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

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad 1)$$

where m is the number of components in the mixture, $f_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the MVN density with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ associated with the ith component, the λ_i, $1, \ldots, m$ are the mixing proportions, and $\boldsymbol{x}$ is a d-dimensional vector of variables.

Letting the training sample be denoted by $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and the new observation (whose distribution is unknown) by $\boldsymbol{X}_u$, then we wish to test the hypotheses

$$H_0 : \boldsymbol{X}_u \in \Pi$$

$$H_1 : \boldsymbol{X}_u \notin \Pi$$

where Π 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 $\boldsymbol{X}_u \in H_0$) is denoted by $L_0(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is an unknown vector-valued parameter associated with the distribution of $\boldsymbol{X}$ under H_0. Likewise, let $\widetilde{L}_1(\boldsymbol{\theta}) = \prod_{s=1}^{n} f(\boldsymbol{X}_s; \boldsymbol{\theta})$ denote the likelihood based only on the training sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. Wang, et al. (1997) and Sain, Gray, Woodward, and Fisk (1999) used the modified likelihood-ratio test statistic

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

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

It is easily seen in (2) that if $\boldsymbol{X}_u$ does not belong to Π, then W will tend to be small. Hence the rejection region is of the form $W \leq W_\alpha$ for some W_α picked to provide a level α 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_α. Whenever some of the training data are unlabeled, the parameters λ_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ 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 $\boldsymbol{X}_u = (X_{u1}, X_{u2}, ..., X_{ud})'$, where X_{uj} denotes variable j observed on the new observation. We further assume that there exists a training sample

$$\boldsymbol{X}_1 = (X_{11}, X_{12}, X_{13}, \ldots, X_{1d})'$$

$$\boldsymbol{X}_2 = (X_{21}, X_{22}, X_{23},$$

$$\boldsymbol{X}_n = (X_{n1}, X_{n2}, X_{n3},$$

from Π. When some of the training data includes missing data, the training sample has the general appearance

$$\boldsymbol{X}_1 = (X_{11}, \quad , X_{13}, \ldots X_{1j}, \quad , X_{1,j+2}, \ldots X_{1d})'$$

$$\boldsymbol{X}_2 = (-, X_{22}, X_{23}, \ldots X_{2j}, X_{2,j+1}, \quad \ldots X_{1d})'$$

$$\boldsymbol{X}_n = (X_{n1}, -, X_{n3}, \ldots X_{nj}, X_{n,j+1}, X_{n,j+2}, \ldots X_{nd})'$$

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 $< n$) 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 (Π) 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 Outlie Testing Api when So D issing

this be the cedure fo to tes hether
hse X II. the po ulatio th traming data W consid ths
whic all labe th aining sample missing, the umber ompo unkn
and the bservat ec ors ha missin mformat

(a) of he M hm E Mode Par

The algorithm eral ce taining likelih
ti ute: th data are missing co nly used th estima
parame h labels missing he licatic considere here the
sq rithm nly eal th missing labe ut also ing da the bservat
tors. The M alg rithn may be tho gh as Prmal pro edur fo performing th
ng uit appealing app to ealing ng data.
missing val b stimate th parameter using the missing val stimat
these param ter imates, re th estimat ng al find
the paramete using the sed val stimates, tc. this
eral ed co ice

Th M alg rithm directly maximizin likelihood func As such,
ft ssary actually tima ng data val rathe nly
sary im th suffi statisti base ng and aila le da tha
ded the aluatio th likelih fim The al thm thus has step
the *expe ep (E ep,* hic the ndi nal xpectatio th uffici
statistic aila data the timat th are cal d,
and th *ep M-step,* hic the lo lik lihood functio 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, $\boldsymbol{Y}$, 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 $y_{i'k} = 0$ for $i' \neq i$. Now, for the kth observation we have

$$\begin{aligned} &\mathrm{P}(Y_{ik} = 1) = \lambda_i \\ &(\boldsymbol{X}|Y_{ik} = 1) \sim \mathrm{MVN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \end{aligned}$$

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

$$\begin{aligned} N_i &= \sum_{k=1}^{n} Y_{ik} \\ S_i &= \sum_{k=1}^{n} Y_{ik}\boldsymbol{X}_k \\ SS_i &= \sum_{k=1}^{n} Y_{ik}\boldsymbol{X}_k\boldsymbol{X}'_k \end{aligned} \tag{3}$$

