TESTING FOR OUTLIERS FROM A MIXTURE DISTRIBUTION
WHEN SOME DATA ARE MISSING

Wayne A. Woodward* and Stephan R. Sain**
* Southern Methodist University
** University of Colorado at Denver

ABSTRACT

We consider the problem of multivariate outlier testing from a population from which a training sample is available. We assume that a new observation is obtained, and we test whether the new observation is from the population of the training sample. Problems of this sort arise in a number of applications including nuclear monitoring, biometrics (including fingerprint and handwriting identification), and medical diagnosis. In many cases it is reasonable to model the population of the training sample using a mixture-of-normals model (e.g. when the observations come from a variety of sources or the data are substantially non-normal). In this paper we consider a modified likelihood ratio test that is applicable to the case in which: (a) the training data follow a mixture-of-normals distribution, (b) all labels in the training sample are missing, (c) some of the observation vectors in the training sample have missing information, and (d) the number of components in the mixture is unknown.

The approach often used in practice to handle the fact that some of the data vectors have missing observations is to perform the test based only on the data vectors with full data. When large amounts of data are missing, use of this strategy may lead to loss of valuable information, especially in the case of small training samples which, for example, is often the case in the nuclear monitoring setting mentioned previously. An alternative procedure is to incorporate all \( n \) of the data vectors using the EM algorithm to handle the missing data. We use simulations and examples to compare the use of the EM algorithm on the entire data set with the use of only the complete data vectors.

Key Words: EM Algorithm, Mixture Model, Missing Data, Outlier Detection
1. Introduction

We consider the problem of testing a new data value to determine whether it should be considered an outlier from a distribution for which we have a training sample, i.e. "outlier testing." Fisk, Gray, and McCartor (1996) and Taylor and Hartse (1997) have used a likelihood ratio test for detecting outliers from a multivariate normal (MVN) distribution fit to the training data when no data were missing. These authors applied the test to the problem of detecting seismic signals of underground nuclear explosions when a training sample of non-nuclear seismic events is available.

Our focus in this paper will be the case in which the training data are modeled as a mixture of normals. A mixture model is an obvious choice for a wide variety of settings. For example, in the seismic setting discussed above, the population of non-nuclear observations in a particular region may consist of observations from a variety of sources such as earthquakes and mining explosions, and differing types of earthquakes. Additionally, in the area of medical diagnosis, benign tumors may be of several types, etc. The flexibility of the mixture-of-normals model also makes it useful for modeling non-normality even if distinguishable components are not present. The training data will be considered a sample of size \( n \) from a mixture distribution whose density is given by

\[
f(x) = \sum_{i=1}^{m} \lambda_i f_i(x; \mu_i, \Sigma_i)
\]

where \( m \) is the number of components in the mixture, \( f_i(x; \mu_i, \Sigma_i) \) is the MVN density with mean vector \( \mu_i \) and covariance matrix \( \Sigma_i \) associated with the \( i \)th component, the \( \lambda_i, \ldots, m \) are the mixing proportions, and \( x \) is a \( d \)-dimensional vector of variables.

Letting the training sample be denoted by \( X_1, \ldots, X_n \) and the new observation (whose distribution is unknown) by \( X_u \), then we wish to test the hypotheses
\[ H_0 : \mathbf{X}_u \in \Pi \]
\[ H_1 : \mathbf{X}_u \notin \Pi \]

where \( \Pi \) denotes the population of the training data.

We consider the case in which data may be missing in the training data. In the case of a mixture model, there are at least three different ways in which "data" may be missing:

(a) missing labels
(b) unknown number of components
(c) missing data in the data vectors

A "label" is said to be known for a given observation if it is known to which component in the mixture that observation belongs. Wang, Woodward, Gray, Wiechecki, and Sain (1997) developed a modified likelihood ratio test for the case in which some but not all of the labels may be missing. The authors assumed that the number of components, \( m \), is known and that there is no missing data in the data vectors. The likelihood function under \( H_0 \) (i.e. under the assumption that \( \mathbf{X}_u \in H_0 \)) is denoted by \( L_0(\theta) \) where \( \theta \) is an unknown vector-valued parameter associated with the distribution of \( \mathbf{X} \) under \( H_0 \). Likewise, let \( \bar{\mathcal{L}}_1(\theta) = \prod_{s=1}^{n} f(X_s; \theta) \) denote the likelihood based only on the training sample \( \mathbf{X}_1, ..., \mathbf{X}_n \).

