OUTLIER TESTING WHEN SOME DATA ARE MISSING

Wayne A. Woodward*, Stephan R. Sain*, H.L. Gray*, and Mark D. Fisk**

* Southern Methodist University
** Mission Research Corporation

Sponsored by Defense Threat Reduction Agency
Contract No. DTRA01-99-C-0018

ABSTRACT

We consider the problem of multivariate outlier testing for purposes of distinguishing seismic signals of underground nuclear events from training samples based on non-nuclear seismic events when some of the data are missing. We first consider the case in which the training data follow a multivariate normal distribution. Suppose that such a training set of n observations based on k features is available but that some of the observations are missing. The approach currently used in practice is to perform the outlier testing using a generalized likelihood ratio test procedure based only on the data vectors with full data. Critical values for this test are usually obtained using Hotelling's T^2 distribution. When large amounts of data are missing, use of this strategy may lead to loss of valuable information. An alternative procedure is to incorporate all n of the data vectors using the EM algorithm to appropriately handle the missing data. In this case, the Hotelling's T^2 procedure no longer applies, and resampling methods are used to find appropriate critical regions. We use simulation results and analysis related to seismic data to compare these two strategies for dealing with missing data.

In some cases it may be best to model the non-nuclear events in the region using a mixture-of-normals model (e.g. when the events come from a variety of sources or the data are substantially non-normal). We describe an EM-algorithm based procedure for using the modified likelihood ratio test to test for outliers when the training data follow a mixture distribution and when some of the observations are missing. In this setting, we again use simulations and analysis of seismic data to compare the use of the EM algorithm on the entire data set with the use of only the complete data vectors. We also consider a new alternative to the likelihood ratio test that provides a useful reduction in the computational complexity.

Key Words: outlier testing, missing data, mixture model, EM algorithm

OBJECTIVE

The objective of this research is to develop and implement statistical methodology for outlier testing when some data are missing. In particular we examine the use of the EM algorithm in the case of missing observations and compare its use with that of only using those data vectors for which complete data are available.

RESEARCH ACCOMPLISHED

BACKGROUND

Several authors have considered the problem of outlier testing for purposes of distinguishing seismic signals of underground nuclear events from training samples based on non-nuclear seismic events such as earthquakes, mining explosions, etc. Fisk, Gray, and McCartor (1996) considered the problem of using a likelihood ratio test for detecting outliers from a multivariate normal (MVN) distribution fit to the training data. They applied the method to a variety of seismic data sets and demonstrated excellent results. Taylor and Hartse (1997) also successfully applied this procedure to seismic data from the WMQ station in western China. These papers, however, require complete or non-missing data. That is, if k features are selected for use on a training sample of n events, then these results assume that all $k \times n$ observations are available. When some of the features for some of the events are missing, then the question arises concerning the optimal method for analyzing the data in such a way as to make the best use of the available data. One common technique for dealing with this issue is to use only those events in the training sample for which all k features were observed. It should be noted that such a procedure can result in a substantial loss of data and associated information. For example, consider the case in which 25% of the data are missing at random from a training sample of size $n = 50$ events. If, for example, we want to use four features, then the expected number of events for which all four features are observed is 16. Intuitively, however, it seems that we should not "throw away" the partial data on the other 34 events. In this research we compare the use of the "complete vector" approach which uses only these 16 vectors of complete data with the use of the Expectation-Maximization (EM) algorithm (see Dempster, Laird, and Rubin, 1977) which attempts to make optimal use of all data observed. These analysis strategies are compared via detection probabilities. Another problem can occur if there are no cases or only a very few cases in which all k of the features are observed. If the "complete-vector" strategy is used, then some key features may need to be removed before analysis can be done.

