \times \{\mathbf{A}(\lambda) + T^{-1/2}\mathbf{B}(\lambda)\}\mathbf{\Omega}\{\mathbf{A}(\lambda) + T^{-1/2}\mathbf{B}(\lambda)\}^*, \quad (36)$$

with $\mathbf{A}(\lambda) = \sum_{s=0}^{\infty} \mathbf{G}(s)\exp\{i\lambda s\}$ and $\mathbf{B}(\lambda) = \sum_{s=0}^{\infty} \mathbf{H}(s)\exp\{i\lambda s\}$. The hypothesis (36) is different from Kakizawa's (1996b), but it includes the parametric situation such as his corollary 1. As in Kakizawa's paper, the quantities $D_H(\mathbf{f}_j; \mathbf{f}_k)$ and $V_H^2(j,k)$ [see (9) and (30)] are determined by the local property of the function $H(\mathbf{Z})$ at $\mathbf{Z} = \mathbf{I}_p$. By assumption A1, $H(\mathbf{I}_m) = 0$ and $\mathbf{H}^{(1)}(\mathbf{I}_m)$ is a zero matrix. We assume further that:

A3. The $m^2 \times m^2$ Hessian matrix of $\mathbf{H}(\mathbf{Z})$ at $\mathbf{Z} = \mathbf{I}_m$ is $c\mathbf{K}_m$, where $\mathbf{K}_m$ is the commutation matrix (Magnus and Neudecker 1988) and $c > 0$.

Note that $\mathbf{H_I}$, $\mathbf{H}_B$, $\mathbf{H_Q}$, and their symmetric versions $\tilde{\mathbf{H}}$ in (13) satisfy assumptions A1–A3. Then we have the following result, the proof of which is given in the Appendix.

Theorem 2. Let $\mathbf{f}_1$ and $\mathbf{f}_2$, defined by (35) and (36), be the hypothetical spectral density matrices of m-variate linear processes (33) and (34), respectively. If $\mathbf{H}(\mathbf{0})$ is an $m \times m$ zero matrix, the asymptotic error rates are independent of the non-Gaussianity of the process.

6. DISCUSSION

It is interesting that the focus of clustering and classification shifts when one moves from the realm of classical multivariate analysis to multivariate time series analysis. Although there are some instances in which the model that assumes a common mean series over replicates within a group makes sense, it is more often the case that phase shifts between replicates make such analyses meaningless. It these cases the emphasis of the methodology shifts from mean differences to differences in the covariance structure or, equivalently, to difference between the spectral matrices. When the data are stationary, or can be made so by transforming or detrending, it is natural to utilize the periodic structure of multivariate time series as the predominant setting in which to characterize similarities and differences. This makes the spectral matrix the focus of the analysis; one also achieves a great reduction in dimensionality, because matrix computations in time are reduced by Fourier transformation. The increase in computing speed and new sophisticated data collection methods mean that multivariate time series are much more common, and we believe that developing methods for analyzing such series will be of increasing importance.

With motivation provided by the foregoing remarks, we have extended information-theoretic techniques based on the spectrum to the multivariate case by defining a general measure of disparity between two multivariate time series. The Kullback–Leibler and Chernoff information measures were special cases that turned out to be especially useful for clustering and discriminating between seismic records generated by earthquakes and mining explosions. In general, both the Kullback–Leibler and Chernoff measures were useful for clustering similar events in these two populations. The time series versions of Kullback–Liebler discrimination information and the Chernoff information also were used effectively to discriminate between explosions and earthquakes. Theoretical performance of the general disparity criterion was couched in terms of the misclassification error probabilities that were shown to be computable from the limiting normal distribution of the disparity between the sample and theoretical spectrum. A general robustness condition for departures from normality was derived by specifying that the limiting error rate not depend on higher order moments.

It is interesting that the Chernoff measure behaves more and more like the Kullback–Leibler information as the parameter α gets smaller. In this sense, the behavior of α exerts a *regularizing influence* on the likelihood or Kullback–Leibler approaches in much the same way as classical regularized discriminant analysis in the sense of Friedman (1989), where the covariance mixing parameters are varied to achieve a balance between linear and quadratic discriminant analysis. A good discussion of regularization methods

and the estimation of the weights using cross validation has been provided by McLachlan (1992, pp. 144–151).

It should be noted that although we have concentrated on nonparametric spectral disparities for clustering and classification, there can be a parallel approach based on parametric models for the spectrum. In particular, autoregressive spectral estimators can be used following the approach of Dargahi-Noubary (1995), who used standard explosion source theory to motivate these models. Parametric modeling of earthquake and explosion spectra has not been common in the seismological literature, and we note the increased difficulty due to nonlinear estimation required when the spectral matrix $\mathbf{f}(\lambda) = \mathbf{f}_{\boldsymbol{\theta}}(\lambda)$ is a function of a $q \times 1$ parameter vector $\boldsymbol{\theta}$. Kakizawa (1996a,b, 1997) provided some asymptotic results for this case.