Wang, et al. (1997) and Sain, Gray, Woodward, and Fisk (1999) used the modified likelihood-ratio test statistic

\[ W = \frac{\sup_{\theta \in \Theta} L_0(\theta)}{\sup_{\theta \in \Theta} \bar{\mathcal{L}}_1(\theta)} \]  

(2)
The usual likelihood ratio involves a second factor in the denominator, $h(x_u)$, where $h(x)$ is the density function of the outlier population and $x_u$ denotes a single observation available from that population. However, estimating $h(x)$ is very difficult with only one observation available and when a priori information is not available concerning the outlier distribution. Thus, it makes sense to estimate $h(x)$ nonparametrically. Moreover, given any of the potential nonparametric density estimators of $h(x)$ in the case of only one data point, the factor $h(x_u)$ will not vary with $\theta$ in the maximization process nor will $h(x_u)$ vary as $x_u$ varies from sample to sample. Consider, for example, a histogram estimator of $h(x)$. With a single data value, such an estimator would be a constant regardless of the value of $x$. Thus, for simplicity we use $W$ in (2).

It is easily seen in (2) that if $x_u$ does not belong to $\Pi$, then $W$ will tend to be small. Hence the rejection region is of the form $W \leq W_\alpha$ for some $W_\alpha$ picked to provide a level $\alpha$ test. Since the null distribution of $W$ has no known closed form, Wang, et al. (1997) used a bootstrap procedure (see Efron and Tibshirani, 1993) to derive the critical value $W_\alpha$. Whenever some of the training data are unlabeled, the parameters $\lambda_i, \mu_i$, and $\Sigma_i$ of the mixture model are estimated via the Expectation-Maximization (EM) algorithm (see Dempster, Laird, and Rubin, 1977, McLachlan and Krishnan, 1997, and Redner and Walker, 1984). Based on simulations, Wang, et al. (1997) showed that in this setting, the modified likelihood ratio test can be used successfully for outlier detection.

Sain, et al. (1999) extend the results of Wang, et al. (1997) 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 labeled. 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 normal model due to apparent non-normality of the data.
even when there are not any identifiable groups of observation types represented in the training data.

In this paper we consider the case in which some of the variables may be missing for some observations in the training sample. For example, in the seismic setting log(Pg/Lg) ratios at higher frequency bands are often missing because of attenuation effects on high frequencies. We consider the case in which \( d \) variables are observed on the new observation, and we denote this observation by \( X_u = (X_{u1}, X_{u2}, ..., X_{ud})' \), where \( X_{uj} \) denotes variable \( j \) observed on the new observation. We further assume that there exists a training sample

\[
X_1 = (X_{11}, X_{12}, X_{13}, ..., X_{1d})' \\
X_2 = (X_{21}, X_{22}, X_{23}, ..., X_{2d})' \\
\vdots \\
X_n = (X_{n1}, X_{n2}, X_{n3}, ..., X_{nd})'
\]

from \( \Pi \). When some of the training data includes missing data, the training sample has the general appearance

\[
X_1 = (X_{11}, X_{12}, X_{13}, ..., X_{1j}, X_{1,j+2}, ..., X_{1d})' \\
X_2 = (-, X_{22}, X_{23}, ..., X_{2j}, X_{2,j+1}, ..., X_{2d})' \\
\vdots \\
X_n = X_{n1}, -, X_{n3}, ..., X_{nj}, X_{n,j+1}, X_{n,j+2}, ..., X_{nd})'
\]

where " - " denotes that the particular variable is missing for that observation. Thus, to apply standard or recently developed outlier detection methodology, one must reduce the data to a subset of the original training data that includes only those \( l \) (where \( l < n \)) data
vectors for which all of the variables were observed. It is clear that such a procedure can result in a loss of information and should lead to a reduction in detection power. To demonstrate the extent of this problem, Woodward, Sain, Gray, Zhao, and Fisk (2002) show that, depending on the missing data probabilities and the number of variables, the number of complete vectors available for analysis may be dramatically smaller than the number of original cases. For example, in the case of $d = 4$ variables and a 25% chance that an observation will be missing, we expect fewer than one-third of the data vectors to be retained for analysis using the strategy of analyzing only the complete data vectors. This is in spite of the fact that about 75% of the original data set should be available for use.

Another problem arises if there are no cases or only a very few cases in which all $d$ of the variables are observed. If the strategy of using only the complete vectors is used, then some of the variables may need to be deleted. This may also result in loss of some important information. It should also be pointed out that we are not considering the case in which there is missing data in the outlier. In an application of the techniques developed here, the variables observed in the outlier determine the variables to be used in the outlier test.

The purpose of this paper is to examine the extent to which detection power can be improved by retaining all of the available data as compared to using only the complete data vectors in the mixture-of-normals case outlined above. Woodward, et al. (2002) consider this problem in the case in which the population of the training data ($\Pi$) is multivariate normal. They show that substantial increases in power can be obtained by the use of all available data via the EM algorithm compared to the use of only complete data vectors. Section 2 describes this method for the case in which the population of the training data is a mixture of multivariate normals. In Section 3 we discuss three simulation studies comparing the use of only vectors having complete data with the use of the EM algorithm on data with some missing observations.
An Outlier Testing Approach when Some Data is Missing

This is the procedure for testing whether \( X \) lies in the population. The training data \( W \) considers the missing labels in the training sample missing, and the base vectors that are missing.