In some cases it may be best to model the non-nuclear events in the region using a mixture-of-normals model (e.g. when the events come from a variety of sources or the data are substantially non-normal). In this case, the training data can be considered to be a sample of size n from a mixture distribution whose density is given by

$$f(\mathrm{x}) = \sum_{i==1}^{m} p_i g_i(\mathrm{x}; \mu_i, \Sigma_i) \tag{1}$$

where m is the number of components in the mixture, $g_i(\mathrm{x}; \mu_i, \Sigma_i)$ is the MVN density associated with the ith component, the p_i, $i = 1, \ldots, m$ are the mixing proportions, and $\mathbf{x}$ is a d-dimensional vector of feature variables. Wang, Woodward, Gray, Wiechecki, and Sain (1997) developed a modified likelihood ratio test that is applicable to the mixture and non-mixture settings and, in essence, requires no distributional assumptions concerning the outlier distribution. These authors demonstrated via simulations that the modified likelihood ratio test can be used successfully for outlier detection when m is known and at least some of the training data are labeled. Sain, Gray, Woodward, and Fisk (1999) extend these results to the case in which no data are labeled and in which the number of components in the mixture is unknown. They demonstrated their results using simulations similar to those of Wang, et al. (1997) and showed little or no loss of power when no training data are unlabeled. Sain, et al. (1999) obtained excellent results using their procedure on actual seismic data from the Vogtland region near the Czech-German border and from the WMQ station in western China. Using the China data, the authors demonstrated that a mixture model may be preferable to the use of a single multivariate model even when there are not any identifiable groups of event types represented in the training data.

However, the mixture-model results described above also make the assumption that there is no missing data (except for missing labels.) That is, the assumption is made that all k features are observed on all of the n events in the training sample. Again, because of the fact that it is very common for the seismic data

collected for events in a training sample to contain some missing data, it is important in this setting to also consider the issue of optimal handling of missing feature data. For further discussion of outlier testing in the presence of missing data see Miller, Gray, and Woodward (1993), Miller, Woodward, Gray, Fisk, and McCartor (1994), and Woodward, Sain, Gray, and Fisk (1999).

MODIFIED LIKELIHOOD RATIO TEST

The training sample is denoted by

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \Pi,$$

and it is a sample of size n from the population (Π) of non-nuclear events in the region of interest. A new observation, $\boldsymbol{X}_{n+1}$, is obtained, and given the training sample we wish to test the hypotheses

$$H_0\colon \boldsymbol{X}_{n+1} \in \Pi$$
$$H_1\colon \boldsymbol{X}_{n+1} \notin \Pi.$$

The classical likelihood ratio test statistic is the ratio of the maximized likelihood functions under H_0 and H_1. We let $L_0(\theta) = \prod_{s=1}^{n} f(\mathbf{X}_s;\theta) \times f(\mathbf{X}_{n+1};\theta)$ denote the likelihood function under H_0 (i.e. under the assumption that $\boldsymbol{X}_{n+1} \in H_0$) where θ is an unknown vector-valued parameter associated with the distribution of $\boldsymbol{X}$ under H_0. We also let $\tilde{L}_1(\theta) = \prod_{s=1}^{n} f(\mathbf{X}_s;\theta)$ denote the likelihood based only on the training sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the mixture. Wang, et al. (1997) and Sain, et al. (1999) used the modified likelihood-ratio test statistic

$$W = \frac{\sup_{\theta \in \Theta} L_0(\theta)}{\sup_{\theta \in \Theta} \tilde{L}_1(\theta)}. \tag{2}$$

The usual likelihood ratio would have involved another factor in the denominator consisting of the likelihood function associated with the outlier population. However, we use the form in (2) because of the fact that little is known about the outlier population from which we have only one observation, $\boldsymbol{X}_{n+1}$. It is easily seen in (2) that if $\boldsymbol{X}_{n+1}$ does not belong to Π, this will tend to make W smaller. Hence the rejection region is of the form $W \le W_\alpha$ for some W_α picked to provide a level α test. The null distribution of W has no known closed form in the case of a training sample from a multivariate normal or a mixture of normals, so we use a bootstrap procedure (see Efron and Tibshirani, 1993) to approximate it. The algorithm is given in Sain, et al. (1999).