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

$$\widehat{\lambda}_i = \frac{N_i}{n}$$

$$\hat{\mu}_i = \frac{S_i}{N_i} \qquad (4)$$
$$\hat{\Sigma}_i = \frac{SS_i}{N_i} - \hat{\mu}_i \hat{\mu}_i',$$

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 jth iteration of the algorithm. Then, the conditional expectations calculated in the E-step are :

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

$$\mathrm{E}\left[S_i \mid \hat{\theta}^{(j)}, X_{obs}\right] \qquad (5)$$

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

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, $m = 1,$, 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 (Δ_{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 ($L_{max}(m)$) using the procedure described in Section 2(a) above

AIC and BIC are calculated using the formulas

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

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

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_{AIC} or m_{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 $\widehat{\lambda}_i$'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_α to be the (100α)th percentile of the W_b^*'s . Reject H_0 and conclude that the $(n + 1)$st point is an outlier if $W \leq W_\alpha$.

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, and 60. In the first setting, we let the training data be from a mixture where $\mu_1 = (0, 0)'$, $\mu_2 = (6, 4)'$, $\Sigma_1 = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$, and $\Sigma_2 = \begin{pmatrix} 1 & -.5 \\ -.5 & 1 \end{pmatrix}$. The outlier population is MVN(μ_o, Σ_1) where μ_o takes on the values $\mu_o = (0,5)'$, $(1,4.5)'$, $(2,4)'$, and $(5,8.5)'$. In the second scenario, the training data are from a mixture where μ_1, Σ_1, and Σ_2 are as before and where $\mu_2 = (0, 6)'$. In this case we consider outlier populations that are MVN(μ_o, Σ_1) where $\mu_o = (-4,3)'$, $(-2,3)'$, $(4,3)'$, $(5,3)'$ and (-

1,10.5)′. 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 MVN(μ_0, Σ_1). Denoting the ith 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_{mis} 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_{mis}, the expected number of vectors for which all of the observations are available, is given in the case of d variables by

$$n_F = \frac{n(1-p_{mis})^d}{(1-p_{mis}^d)}.$$

The simulations shown in Tables 1 and 2 are based on the case in which p_{mis} .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 ed that the the alg rithm enhanced the likelihood that AIC uld orrectly ho t mponents Whe using only the full ectors. IC icked compo ents about 0% the h while the EM approach using all vailable data esults compo being selected about 90% the ulated sample It should he oted that this the expected number f full vectors and th surprising that AI ften lects nly mponent. When 40 AIC tends pick mponents bo 3% of the sample using m full vectors and about the samples using the EM algorithm approach. Wh 60 AIC co ectly selected tw mponents er 3% f the time fo each

In Table that all ailable data using the EM algorithm gives substantially better tectio po all cases In Table not see th marked mpr ement etectio po using the EM algorithm tha Table Examinati Figure sho that the two mponents the mixture differ only with respe to nc nd ariable That is, if the second ariable missing bservatio then th really nformatio fr th first ble oncerning en mpon membership the bservatio

It lear upo reflectio that asking of the algorithm to impro parameter estimatio and conseq tly detecti po the training data from mixture population wher he data unlabeled, the ar assumed to be equal, and th number components unkno The abo disc suggests that the the M should co expec ed to vid improved power h the components f he mixtur accompanie by separatio both the ariabl the bivariate

Tables and sho that the bserved significance rate so what high fo the EM th the smalle sampl sizes based tendency fo the significance le using both techni ues be slightly larg than the ominal el fo 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 (μ_i, Σ_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 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 Σ_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 α .05 level of significance. The missing data probability is p_{mis} 0.5 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

compo its simular to that Tables and 4 fo the full :tor appr Using the EM procedur BIC pick compo and cks three mponent only abo the sampl Thus, surprisingly nt etect po fo th EM app in this simular to those Table Thus, this setting C be preferre er AIC. The bserved significanc els based 000 nsdicatio fo the full ector an EM app ach respectively

A ew lected ulatio results at p_{mis} here CO expecte number full ecto sh that the EM algorithm substantially detectio wer the full vect approac this although as dramatically the p_{mis} sho Tabl -4.