\((a)\) of the M-Step Mode Pair

The algorithm is as follows: training the likelihood ti etc. of data are missing co only used to estimate the parameters. The labels missing the likelihood considers here the small algorithm only real missing labels but also using the set of observations. The algorithm may be thought of as a primal procedure for performing the M-Step appealing again to solving using data.

Missing val. Estimate the parameter using the missing value estimates these parameter estimates, repeat the estimates using all estimates, etc. This primal co existence

The algorithm directly maximizes the likelihood function. As such, the necessary actually using data can be only necessary in the sufficient statistic base using and data ded the sufficient the likelihood function. The algorithm thus has the step the expected \( p(E|ep) \)hic the individual expected insufficient statistic aila data the item th are called th and th \( ep \) M-step, hic the likelihood function based
sufficient statistics calculated in (i) is maximized to give revised parameter estimates. See Little and Rubin (1987) for details of the basic method.

The mixture-of-normals setting considered here is a special case of the general location model discussed by Olkin and Tate (1961) involving both continuous and categorical variables. Conceptually, the component label can be thought of as a matrix, $Y$, of categorical variables specifying component membership for each sample value. If observation $k$ is known to be from component $i$, then $y_{ik} = 1$ and $y_{ik}' = 0$ for $i' \neq i$. Now, for the $k$th observation we have

$$P(Y_{ik} = 1) = \lambda_i$$

$$X|Y_{ik} = 1) \sim \text{MVN}(\mu_i, \Sigma_i)$$

When all data are available, then the sufficient statistics for the mixture model parameters are given by

$$N_i = \sum_{k=1}^{n} Y_{ik}$$

$$S_i = \sum_{k=1}^{n} Y_{ik} X_k$$

$$SS_i = \sum_{k=1}^{n} Y_{ik} X_k X'_k$$

In words, $N_i$ counts the number of observations from component $i$ while $S_i$ and $SS_i$ are the sum and the sum of squares and cross products, respectively, of observations from component $i$. When there are no missing data, the corresponding maximum likelihood estimates are

$$\hat{\lambda}_i = \frac{N_i}{n}$$
where

\[
\hat{\mu}_i = \frac{S_i}{N_i},
\]
\[
\hat{\Sigma}_i = \frac{SS_i}{N_i} - \hat{\mu}_i \hat{\mu}_i',
\]
for \(i = 1, \ldots, m\).

Now we consider the case in which some observations can be missing, i.e. the sufficient statistics in (3) cannot be calculated. Wang, et al. (1997) and Sain, et al. (1999) considered the case in which some or all of the labels, \(Y\), are missing. In this paper, we consider the case in which not only the labels but some of the continuous variables can be missing. When some data are missing, the E-step of the EM algorithm consists of finding conditional expectations of the sufficient statistics in (3), and the M-step involves simply calculating the estimates in (4) using the conditional expectations of the sufficient statistics in place of the sufficient statistics themselves. Specifically, let \(X_{obs}\) denote the actual observed data, and let \(\hat{\theta}^{(j)}\) denote current estimates of the parameters entering the \(j\)th iteration of the algorithm. Then, the conditional expectations calculated in the E-step are:

\[
E\left[ N_i \mid \hat{\theta}^{(j)}, X_{obs} \right]
\]
\[
E\left[ S_i \mid \hat{\theta}^{(j)}, X_{obs} \right]
\]
\[
E\left[ SS_i \mid \hat{\theta}^{(j)}, X_{obs} \right]
\]

(5)

More details concerning the calculation of the conditional expectations can be found in Little and Rubin (1987), McLachlan and Peel (2000), McLachlan and Krishnan (1997), and Miller, Woodward, Gray, Fisk, and McCartor (1994). Upon convergence of the EM algorithm, the estimates obtained by using the conditional expectations in (5) in the final
iteration to solve (4) are called the EM estimates. It should be noted that the likelihood function depends on the data only through the sufficient statistics. Thus, upon convergence, the maximized likelihood function can be calculated using the final EM parameter estimates and final conditional expectations of the sufficient statistics.