A SIMPLIFIED TEST

As part of our research effort, we have considered an alternative approach, which is less computational than the modified likelihood ratio testing procedure and is asymptotically equivalent to it (see Wang, et al., 1997). Specifically, in the new approach we fit a density (mixture or non-mixture depending on the situation) to the training data and estimate the parameters (using the EM algorithm if some data are missing). We denote the estimated density as $\hat{f}$. We then simply evaluate this estimated density at the potential outlier point, i.e. we find $\hat{f}(\mathbf{X}_{n+1})$. Obviously, the smaller the value of $\hat{f}(\mathbf{X}_{n+1})$, the more "outlier-like" the new event is. In order to find the distribution of the density function under the null hypothesis $H_0\colon \mathbf{X}_{n+1} \in \Pi$, we use a bootstrap procedure analogous to that used for W. This procedure is much faster and avoids the necessity of calculating the likelihood function in the numerator of (2).

THE EM ALGORITHM

The EM algorithm is a very general procedure for obtaining maximum likelihood estimates when some data are missing, and it is applicable to the mixture-of-normals case considered here. Actually, the EM algorithm may be thought of as a formal procedure for performing the following intuitively appealing approach to dealing with missing data: (a) estimate missing values using the current parameter estimates, (b) revise parameter estimates using the available data and the estimates of the missing values, and (c) iterate on (a) and (b) until convergence of parameter estimates is obtained.

The EM algorithm is directly tied to maximizing the likelihood function. As such, it is often not necessary to actually estimate missing data values but rather is only necessary to estimate the sufficient statistics (based on missing and available data) that are needed in the evaluation of the likelihood function. The EM algorithm thus has two steps: (1) the *expectation step (E-step)* in which the conditional expectation of the sufficient statistics given available data and the current estimates of the parameters are calculated, and (2) the *maximization step (M-step)* in which the log-likelihood function (based on sufficient statistics calculated in Step 1) is maximized to give revised parameter estimates. For more details concerning the use of the EM algorithm in these settings, see Miller, et al. (1993, 1994), and Woodward, et al. (1999).

SIMILATION RESULTS

Consider the situation in which the event in question is measured at three feature variables which we want to simultaneously use in the outlier testing. The question is whether it is preferable to do the outlier testing: (i) using only those data vectors in the training set for which all three features were observed, or (ii) using all data vectors for which at least one of the three feature variables was observed.

Consider the case in which a training sample of size $n = 40$ is obtained from the multivariate normal population with mean μ and covariance Σ. Further suppose that the new event comes from a population that is multivariate normal with mean μ_0 covariance Σ_0. In Table 1 we show results for two separate scenarios. In the table we show the empirical powers based on 1000 replications. We show results using the modified likelihood ratio test and the simplified test based on $\hat{f}$. The first column of each table corresponds to the case in which only complete data vectors are retained and the second column corresponds to the case in which the EM algorithm is used on the complete data set consisting of all observations for which at least one of the three features is available. The rows in the table correspond to the missing data scenarios in which the probability, p_m, of a missing feature takes on the values 0 (i.e. there is no missing data), .38, and .5. In the simulations each data item has a p_m probability of being declared missing. However, if all three features are declared missing then the procedure is repeated until at least one of the observations is declared non-missing. Using this procedure, the expected number of complete vectors out of the 40 observations in the data set is 10 and 6 for p_m equal to .38 and .5 respectively. For the first column in Table 1, training samples of size indicated in parentheses (i.e. 40, 10, or 6) are generated with no missing data. This corresponds roughly to the situation in which the analysis is run only on the vector observations with complete data. The second column in the table corresponds to the case in which all observations, some of which may contain missing features, are analyzed using the EM algorithm. The numbers given in the body of the table are empirical detection probabilities, i.e. they are the proportion of the 1000 replications for which the outlier was detected. The EM entry in the second column of the first row is not given since there is no missing data, and obviously, in this case the EM algorithm requires no iteration and would produce the results given in the first column of this row. Additionally, in parentheses below the detection probabilities are given estimated false alarm rates. That is, in this case we generated the "outlier" from the population of the training data, i.e. MVN(μ, Σ), and we are interested in the proportion of times for which the test (incorrectly) detected an outlier. The test in the simulations is designed to have a false alarm rate of .05, so the empirical false alarm rates provide information concerning whether the actual false alarm rate is near .05. The standard error for detection probabilities shown in this report is .016 while the standard error for the false alarm rates is .007.