(c) Simulatio bas d Se Data

In the pr bl detectio f uclear exp certain ariab calculated from the associated ith the ent sti Training data these variabl llected on non-nu lear seismic events. ns bserved, the probl of ts est to ecide hether should be onsidered ther m the popula he training data and therefo potentially lear T lese variables variables tha he sured by sta veral thousand kil meters the ar ed to depth based hypo calculatio and *(i)* the difference between body and surface magni udes simplicity will deno these variables depth m_b M respectively In this example, the mixture model the training data to eep earthquake D Q and shall earthquakes (Q) surprising that eep earthquakes end ha larger depth estimates than either shallo earthquake lear xpl ns, whil the depths of shallo earth uakes and exp similar each the The alue er ariable m_b end ai larger fo deep earthquake and explosi han for hall ow earthquakes explo pro low

mag surface and the surfa magnitudes f earthquak tend to ecrease as depth increases. A tual alue: of these ariables databases of earthquakes and exp ons tend ze classified H in igure illustrate th situation describe bove scaled versions the ariabl showing iative positioning th The large ed with epth deep earth uakes flects the fact that eep earthquak may be many kilomet th while the shallo earthquak by definitio earth uakes relat ely lo the surface

In ur ulations, onsid tw separate utlier populations These two populati ns illustrated Fig with tw styles f dashed onto The ther populatio ite the lo right igure and denoted by EX epresentativ f the locatio xpl The ut po ulatio denoted by "HYP" used for illustrati urpo only and ill be eferre the hypothetical utlier po ulation. We co der the distributio composed eep and shallo earthquake to be mode composed equal mixture the t components and we assume that 0% the bservations missing W also tha bserved from the uther popula ulati 000 replications fr this scenario ty samp and the proportion of the replicates for which ther de ected giv in Table In th imulati ns, AIC used select between and omponents. W sh simula esults fo the tw utlier populatio nsidered In Table an be tha the M approach ga consistently higher detecti po er than that btained using full vect the explo popula the improvement using the EM most pr unced at the hypothetical utlier populati n, there ery pronounced impr ement using the EM ith and Examinatio the sho that he full results in AIC correctly selecting distribut nly abo 0% of the time hile fo this sample and th M appr h, AIC selects omponent mixture 90% f the igure 4 it be seen that impr per 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 3*c*. Example 3*b* 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

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.

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis, Fourth Edition,* Upper Saddle River, New Jersey: Prentice Hall.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data.* New York: John Wiley and Sons, Inc.

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

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, New York: John Wiley and Sons, Inc.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*, New York: John Wile and Sons, Inc.

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.

Redner, R.A. and Walker, H.F. (1984). "Mixture Densities, Maximum Likelihood and the EM Algorithm." *SIAM Review*, 26, 195-239.

Sain, S.R., Gray, H.L., Woodward, W.A., and Fisk, M.D. (1999). 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., Zhao, B., and Fisk, M.D. (2002). Testing for Multivariate Outliers in the Presence of Missing Data, *Journal of Applied and Theoretical Geophysics*, 159, 889-903.

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

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

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

	$n = 30$		$n = 40$		$n = 60$	
	Full Vectors	EM	Full Vectors	EM	Full Vectors	EM
(-4,3)	.804	.836	.899	.901	.969	.955
(-2,3)	.418	.552	.595	.638	.808	.786
(4,3)	.609	.681	.654	.723	.814	.847
(5,3)	.738	.823	.795	.872	.930	.930
(-1,10.5)	.509	.646	.710	.781	.885	.916
Sig. Level	.051	.069	.066	.085	.057	.066

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

		-1.5		2.0		5.5	
		Full		Full		Full	
X	Y	Vectors	EM	Vectors	EM	Vectors	EM
	-1.5	.147	.277	.629	.839	.942	1.000
-1.5	2.0	.626	.836	.510	.846	.894	.997
	5.5	.940	1.000	.894	.999	.819	.994
	-1.5	.568	.841	.613	.847	.906	1.000
2.0	2.0	.666	.914	.121	.249	.618	.856
	5.5	.915	.999	.606	.864	.321	.411
	-1.5	.893	.997	.841	.990	.915	1.000
5.5	2.0	.849	.993	.324	.482	.593	.806
	5.5	.921	.999	.586	.831	.317	.633