(b) The Outlier Testing Algorithm

In this section we discuss an algorithm for outlier testing in the mixture-of-normals setting when some data are missing. We assume that the distribution of the training data can be approximated by a mixture-of-normals where the number of components, \( m \), is unknown, but a maximum number of components to be considered is given (MAXM). Note also that we assume that the outlier contains no missing data. The algorithm is as follows:

**Step 1:** Using the \( d \) variables and for each \( m, m = 1, \ldots, \text{MAXM} \), fit an \( m \)-component mixture to the training data and calculate AIC. Specifically

- For each \( m \) we use a hierarchical/k-means clustering routine to find starting values for the mixture parameters (see Sain, et al., 1999 and Kaufman and Rousseeuw, 1990). Distance is calculated using a 'normalized' Euclidean distance metric that takes into account missing values. The distance \((\Delta_{jk})\) between points \(X_j\) and \(X_k\) is measured by first defining \(\Delta_{jk}(l) = 0\) if \(X_{jl}\) or \(X_{kl}\) is missing and \(\Delta_{jk}(l) = X_{jl} - X_{kl}\) otherwise. We then calculate the distance as \(\Delta_{jk} = \frac{d}{d-d_m}\sum_{l=1}^{d} \Delta_{ij}^2(l)\) where \(d_m\) is the number of \(\Delta_{ij}(l)\)'s that were set equal to zero because of missing data. The variables are prescaled to have mean zero and unit variance in each dimension.

- Using these starting values, we obtain EM estimates of the parameters
and find the associated maximized likelihood \( L_{\text{max}}(m) \) using the procedure described in Section 2(a) above.

AIC and BIC are calculated using the formulas:

\[
\text{AIC}(m) = -2 \ln(L_{\text{max}}(m)) + 2(\text{# of free parameters}),
\]

\[
\text{BIC}(m) = -2 \ln(L_{\text{max}}(m)) + \ln(n)(\text{# of free parameters}),
\]

where the number of free parameters is \( m - 1 + dm + dm(m - 1)/2 \)

and where \( n \) is the number of observations in the training sample.

**Step 2:** Select the number of components for which AIC/BIC is minimized. The number of components selected will be denoted \( m_{\text{AIC}} \) or \( m_{\text{BIC}} \). Note that a number of components, \( m \), will not be considered as a candidate for the number of components if:

- any computational problems are encountered while obtaining EM estimates based on an \( m \)-component-model because of singular covariance estimates, etc.
- when fitting an \( m \)-component model, any of the \( \hat{\lambda}_i \)'s, \( i = 1, \ldots, m \) are less than the maximum of 0.05 and \((d + 2)/n\). This restriction is imposed to avoid instability encountered when one of the mixture components, and the resulting estimates, are based on a very small number of data values.

**Step 3:** The modified likelihood ratio statistic, \( W \), is calculated for the data, using the number of components, \( m \), found by AIC or BIC. The denominator of \( W \) is calculated for the \( n \) observations in the training sample, while the numerator of \( W \) is obtained by augmenting the training sample with the outlier point and recalculating the EM estimates and associated likelihood function. It should be noted that in this case, the number of components,
$m$, and the starting values for the parameters in the EM algorithm are those obtained from the $n$ observations in the training sample.

**Step 4:** The bootstrap is used to find the distribution of $W$. At each bootstrap iteration, $b, b = 1, ..., B$, we use the parametric bootstrap to obtain $n + 1$ observations from the distribution of the training data. Data from a mixture distribution are generated where the number of components, $m$, and parameter values are those estimated from the $n$ observations in the training sample. Each bootstrap sample is generated so that it involves the same missing data structure as the original sample. We perform the modified likelihood ratio test on each bootstrap sample using parameter estimates based on the number of components and the starting values obtained from the training data. The associated test statistic is denoted $W^*_b$.

Note that the $(n + 1)$st observation must have complete data for the variables under consideration. If the nonparametric bootstrap were used, only those training sample values with complete data would be available for resampling as the $(n + 1)$st observation. If there were only a few observations in the training data with complete data, then it is clear that nonparametric bootstrapping would not be desirable, and thus we use the parametric bootstrap.

**Step 5:** Define $W_\alpha$ to be the (100$\alpha$)th percentile of the $W^*_b$'s. Reject $H_0$ and conclude that the $(n + 1)$st point is an outlier if $W \leq W_\alpha$. 

12
It should be noted that prior to performing the likelihood ratio test on the training data in Step 3 above, we check to determine if the potential outlier is "super extreme". Numerical problems in computing likelihoods can occur if the outlier is too far removed from the training data. Of course, if a potential outlier is sufficiently far away from the training data, there is actually no reason to perform the likelihood ratio test. Currently, a new observation $X_u$ is considered to be a "super-extreme" outlier if each of the estimated component density functions evaluated at the new observation is less than $e^{-25}$ (i.e. $\approx 10^{-11}$). If the new observation is "extreme" by this criterion, then it is declared an outlier and the algorithm terminates.

3. Simulations

In this section we report the results of simulation studies that examine the effect of missing data, missing labels, and unknown number of components on the detection power of the outlier test based on $W$. In each case the training data are generated from a mixture distribution as in (1) with $m = 2$, where $\lambda_1 = \lambda_2 = 0.5$ and the component distributions are multivariate normal.