From the left-hand column of the table it can be seen that, as would be expected, the presence of missing data has reduced the detection probability both for the modified likelihood ratio test and the test based on $\hat{f}(\mathbf{X}_{n+1})$. Also, in the eight cases shown in the table for which a direct comparison can be made between

the use of the EM algorithm and the use of only the complete vectors, the detection probabilities using the EM-algorithm are always higher, and for p_m =.5 the improvement is substantial. It can also be seen that the detection probabilities for the modified likelihood ratio test and the test based on $\hat{f}(\mathbf{x}_{n+1})$ are similar. It should also be noted that the false alarm rate for both tests have a tendency to be elevated, especially for the case of smaller sample sizes. This phenomenon has been previously observed by Wang, et al. (1997) and Sain, et al. (1999).

OUTLIER TESTING BASED ON WMQ DATA

We next consider the application of these outlier tests in settings dictated by actual seismic data. In particular we consider simulations based on data from the WMQ station in western China (see Hartse, et al., 1997). This distance corrected data consists of n = 134 events which are primarily earthquakes. Data is also available on a few nuclear events observed at WMQ. Our analysis will be based on the three features corresponding to log(Pg/Lg) ratios in the frequency bands 0.5-1, 1.5-3, and 4-8 Hz. There is no missing data on these three features in the data set that was analyzed. Using the three features, a multivariate normal fit to earthquake training data has mean $\mu_{EQ} = (-.024, .016, .015)'$, where the features are in the order 0.5-1, 1.5-3, and 4-8 Hz respectively, and covariance

$$\Sigma_{EQ} = \begin{pmatrix} .0468 & .0268 & .0134 \\ .0268 & .0319 & .0170 \\ .0134 & .0170 & .0188 \end{pmatrix}.$$

Additionally, the multivariate normal fit to the nuclear explosions has mean $(.035, .557, .679)'$ and covariance

$$\Sigma_0 = \begin{pmatrix} .068 & -.004 & .033 \\ -.004 & .017 & .003 \\ .033 & .003 & .019 \end{pmatrix}.$$

The explosion population is so far removed from the training data that detection is essentially assured using any reasonable approach. We follow the approach of Sain, et al. (1999) and Woodward, et al. (1999) and artificially move the outlier population closer to the training data. Specifically, in the simulations given here we consider the outlier population to be multivariate normal with mean $\mu_0 = (.015, .337, .358)'$ and with covariance Σ_0 given above.

In Table 2 we report the results of a simulation study based on the multivariate normal distribution fit to the earthquake training data with the outlier population being the multivariate normal distribution with mean and covariance given by μ_0 and Σ_0 above. While there was no missing data in the original data set analyzed by Sain, et al. (1999), we are able to ascertain the effect of missing data in this situation by simulating training samples from the distribution fit to the earthquake data and randomly assigning some of the observations to be missing. In Table 2 we give the empirical detection probabilities based on 1000 replications with B = 199 in the bootstrap-based outlier test. We consider three training sample sizes, n = 50, 100, and 150, and we again consider the cases in which p_m takes on the values 0, .38, and .5. We give separate tables for the modified likelihood ratio test and the simplified test based on $\hat{f}(\mathbf{x}_{n+1})$. In the table we see that in every case, the use of the EM algorithm improves the detection probability over the use of only the complete vectors. Not surprisingly, this improvement is more apparent when the percent missing is higher and when the original sample size is smaller. As in Table 1, it can be seen that the results using the test based on the modified likelihood ratio and those using the test based on $\hat{f}(\mathbf{x}_{n+1})$ are similar. Of particular interest is the fact that for n = 100 and n = 150, the use of the EM algorithm with p_m = .38 and p_m = .50 yields detection probabilities very similar to those given in the first row of the table for the case in which no data are missing. It can also be seen that the observed false alarm rates were somewhat elevated in some cases. Current investigation by these authors deals with understanding and properly adjusting for these somewhat elevated levels.

