Chapter 2. STATISTICAL CLASSIFICATION APPROACH TO DISCRIMINATION BETWEEN WEAK EARTHQUAKES AND EXPLOSIONS

1. INTRODUCTION

Discrimination between weak regional or local earthquakes and explosions is a serious challenge for state-of-the-art monitoring of the Comprehensive Test Ban Treaty (CTBT). In particular, it is known that discrimination features effective in one seismic region are often useless in another region. This is a serious obstacle to refinement and standardization of source discrimination techniques. Another important unsolved task is consistent estimation of the probability of event misclassification inherent to the seismic region being monitored.

Automation of procedures for weak local explosion and earthquake discrimination is important in regions where commercial mining operations and quarry blasting generate a large number of seismic recordings on a daily basis, and numerous events have to be processed in a near-real-time mode. An automated or semi-automated computer discrimination technique should contain procedures for selection of the discrimination features which are most informative and can be extracted from seismograms automatically with minimum intervention by an analyst.

In recent years success has been achieved in techniques for discriminating between earthquake and explosion sources by artificial neural networks (Dowla, 1995; Dyssart and Pulli, 1990). Nonetheless, the potential of statistical approaches to discrimination has not been exhausted yet (Anderson et al., 1995; Kakizawa et al., 1997; Shumway, 1995). In this study we follow the conventional feature extraction—feature discrimination approach of testing statistical hypotheses formulated for probability distributions of discrimination features extracted from seismograms. The discrimination problem is solved by processing feature vector sets collected from learning events and events to be classified with the help of classical statistical pattern recognition procedures.

The features used in this discrimination technique are choosen as a rule by heuristic considerations. Often spectral characteristics of various seismogram phases that are typically different for earthquakes and explosions serve as discrimination features (Gupta, 1995; Kim et al., 1995). Numerous investigations in discrimination analysis (Deev, 1970; Levin et al., 1970; Raudis, 1975; Tsvang et al., 1993) demonstrate that selection of a small number of the most informative features is extremely useful in this approach. It has been theoretically and experimentally proved that a few carefully selected features may provide a smaller error classification probability than the set of all features. This is the so-called “pick-effect” or “multivariate effect”.

The optimal feature subset selected using data from one region often do not coincide with the equivalent subset for another region. To overcome the difficulties of “manually” selecting the best features we have developed an automatic feature selection procedure which chooses the optimal feature subset specific for a region of interest.

Statistical methods for selecting optimal earthquake—explosion discrimination features and for consistent estimation of misclassification probability for events from a given region are not widely used in seismic monitoring practice. The authors know of only a few papers (e.g., Tsvang et.al., 1993, Pinsky et al., 1997) in which these methods are applied to discrimination of regional and teleseismic explosions and earthquakes. In this study we implemented these statistical methods for discrimination between weak local earthquakes and chemical explosions recorded by the Israel Seismic Network (ISN) (Figure 1) operated by Geophysical Institute of Israel (GII). The data base with ground truth information is described in [Gitterman and van Eck (1993)] and [Pinsky et al. (1997)]. The list of event parameters is given in Table 1.

Discrimination features based on relative power spectral distributions of P and S phases were extracted from the event waveforms and processed by automated computer procedures to: a) select the most informative features providing the minimum probability of discrimination errors; b) estimate the probability of misclassifications in a consistent manner.

Power spectra of seismic noise were estimated for intervals preceding the event waveforms and were subtracted to improve the quality of discrimination feature measurements under conditions of poor signal-to-noise ratio typical for weak events. The feature selection procedure allowed us to extract automatically the 3–8 most informative features from more than 20 deemed to be relevant from heuristic considerations. The feature selection procedures featured nonlinear transformation and quadratic discrimination; after subtracting noise, we achieved an average misclassification probability, estimated with the help of a statistically consistent cross-validation method, of 3.8%, i.e., only 2 events (explosions) were incorrectly classified out of 53 events.