(a) Bivariate Examples

In this section we consider two mixture scenarios and in each case we use training sample sizes of $n = 30, 40, \text{and } 60$. In the first setting, we let the training data be from a mixture where $\mu_1 = (0, 0)'$, $\mu_2 = (6, 4)'$, $\Sigma_1 = \left(\begin{array}{cc} 1 & .5 \\ .5 & 1 \end{array}\right)$, and $\Sigma_2 = \left(\begin{array}{cc} 1 & -.5 \\ -.5 & 1 \end{array}\right)$. The outlier population is $\text{MVN}(\mu_0, \Sigma_1)$ where $\mu_0$ takes on the values $\mu_0 = (0,5)'$, $(1,4.5)'$, $(2,4)'$, and $(5,8.5)'$. In the second scenario, the training data are from a mixture where $\mu_1$, $\Sigma_1$, and $\Sigma_2$ are as before and where $\mu_2 = (0, 6)'$. In this case we consider outlier populations that are $\text{MVN}(\mu_0, \Sigma_1)$ where $\mu_0 = (-4,3)'$, $(-2,3)'$, $(4,3)'$, $(5,3)'$ and (-
These mixture distributions and outlier means are shown in Figures 1 and 2. The contours of the mixture components are shown with solid contours while the means of the simulated outlier populations are shown with "x". The outlier population with mean \((0,5)\)' is shown with dashed contours.

In Tables 1 and 2 we show the results of simulations based on 1000 replications from the scenarios described above in which the testing procedure is run at the \(\alpha = 0.05\) level of significance. In each case we generate a training sample (that has some missing data) along with an outlier from \(\text{MVN}(\mu_0, \Sigma_1)\). Denoting the \(i\)th observation in the training sample by \(x_i \ (x_{i1}, x_{i2})'\), then a random procedure is used to give each of the \(x_{ij}\) a \(p_{\text{miss}}\) probability of being declared missing and thus replaced in the data set by a missing data indicator. If, however, by using this procedure both variables in an observed vector are missing, then we repeat the procedure of randomly assigning these individual features as missing until at least one of \(x_{i1}\) or \(x_{i2}\) is not declared to be missing. Based on a given missing value probability, \(p_{\text{miss}}\), the expected number of vectors for which all of the observations are available, is given in the case of \(d\) variables by

\[
\eta_F = \frac{n(1-p_{\text{miss}})^d}{(1-p_{\text{miss}}^d)}.
\]

The simulations shown in Tables 1 and 2 are based on the case in which \(p_{\text{miss}} = 0.5\). In the tables we show the proportion of the 1000 replicated outliers that were detected. These detection proportions are found using two approaches. First we consider the strategy of using only those vectors for which both variables are observed, and we denote this the "full vector" approach. It should be noted that the expected number of complete data vectors in this case is one-third of the sample size. As a second approach we use all available data in the training sample through the use of the EM algorithm. In both cases we let AIC select the number of components up to a maximum of two components. It
should be noted that the algorithm enhanced the likelihood that AIC would correctly select components when using only the full vectors. AIC selected components about 0% of the time, while the EM approach using all available data resulted in components being selected about 90% of the time. It should be noted that this is the expected number of full vectors.

In Table 1, all available data using the EM algorithm gives substantially better detection than all other cases. In Table 2, not seen is the marked improvement in detection using the EM algorithm.

Examination of the EM algorithm shows that the two components of the mixture differ only with respect to one variable. That is, if the second variable is missing, then the really informative first variable concerning the component membership of the observation.

It is clear from the figure that detecting the algorithm to improve parameter estimation and consequently detect the training data from the mixture population when the data are unlabeled, the data are assumed to be equal, and the number of components unknown. The above discussion suggests that the EM should be expected to aid improved power in the components of the mixture, accompanied by separation both the marginal and bivariate.

Tables 1 and 2 show that the observed significance rate of the EM is the smallest sample size-based tendency for the significance levels using both techniques. The slightly larger than the nominal level for these sample sizes.
(b) A Three Variable Example