APPLICATION TO THE MIXTURE MODEL

We briefly consider the use of the EM algorithm in the mixture model for the case in which the training data are from the model in (1) with $m = 2$, $p_1 = p_2 = 05$ and in which $f_1(x)$ and $f_2(x)$ are multivariate normal densities with

$$\mu_1 = (-25,\ 0)',\ \ \Sigma_1 = \begin{pmatrix} 1 & -.75 \\ -.75 & 1 \end{pmatrix}$$

and

$$\mu_2 = (25,\ 0)',\ \ \Sigma_2 = \begin{pmatrix} 1 & .75 \\ .75 & 1 \end{pmatrix}.$$

Additionally, we assume that the outlier population is MVN(μ_0, I) where $\mu_0 = (0,2)'$ and I is the 2 x 2 identity matrix. We consider the cases $p_m = 0$, 1/3, and .50. In Table 3 we show the results of the outlier test for the case in which, as before, the results are based on 1000 repetitions. The results shown here are for the simplified test based on $\hat{f}(\mathbf{x}_{n+1})$. In the simulations, we let AIC select the number of components with the restriction that $m = 1$ or 2. (See Sain, et al., 1999).

In the table it can be seen that for the $p_m = .33$ case, the use of the EM algorithm resulted in a marginal decrease in detection probability whereas in the case of $p_m = 0.5$ there was a marked improvement using the EM algorithm. Other simulations (not shown here) also show that observed false alarm rates may be high, especially in the cases of smaller number of available data values. Our experience has shown that much care must be exercised in applying the EM algorithm in this setting. It is a computationally difficult problem to require estimation of the number of components and parameters of each component in the case in which a fairly substantial amount of data are missing. Further investigation in this area is needed.

CONCLUSIONS AND RECOMMENDATIONS

Our results indicate that when the training data can be reasonably modeled as a multivariate normal, then the use of all available data using the EM algorithm provides improved detection probability over that obtained using only those data vectors for which we have complete data. Our results also show that the use of the simplified outlier test based on $\hat{f}(\mathbf{x}_{n+1})$ should be considered since it is computationally faster and gives detection probability results similar to those obtained using the test based on the modified likelihood ratio. Initial results indicate that the use of the EM algorithm may be plausible when the training data can be modeled as a mixture of normals, but further work is required in this area. When the training data cannot be reasonable well fit using a multivariate normal, possible modifications include using a mixture approach in which it is assumed that the covariance matrices are equal or possibly using some sort of nonparametric kernel density estimator.

REFERENCES

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society* B39, 1-38.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Fisk, M.D., Gray, H.L., and McCartor, G. (1996). Regional Event Discrimination without Transporting Thresholds, *Bulletin of the Seismological Society of America*, 86, 1545-1558.

Hartse, H.E., Taylor, S.R., Phillips, W.S., and Randall, G.E. (1997). A Preliminary Study of Regional Seismic Discrimination in Central Asia with Emphasis on Western China, *Bulletin of the Seismological Society of America* 87, 551-568.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley and Sons, Inc.

Miller, J.W., Gray, H.L., and Woodward, W.A. (1993). Discriminant Analysis and Outlier Testing when Data are Missing, Phillips Laboratory Technical Report, ARPA F29601-91-K-DB25.

Miller, J.W., Woodward W.A., Gray, H.L., Fisk, M.A., and McCartor, G.D. (1994). A Hypothesis-Testing Approach to Discriminant Analysis with Mixed Categorical and Continuous Variables when Data are Missing," Technical Report No. SMU/DS/TR-273, Department of Statistical Science, Southern Methodist University.

Sain, S.R., Gray, H.L., Woodward, W.A., and Fisk, M.D. Outlier Detection when Training Data are Unlabeled, *Bulletin of the Seismological Society of America* 89, 294-304.