An automated computer tool was developed to implement the proposed seismogram discrimination technique. It was designed with the help of the Seismic Network Data Analysis (SNDA) System, a problem-oriented programming shell developed at the IRIS Moscow Data Center/SYNAPSE Science Center.

2. OVERVIEW OF THEORETICAL METHODS FOR STATISTICAL FEATURE SELECTION FOR CLASSIFICATION AND ERROR PROBABILITY ESTIMATION

2.1. Statistical approach to the classification problem

There is an extensive bibliography on statistical methods of discriminant analysis. Reviews of the problem may be found in Gupta (1973) and Ivazian (1989). The theoretical background is contained in Anderson (1984) and Johnson and Wichern (1992), and an application to the seismic discrimination problem in Fisk et al. (1993)(. Below we give a very brief sketch of a statistical approach to discriminant analysis.

Denote $\boldsymbol{x} \in R^p$ (R^p is a p-dimensional space) a vector of discrimination features, considered to be a random vector having different (but unknown) distributions P_j, $j=1,2,$ in R^p for earthquakes and explosions. Assume that the set of all possible distributions P_j can parametrized by a k-dimensional parameter $\boldsymbol{\theta}$ and consider the parametric set $\{f(\boldsymbol{x};\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta \subset R_k\}$ of probability density functions (PDFs) of Pj. Two PDFs $f(\boldsymbol{x};\boldsymbol{\theta}_1)$ and $f(\boldsymbol{x};\boldsymbol{\theta}_2)$ with different (but unknown) parameters $\boldsymbol{\theta}_1$,$\boldsymbol{\theta}_2$ represent probabilistic models for two classes of feature vectors related to earthquakes and explosions.

The learning sets of feature vectors $\boldsymbol{X}_{n1}=\{\boldsymbol{x}_1(1),\boldsymbol{x}_2(1),\dots,\boldsymbol{x}_{n1}(1)\}$ and $\boldsymbol{X}_{n2}=\{\boldsymbol{x}_1(2),\ \boldsymbol{x}_2(2),\dots,\boldsymbol{x}_{n2}(2)\}$ collected from earthquakes and explosions are regarded as samples of independent random vectors with PDFs $f(\boldsymbol{x};\boldsymbol{\theta}_1)$ and $f(\boldsymbol{x};\boldsymbol{\theta}_2)$ respectively. The feature vector $\boldsymbol{x}$ which has to be classified (i.e., identified as earthquake or explosion) is regarded as a random vector independent of the learning samples $\boldsymbol{X}_{n1}$ and $\boldsymbol{X}_{n2}$ and having a PDF $f(\boldsymbol{x};\boldsymbol{\theta}_0)$, where the parameter $\boldsymbol{\theta}_0$ is unknown but can only be equal to $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$. The discrimination problem is then interpreted in the terminology of theoretical statistics as the testing of composite statistical hypotheses on the basis of the observations $\boldsymbol{X} = \{\boldsymbol{x},\boldsymbol{X}_{n1},\boldsymbol{X}_{n2}\}$. The hypotheses are: H_1: $\boldsymbol{\theta}_0=\boldsymbol{\theta}_1$; H_2: $\boldsymbol{\theta}_0=\boldsymbol{\theta}_2$. This most general so-called “three-samples” statistical interpretation of the classification problem was suggested by Rao (1954).

There are several statistical strategies for finding the best decision rule for testing hypotheses H_1, H_2 , such as (a) Bayesian, (b) Maximum Likelihood ratio, (c) adaptive, and additional methods (Anderson, 1984; Ivazian et al., 1989; Troitsky, 1986a). The first two strategies are briefly discussed in the Appendix, where it is shown that if the numbers $n1$ and $n2$ of learning observations are large enough these strategies lead to discrimination algorithms which are close to those derived by the adaptive approach.