In this section we consider a simulation study in which there are three variables to be used for outlier testing and the population of the training data is described by a mixture model with \( m = 2 \) components. Specifically, the components are MVN \((\mu_i, \Sigma_i)\), \(i = 1, 2\) where \(\mu_1 = (0, 0, 0)'\), \(\Sigma_1 = \begin{pmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{pmatrix}\), \(\mu_2 = (4, 4, 4)'\), \(\Sigma_2 = \begin{pmatrix} 1 & -.5 & -.5 \\ -.5 & 1 & .5 \\ -.5 & .5 & 1 \end{pmatrix}\), and where \(p_1 = .5\), \(p_2 = .5\). In each case the training sample size is \(n = 100\) with \(p_{mis} = .5\). We considered 27 outlier populations each with covariance matrix equal to \(\Sigma_1\). In Figure 3 we show the 3-dimensional (solid) contours of the component distributions. Also shown in the figure are the 27 means used for the outlier populations in the simulation study. We also show the (dashed) contour for the outlier population centered at \((2, 2, 5.5)'\). In Tables 3 and 4 we show the simulation results based on 1000 replications of training samples of size \(n = 100\) using the \(\alpha = .05\) level of significance. The missing data probability is \(p_{mis} = .5\) for each variable and the expected number of full vectors is about 14. In Table 3 we show the results for the case in which AIC was used to choose between the options of 1 and 2 components for the mixture distribution. Of course, the actual number of components is 2. In the table we see that the detection results using the EM algorithm are substantially higher than those obtained using only the full vectors. This provides strong evidence of the fact that in this case, it would be a serious mistake to ignore the information contained in the incomplete vectors. The dramatic increase in detection power is explained by the separation in all variables as shown in Figure 3 along with the dramatic reduction in sample size when restricting to full vectors. Also impacting the detection power is the fact that because of the reduced sample size, AIC tends to incorrectly select only one component in about one-third of the samples when using the full vector approach. This is somewhat surprising.
based on the fairly wide separation between the two components in the mixture. On the other hand, AIC nearly always correctly chooses two components using the EM algorithm with all available data. Estimates of observed significance level based on 5000 replications are .066 for the full vector approach and .061 using the procedure based on the EM algorithm. The standard error for the significance level estimates is .003 indicating that the observed significance levels are slightly inflated over the nominal .05 level.

In Table 4 the detection power results are shown for the case in which the maximum allowable choice for AIC is three components. For the full vector procedure, when three components are allowed, AIC continues to choose one component in about one-third of the samples and three components in less than 2% of the cases. Thus, the detection results in Table 4 for the full vector case are very similar to those in Table 3. However, when choosing among one, two, and three components in the EM procedure, AIC almost never chose one component, correctly chose two components 60% of the time, and incorrectly chose three components in about 40% of the cases. It can be seen in Table 4 that the result of allowing a possible third component in the model fit to the training data is to somewhat reduce the detection power for the EM procedure in Table 4 as compared to Table 3. However, it should be noted that in Table 4, the detection power for the EM procedure is still substantially higher than that for the full vector approach. Estimates of observed significance level based on 5000 replications are .064 for the full vector approach and .075 using the EM and all available data. Thus, the effect of the allowable third component is to somewhat increase the observed significance using the EM approach.

It is well known that the model order selected by AIC tends to be high when the sample size is large. For this reason, we considered the use of BIC to pick the number of components, allowing from one to three components. In this case using the full vector approach, BIC incorrectly picks one component a little over 40% of the time and rarely picks three components. The detection power using the BIC allowing up to three
composites similar to that of Tables and for the full \texttt{anova} model. Using the EM procedure, BIC pick composite and the sample. Thus, surprisingly \texttt{if} effect is for the EM app in this similar to those Table Thus, this setting \texttt{C} be preferred over AIC. The observed significance is based on \texttt{ns} statistic for the full factor an EM app each respectively.

A selected simulation results at $p_{mis}$ where \texttt{no} expected number full recto sh that the EM algorithm substantially detected over the full vector approach this although as dramatically the $p_{mis}$ sho Tahl 1-4.

\textit{(c) Simulatio based Se Data}

In the prbl detection of nuclear \texttt{exp} certain varia calculated from the data these variable collected on non-nuclear seismic events. \texttt{ns} observed, the probl ofts est to decide whether should be considered either on the populato the training data and therefore potentially learn. To these variables variables that be asked by sta several thousand kilo meters the are ed to depth based hys calculatio and \texttt{i}) the difference between body and surface magnitudes simplicity will deno these variables depth \texttt{m_i} respectively. In this example, the mixture model the training data to deep earthquake \texttt{D} Q and shall earthquake (\texttt{Q}) surprising that deep earthquakes end ha larger depth estimates than either shall earthquake \texttt{e} xpl \texttt{ns}, while the depths of shall earthquakes and \texttt{exp} similar each the. The value of\texttt{P1} variable $m_i$ end \texttt{R1} larger for deep earthquake and explosion, than for hallow earthquakes.\texttt{explo pro low
magnitudes of earthquakes tend to increase as depth increases. Actual values of these variables, databases of earthquakes and experiments, are classified. How in Figure 1 illustrate this situation described above. Seismological versions of the variables showing relative positioning. The large number of depth-related earthquakes reflects the fact that deep earthquakes may be many kilometers while shallow earthquakes by definition earthquakes relatively low on the surface.

In our simulations, consider two separate outlier populations. These two populations illustrated in Figure 2 with two styles of dashed lines. The other population is on the lower right of the figure and denoted by FX. The population, denoted "HYP" used for illustrating purpose only and ill be referred to as the hypothetical outlier population. We considered the distribution of deep and shallow earthquakes to be modeled as a distribution of equal mixture of components and we assume that 0% of the observations are missing. We also observed from the other population. We simulate 1000 replications for this scenario.