Taylor, S.R. and Hartse, H.E. (1997). An Evaluation of Generalized Likelihood Ratio Outlier Detection to Identification of Seismic Events in Western China, *Bulletin of the Seismological Society of America*, 87, 824-831.

Wang, S., Woodward, W.A., Gray, H.L., Wiechecki, S., and Sain, S.R. (1997). A New Test for Outlier Detection from a Multivariate Mixture Distribution, *Journal of Computational and Graphical Statistics*, 6, 285-299.

Woodward, W.A., Sain, S.R., Gray, H.L., and Fisk, M.D. (1999). Outlier Testing when Some Data are Missing, submitted to *Pure and Applied Geophysics*.

Table 1. Empirical Detection Probabilities for 2 Simulated Models

(a) Training sample of size $n = 40$ from MVN(μ,Σ) where $\mu = (0,\ 0,\ 0)'$ and $\Sigma = \begin{pmatrix} 1 & .2 & .2 \\ .2 & 1 & .2 \\ .2 & .2 & 1 \end{pmatrix}$

Outlier from MVN(μ_0, Σ) where $\mu_0 = (25,\ 25,\ 25)'$

	Modified Likelihood Ratio Test			Test based on $\hat{f}(X_{n+1})$	
p_m	Complete Vectors	EM		Complete Vectors	EM
0.50	.829 (40) (.046)	---		.836 (40) (.066)	---
0.38	.684 (10) (.073)	.790 (.060)		.622 (10) (.068)	.781 (.074)
0.50	.523 (6) (.096)	.790 (.069)		.332 (6) (.042)	.730 (.063)

(b) Training sample of size $n = 40$ from MVN(μ,Σ) where $\mu = (0,\ 0,\ 0)'$ and $\Sigma = \begin{pmatrix} 1 & -.7 & -.7 \\ -.7 & 1 & .7 \\ -.7 & .7 & 1 \end{pmatrix}$

Outlier from MVN(μ_0, Σ_0) where $\mu_0 = (1,\ 1,\ 1)'$ and $\Sigma_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

	Modified Likelihood Ratio Test		Test based on $\hat{f}(X_{n+1})$	
p_m	Complete Vectors	EM	Complete Vectors	EM
0.00	.666 (40) (.058)	—	.672 (40) (.053)	—
0.38	.582 (10) (.059)	.605 (.059)	.516 (10) (.070)	.628 (.074)
0.50	.442 (6) (.093)	.613 (.083)	.311 (6) (.059)	.606 (.086)

Table 2. Empirical detection probabilities using the models fit to the WMQ data

Modified Likelihood Ratio Test

	n = 50		n = 100		n = 150	
p_m	Complete Vectors	EM	Complete Vectors	EM	Complete Vectors	EM
0.00	.950 (50) (.057)	---	.968 (100) (.049)	---	.955 (150) (.052)	---
0.38	.856 (13) (.079)	.934 (.088)	.915 (25) (.067)	.956 (.043)	.932 (38) (.057)	.962 (.056)
0.50	.675 (7) (.113)	.856 (.076)	.846 (14) (.078)	.945 (.061)	.894 (21) (.055)	.953 (.072)

Test Based on $\hat{f}(X_{n+1})$

	n = 50		n = 100		n = 150	
p_m	Complete Vectors	EM	Complete Vectors	EM	Complete Vectors	EM
0.00	.939 (50) (.059)	---	.959 (100) (.065)	---	.962 (150) (.041)	---
0.38	.813 (13) (.054)	.943 (.084)	.909 (25) (.044)	.947 (.060)	.922 (38) (.060)	.954 (.063)
0.50	.481 (7) (.068)	.906 (.093)	.824 (14) (.066)	.948 (.101)	.883 (21) (.063)	.971 (.071)

Table 3. Detection Results in the Mixture Case Described in the Text

p_m	Complete Vectors	EM
0.00	.922 (50) (.043)	---
0.33	.872 (20) (.053)	.792 (.049)
0.50	.576 (13) (.040)	.722 (.068)