APPENDIX: PROOF OF THEOREM 2

Let $\mathbf{f}_1$ and $\mathbf{f}_2$ be defined as in (35) and (36). For the $m \times m$ Hermitian matrix

$$\boldsymbol{\Delta}(\lambda) = \boldsymbol{\Omega}^{-1/2}\mathbf{A}(\lambda)^{-1}\mathbf{B}(\lambda)\boldsymbol{\Omega}^{1/2} + \boldsymbol{\Omega}^{1/2}\mathbf{B}(\lambda)\mathbf{A}(\lambda)^{*-1}\boldsymbol{\Omega}^{-1/2},$$

we set

$$V = \frac{1}{2}\int_{-\pi}^{\pi} \operatorname{tr}[\boldsymbol{\Delta}(\lambda)]^2 \, \frac{d\lambda}{2\pi}.$$

After some algebra, the quantities $D_H(\mathbf{f}_j; \mathbf{f}_k)$ and $V_H^2(j,k)$ [see (9) and (30)] reduce to

$$D_H(\mathbf{f}_j; \mathbf{f}_k) = \frac{cV}{2T} + o(T^{-1})$$

and

$$V_H^2(j,k) = \frac{c^2}{T}\left(V + \frac{1}{4}\sum \kappa_{abcd}\gamma_{ab}\gamma_{cd}\right) + o(T^{-1}),$$

where the $m \times m$ matrix $\boldsymbol{\Gamma} = \{\gamma_{ab}\}$ is defined by

$$\boldsymbol{\Gamma} = \int_{-\pi}^{\pi} \boldsymbol{\Omega}^{-1/2}\,\boldsymbol{\Delta}(\lambda)\,\boldsymbol{\Omega}^{-1/2}\,\frac{d\lambda}{2\pi}.$$

Substituting them into (32), the asymptotic error rates are given by $\Phi(-\sqrt{V}/2)$ if $\boldsymbol{\Gamma} = 0$. But using $\mathbf{A}(\lambda)^{-1} = \sum_{s=0}^{\infty}\mathbf{C}(s)e^{i\lambda s}$ with $\mathbf{C}(0) = \mathbf{G}(0)^{-1}$, we have

$$\boldsymbol{\Gamma} = \Omega^{-1}\mathbf{G}(0)^{-1}H(0) + H(0)'\mathbf{G}(0)'^{-1}\Omega^{-1}.$$

Thus $\boldsymbol{\Gamma} = 0$ is equivalent to $H(0) = 0$, because we can show

$$\operatorname{vec}\boldsymbol{\Gamma} = 2\{\mathbf{I}_p \otimes \boldsymbol{\Omega}^{-1}\mathbf{G}(0)^{-1}\}\operatorname{vec}\{\mathbf{H}(0)\},$$

and the result is proved.

[Received May 1996. Revised June 1997.]

REFERENCES

Alagón, J. (1989), "Spectral Discrimination for Two Groups of Time Series," *Journal of Time Series Analysis*, 10, 203–214.

Bennett, T. J., and Murphy, J. R. (1986), "Analysis of Seismic Discrimination Capabilities Using Regional Data From Western U.S. Events," *Bulletin of the Seismological Society of America*, 76, 1069–1086.

Bhattacharya, A. (1943), "On a Measure of Divergence Between Two Statistical Populations," *Bulletin of the Calcutta Mathematical Society*, 35, 99–109.

Blandford, R. R. (1993), "Discrimination of Earthquakes and Explosions at Regional Distances Using Complexity," AFTAC-TR- 93-044 HQ, Air Force Technical Applications Center, Patrick Air Force Base, FL.

Blashford, R. K., Alenderfer, M. S., and Morey, L. C. (1982), "Cluster Analysis Software," in *Handbook of Statistics*, Vol. 2, eds. P. R. Krishnaiah and L. N. Kanal, New York: North-Holland.

Booker, A., and Mitronovas, W. (1964), "An Application of Statistical Discrimination to Classify Seismic Events," *Bulletin of the Seismological Society of America*, 54, 951–971.

Chernoff, H. (1952), "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of the Observations," *Annals of Mathematical Statistics*, 25, 573–578.

Dargahi-Noubary, G. R. (1995), "Stochastic Modeling and Identification of Seismic Records Based on Established Deterministic Formulations," *Journal of Time Series Analysis*, 16, 201–219.

Dargahi-Noubary, G. R., and Laycock, P. J. (1981), "Spectral Ratio Discriminants and Information Theory," *Journal of Time Series Analysis*, 2, 71–86.

Dysart, P., and Pulli, J. J. (1990), "Regional Seismic Event Classification at the NORESS Array; Seismological Measurements and the Use of Trained Neural Networks," *Bulletin of the Seismological Society of America*, 80, 1910–1933.