The proportion of the replicates for which the detected gain in Table 1, in the simulations, AIC is used to select between the components, with similar results for the two outlier populations considered. In Table 1, we can see that the EM approach consistently higher detection power than that obtained using full vector of the explorable population. The improvement using the EM is most pronounced at the hypothetical outlier population, there is very pronounced improvement using the EM. The EM approach correctly selected the distribution, only about 0% of the time while for this sample, the EM approach, AIC selects the component mixture of the components. It can be seen that improper selection of
component model has a deleterious effect on detecting the hypothetical outlier, thus providing an explanation for the striking improvement of the EM approach in this case. It can also be seen from Figure 4 that incorrectly choosing a 1-component model would not have the dramatic negative effect on outlier detection from the explosion population, and in Table 5 it can be seen that the full vector and EM results for $n = 30$ are not dramatically different in this case. For $n = 50$ AIC picked two components in about 80% of the cases using full vectors and in about 99% of the cases using the EM approach, while for $n = 75$, AIC picked a 2-component model at least 97% of the time using either approach for handling missing data.

4. Concluding Remarks

In this paper we have examined the use of two techniques for handling missing data in the problem of testing for outliers from a mixture of multivariate normal distributions. The simulations shown here indicate that the utilization of all available data via the EM algorithm can result in higher detection probabilities than those obtained using only the full vectors. Woodward, et. al. (2002) showed similar improvement using the EM algorithm in outlier testing from a multivariate normal distribution. The mixture case discussed here is more complex in nature, and among other factors, detection performance depends on the number of components selected. In general, caution must be used to assure that sufficient sample size is available to provide reasonable estimates of the mixture model parameters. We have also shown in Example 3a that performance of the EM algorithm depends on the amount of information concerning component membership that is available in data values with missing observations.

It is shown that when using the full vector approach, AIC tends to underestimate the number of components for relatively small sample sizes and a substantial amount of missing data, i.e. for cases in which the resulting sample of full vectors is small. This can
lead to very poor discrimination performance. e.g. the \( n = 30 \) case using the hypothetical outlier in Example 3c. Example 3b shows dramatic improvement using the EM algorithm in a three variable case over results obtained using only the full vectors. In this case it is shown that the tendency of AIC to pick too many components for large samples may negatively effect detection power. Thus it may be useful to examine the application of alternative order selection criteria such as BIC.
REFERENCES


Table 1. Detection Power at $\alpha = .05$ Level of Significance for the Bivariate Case -- Components Separated in Each Variable

<table>
<thead>
<tr>
<th>Sig. Level</th>
<th>$n = 30$</th>
<th></th>
<th>$n = 40$</th>
<th></th>
<th>$n = 60$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Vectors</td>
<td>EM</td>
<td>Full Vectors</td>
<td>EM</td>
<td>Full Vectors</td>
<td>EM</td>
</tr>
<tr>
<td>(0,5)</td>
<td>.796</td>
<td>.867</td>
<td>.872</td>
<td>.957</td>
<td>.967</td>
<td>.989</td>
</tr>
<tr>
<td>(1,4.5)</td>
<td>.576</td>
<td>.740</td>
<td>.653</td>
<td>.859</td>
<td>.868</td>
<td>.924</td>
</tr>
<tr>
<td>(2,4)</td>
<td>.339</td>
<td>.560</td>
<td>.502</td>
<td>.677</td>
<td>.689</td>
<td>.786</td>
</tr>
<tr>
<td>(5,8.5)</td>
<td>.741</td>
<td>.819</td>
<td>.797</td>
<td>.927</td>
<td>.904</td>
<td>.960</td>
</tr>
<tr>
<td>Sig. Level</td>
<td>.072</td>
<td>.083</td>
<td>.072</td>
<td>.064</td>
<td>.068</td>
<td>.063</td>
</tr>
</tbody>
</table>

Table 2. Detection Power at $\alpha = .05$ Level of Significance for the Bivariate Case -- Components Separated in Only One Variable

<table>
<thead>
<tr>
<th>Sig. Level</th>
<th>$n = 30$</th>
<th></th>
<th>$n = 40$</th>
<th></th>
<th>$n = 60$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Vectors</td>
<td>EM</td>
<td>Full Vectors</td>
<td>EM</td>
<td>Full Vectors</td>
<td>EM</td>
</tr>
<tr>
<td>(-4,3)</td>
<td>.804</td>
<td>.836</td>
<td>.899</td>
<td>.901</td>
<td>.969</td>
<td>.955</td>
</tr>
<tr>
<td>(-2,3)</td>
<td>.418</td>
<td>.552</td>
<td>.595</td>
<td>.638</td>
<td>.808</td>
<td>.786</td>
</tr>
<tr>
<td>(4,3)</td>
<td>.609</td>
<td>.681</td>
<td>.654</td>
<td>.723</td>
<td>.814</td>
<td>.847</td>
</tr>
<tr>
<td>(5,3)</td>
<td>.738</td>
<td>.823</td>
<td>.795</td>
<td>.872</td>
<td>.930</td>
<td>.930</td>
</tr>
<tr>
<td>(-1,10.5)</td>
<td>.509</td>
<td>.646</td>
<td>.710</td>
<td>.781</td>
<td>.885</td>
<td>.916</td>
</tr>
<tr>
<td>Sig. Level</td>
<td>.051</td>
<td>.069</td>
<td>.066</td>
<td>.085</td>
<td>.057</td>
<td>.066</td>
</tr>
</tbody>
</table>
Table 3. Detection Power at the $\alpha = .05$ Level of Significance for the Trivariate Case using AIC to Select the Number of Components where at most Two Components are Allowed