The adaptive (“plug-in”) approach to the design of discrimination rules has historically received the greatest development. Adaptive rules based on the likelihood ratio (LR) are most commonly used. For such rules the LR $L(\boldsymbol{x})=ln[f(\boldsymbol{x};\boldsymbol{\theta}_2)/f(\boldsymbol{x};\boldsymbol{\theta}_1)]$ is first calculated under the assumption that the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of the distributions of the feature vector $\boldsymbol{x}$ under hypotheses H_1 and H_2 are known. Then consistent estimates $\Theta_1(\boldsymbol{X}_{n1})$, $\Theta_2(\boldsymbol{X}_{n2})$ of these (unknown) parameters are calculated using the learning observations $\boldsymbol{X}_{n1}$,$\boldsymbol{X}_{n2}$. The adaptive rule consists of comparing the statistic $L(\boldsymbol{x},\boldsymbol{X}_{n1},\boldsymbol{X}_{n2})=ln[f(\boldsymbol{x};\Theta_2)/f(\boldsymbol{x};\Theta_1)]$ with the zero threshold.

For Gaussian distributions of the feature vectors with a common covariance matrix $\boldsymbol{S}$ and different mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$: $f(\boldsymbol{x};\boldsymbol{\theta}_l)=\aleph\{\boldsymbol{\mu}_l,\boldsymbol{S}\}$, $l=1,2,$ (where $\aleph$ is the symbol of Gaussian (Normal) distribution) the adaptive approach leads to the result that the Linear Discriminant Function (LDF) is compared with the zero threshold,

$$LDF = [\,\boldsymbol{x} - \frac{1}{2}(\,\overline{\boldsymbol{x}}^{(2)} + \overline{\boldsymbol{x}}^{(1)}\,)]^T\,\boldsymbol{S}_{n1+n2}^{-1}(\,\overline{\boldsymbol{x}}^{(2)} - \overline{\boldsymbol{x}}^{(1)}\,)\,. \tag{1}$$

where $\overline{\boldsymbol{x}}^{(1)}$, $\overline{\boldsymbol{x}}^{(2)}$ are sample mean vectors, $\boldsymbol{S}_{n1+n2}^{-1}$ is an unbiased estimate of the common inverse covariance matrix $\boldsymbol{S}^{-1}$ calculated using the total learning vector set.

For Gaussian distributions of the feature vectors with different covariance matrices $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ and mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $f(\boldsymbol{x};\boldsymbol{\theta}_l)=\boldsymbol{\aleph}\{\boldsymbol{\mu}_l,\boldsymbol{S}_l\}$, $l=1,2$, discrimination is based on comparison of the Quadratic Discriminant Function (QDF) with the zero threshold,

$$QDF = (\boldsymbol{x}-\bar{\boldsymbol{x}}^{(1)})^T \boldsymbol{S}_{n1}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(1)})-(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(2)})^T \boldsymbol{S}_{n2}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(2)})+ln\frac{|\boldsymbol{S}_{n1}|}{|\boldsymbol{S}_{n2}|} \qquad (2)$$

It is shown in the Appendix that in the case of multivariate Gaussian distributions $\boldsymbol{\aleph}\{\boldsymbol{\mu}_l,\boldsymbol{S}_l\}$, $l=1,2$, with different covariance matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, discrimination based on the QDF (2) is asymptotically equivalent to discrimination rules designed using the Bayesian and Maximum Likelihood approaches, if the numbers n_1, n_2 of learning observations tend to infinity (Troitsky, 1986b). In the case of Gaussian distributions with a common covariance matrix $\boldsymbol{S}_1 = \boldsymbol{S}_2 = \boldsymbol{S}$ discrimination based on the LDF (1) is asymptotically equivalent to the Maximum Likelihood discrimination rule. Thus practical application of the simple adaptive LDS and QDF statistics is reasonable from the point of view of rigorous statistical criteria if the numbers of learning samples n_1, n_2 are not small.

The assumption that the multidimensional distributions of the discrimination features are Gaussian is somewhat restrictive for practical applications of the LDF and QDF. However, this restriction can be relaxed by nonlinear transformation of the features (e.g.,, logarithmic, $y=log(x)$; or Box-Cox, $y=(1/\alpha)(x^{\alpha}-1))$ to make their distributions closer to Gaussian. In the studies described below we used discrimination rules based on the LDF and QDF.