Friedman, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, 165–175.

Gersch, W., Martinelli, F., Yonemoto, J., Low, M. D., and MacEwan, J. A. (1979), "Automatic Classification of Electroencephalograms by K–L Nearest Neighbor Rules," *Science*, 205, 193–195.

Hannan, E. J. (1970), *Multiple Time Series*, New York: Wiley.

Hosoya, Y., and Taniguchi, M. (1982), "A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes," *Annals of Statistics*, 10, 132–153; Corr. (1993) 21, 1115–1117.

Johnson, R. A., and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*. (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Kakizawa, Y. (1996a), "Parameter Estimation and Hypothesis Testing in Stationary Vector Time Series," *Statistics and Probability Letters*, 33, 225–234.

——— (1996b), "Discriminant Analysis for non-Gaussian Vector Stationary Processes," *Journal of Nonparametric Statistics*, 7, 187–203.

——— (1997), "Higher Order Asymptotic Theory for Discriminant Analysis in Gaussian Stationary Processes," *Journal of the Japan Statistical Society*, 27, 19–35.

Kazakos, D., and Papantoni-Kazakos, P. (1980), "Spectral Distance Measuring Between Gaussian Processes," *IEEE Transactions on Automatic Control*, AC-25, 950–959.

Kullback, S. (1978), *Information Theory and Statistics*, Gloucester, MA: Peter Smith.

Kullback S., and Leibler, R. A. (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

MacQueen, J. B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceeding of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, Berkeley, Calif.: University of California Press, pp. 181–297.

Magnus, J. R., and Neudecker, H. (1988), *Matrix Differential Calculus With Applications in Statistics and Econometrics*, Chichester, U.K.: Wiley.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

Parzen, E. (1990), "Time Series, Statistics and Information," IMA Preprint Series 663, Institute for Mathematics and Its Applications, University of Minnesota.

Pinsker, M. S. (1964), *Information and Information Stability of Random Variables and Processes*, San Francisco: Holden-Day.

Pulli, J. J. (1995), "Extracting and processing signal parameters for regional seismic event identification," in *Monitoring a Comprehensive Test Ban Treaty*, eds. E. S. Husebye and A. M. Dainty, Doordrecht: Kluwer, pp. 791–803.

Renyi, A. (1961), "On Measures of Entropy and Information," in *Proceeding of 4th Berkeley Symposium on Mathematical Statistics and Probability*, 1, Berkeley: University of California Press, pp. 547–561.

Richards, P. G., Kim, W-Y., and Ekström, G. (1993), "RMS Lg Studies of Underground Nuclear Explosions in the U.S.S.R. and the U.S.," PL-TR-93-2227, Phillips Laboratory, Directorate of Geophysics, Air Force Materiel Command, Hanscom Air Force Base, MA. ADA281016.

Ryall, A. S., Jr., Baumgardt, D. R., Fisk, M. D., and Riviere-Barbier, F. (1996), "Resolving Regional Discrimination Problems: Some Case His-

tories," in *Monitoring a Comprehensive Test Ban Tready*, eds. E. S. Husebye and A. M. Dainty, Doordrecht: Kluwer, pp. 721–741.

Shumway, R. H. (1982), "Discriminant Analysis for Time Series," in *Handbook of Statistics*, Vol. 2, eds. P. R. Krishnaiah and L. N. Kanal, New York: North-Holland.

——— (1996), "Statistical Approaches to Seismic Discrimination," in *Monitoring a Comprehensive Test Ban Treaty*, eds. E. S. Husebye and A. M. Dainty, Doordrecht: Kluwer, pp. 791–803.

Shumway, R. H., and Unger, A. N. (1974), "Linear Discriminant Functions for Stationary Time Series," *Journal of the American Statistical Association*, 69, 948–956.

Taniguchi, M., Puri, M. L., and Kondo, M. (1996), "Nonparametric Approach for Non-Gaussian Vector Stationary Processes," *Journal of Multivariate Analysis*, 56, 259–283.

Taylor, S. R., Denny, M. R., Vergino, E. S., and Glaser, R. E. (1989), "Regional Discrimination Between NTS Explosions and Western U.S. Earthquakes," *Bulletin of the Seismological Society of America*, 79, 1142–1176.

Whittle, P. (1954), "Estimation and Information in Stationary Time Series," *Arkiv för Matematik*, 2, 423–434.

Yuan, J., and Rao, T. S. (1992), "Classification of Textures Using Second-Order Spectra," *Journal of Time Series Analysis*, 4, 91–101.

Zhang, G., and Taniguchi, M. (1994), "Discriminant Analysis for Stationary Vector Series," *Journal of Time Series Analysis*, 15, 117–126.

——— (1995), "Nonparametric Approach for Discriminant Analysis in Time Series," *Journal of Nonparametric Statistics*, 5, 91–101.