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>-1.5</th>
<th>2.0</th>
<th>5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Full Vectors</td>
<td>EM</td>
<td>Full Vectors</td>
</tr>
<tr>
<td>-1.5</td>
<td>2.0</td>
<td>.147</td>
<td>.277</td>
<td>.629</td>
</tr>
<tr>
<td>-1.5</td>
<td>5.5</td>
<td>.626</td>
<td>.836</td>
<td>.510</td>
</tr>
<tr>
<td>5.5</td>
<td>2.0</td>
<td>.940</td>
<td>1.000</td>
<td>.894</td>
</tr>
<tr>
<td>-1.5</td>
<td>2.0</td>
<td>.568</td>
<td>.841</td>
<td>.613</td>
</tr>
<tr>
<td>2.0</td>
<td>5.5</td>
<td>.666</td>
<td>.914</td>
<td>.121</td>
</tr>
<tr>
<td>5.5</td>
<td>2.0</td>
<td>.915</td>
<td>.999</td>
<td>.606</td>
</tr>
<tr>
<td>-1.5</td>
<td>5.5</td>
<td>.893</td>
<td>.997</td>
<td>.841</td>
</tr>
<tr>
<td>5.5</td>
<td>2.0</td>
<td>.849</td>
<td>.993</td>
<td>.324</td>
</tr>
<tr>
<td>5.5</td>
<td>5.5</td>
<td>.921</td>
<td>.999</td>
<td>.586</td>
</tr>
</tbody>
</table>

1000 replications of size $n = 100$
AIC selection for $m \leq 2$
Missing probability is .5 for all 3 variables
SE for tabled estimates of power is .016
Table 4. Trivariate Case using AIC to Select the Number of Components in which at most Three Components are Allowed

<table>
<thead>
<tr>
<th></th>
<th>Z</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.5</td>
<td>2.0</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Vectors</td>
<td>EM</td>
<td>Vectors</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>-1.5</td>
<td>-1.5</td>
<td>.150</td>
<td>.267</td>
<td>.611</td>
</tr>
<tr>
<td>-1.5</td>
<td>2.0</td>
<td>.613</td>
<td>.785</td>
<td>.547</td>
</tr>
<tr>
<td>2.0</td>
<td>2.0</td>
<td>.619</td>
<td>.783</td>
<td>.109</td>
</tr>
<tr>
<td>2.0</td>
<td>5.5</td>
<td>.902</td>
<td>.974</td>
<td>.603</td>
</tr>
<tr>
<td>5.5</td>
<td>2.0</td>
<td>.825</td>
<td>.930</td>
<td>.323</td>
</tr>
<tr>
<td>5.5</td>
<td>5.5</td>
<td>.922</td>
<td>.970</td>
<td>.566</td>
</tr>
</tbody>
</table>

1000 replications of size $n = 100$
AIC selection for $m \leq 3$
Missing probability is .5 for all 3 variables
SE for tabled values is .016
Table 5. Simulation based on Nuclear Monitoring Setting Described in Section 3c

<table>
<thead>
<tr>
<th>Explosion</th>
<th>Hypothetical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Vectors</td>
</tr>
<tr>
<td>30</td>
<td>.538</td>
</tr>
<tr>
<td>50</td>
<td>.669</td>
</tr>
<tr>
<td>75</td>
<td>.781</td>
</tr>
<tr>
<td>100</td>
<td>.849</td>
</tr>
</tbody>
</table>

1000 replications
AIC selection for $m \leq 2$
Missing probability is .5 for each variable
SE for tabled values is .016
Figure 1. Mixture Distribution and Outlier Means for Example 3a with some separation between the Components in each Variable.

Figure 2. Mixture Distribution and Outlier Means for Example 3a with separation between the Components in only One Variable.
Figure 3. Mixture Distribution and Outlier Means for Trivariate Case in Example 3b
Figure 4. Solid Contours Showing the Components of the Mixture Distribution of Shallow and Deep Earthquakes along with Dashed Contours for the Explosion and Hypothetical Outlier Populations