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

		Z					
		-1.5		2.0		5.5	
X	Y	Full Vectors	EM	Full Vectors	EM	Full Vectors	EM
	-1.5	.150	.267	.611	.791	.934	.981
-1.5	2.0	.613	.785	.547	.784	.916	.974
	5.5	.948	.978	.893	.964	.844	.949
	-1.5	.506	.758	.591	.781	.916	.972
2.0	2.0	.619	.783	.109	.252	.614	.789
	5.5	.902	.974	.603	.779	.350	.416
	-1.5	.879	.963	.834	.937	.921	.971
5.5	2.0	.825	.930	.323	.423	.597	.758
	5.5	.922	.970	.566	.727	.284	.582

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 3*c*

	Explosion		Hypothetical	
	Full Vectors	EM	Full Vectors	EM
30	.538	.541	.315	.483
50	.669	.761	.520	.655
75	.781	.821	.647	.719
100	.849	.884	.719	.757

1000 replications
AIC selection for $m \leq 2$
Missing probability is .5 for each variable
SE for tabled values is .016

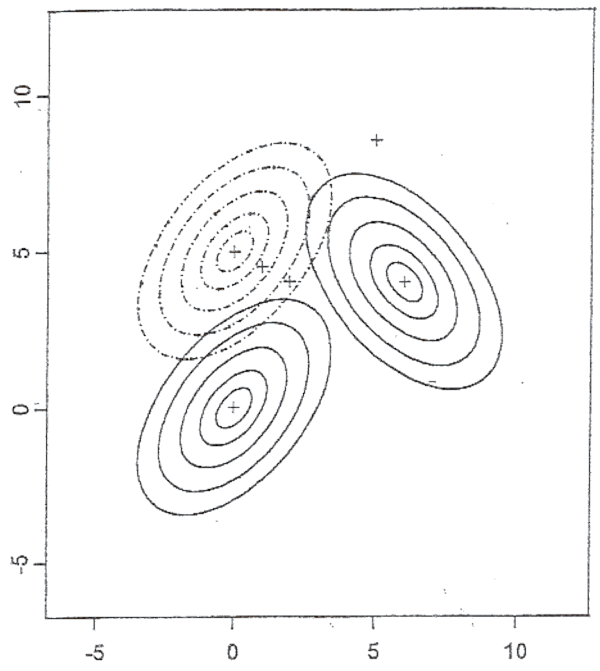

Figure 1. Mixture Distribution and Outlier Means for Example *3a* with some Separation between the Components in each Variable

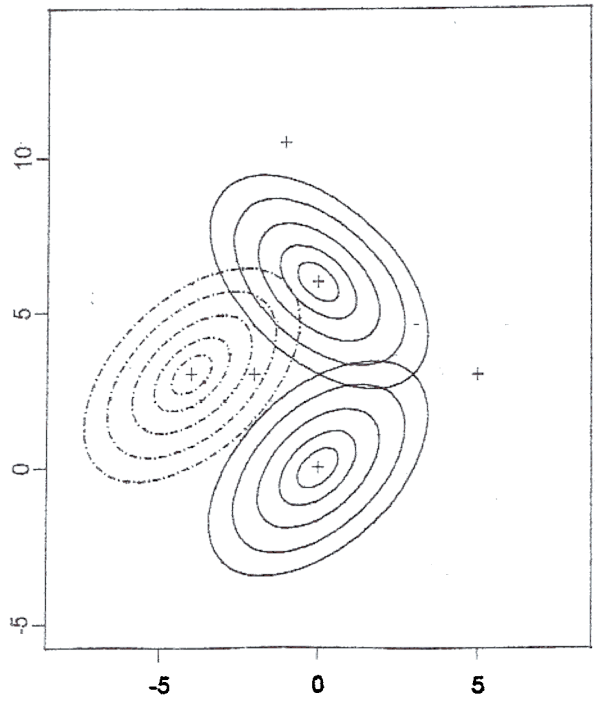

Figure 2. Mixture Distribution and Outlier Means for Example *3a* with Separation between the Components in only One Variable

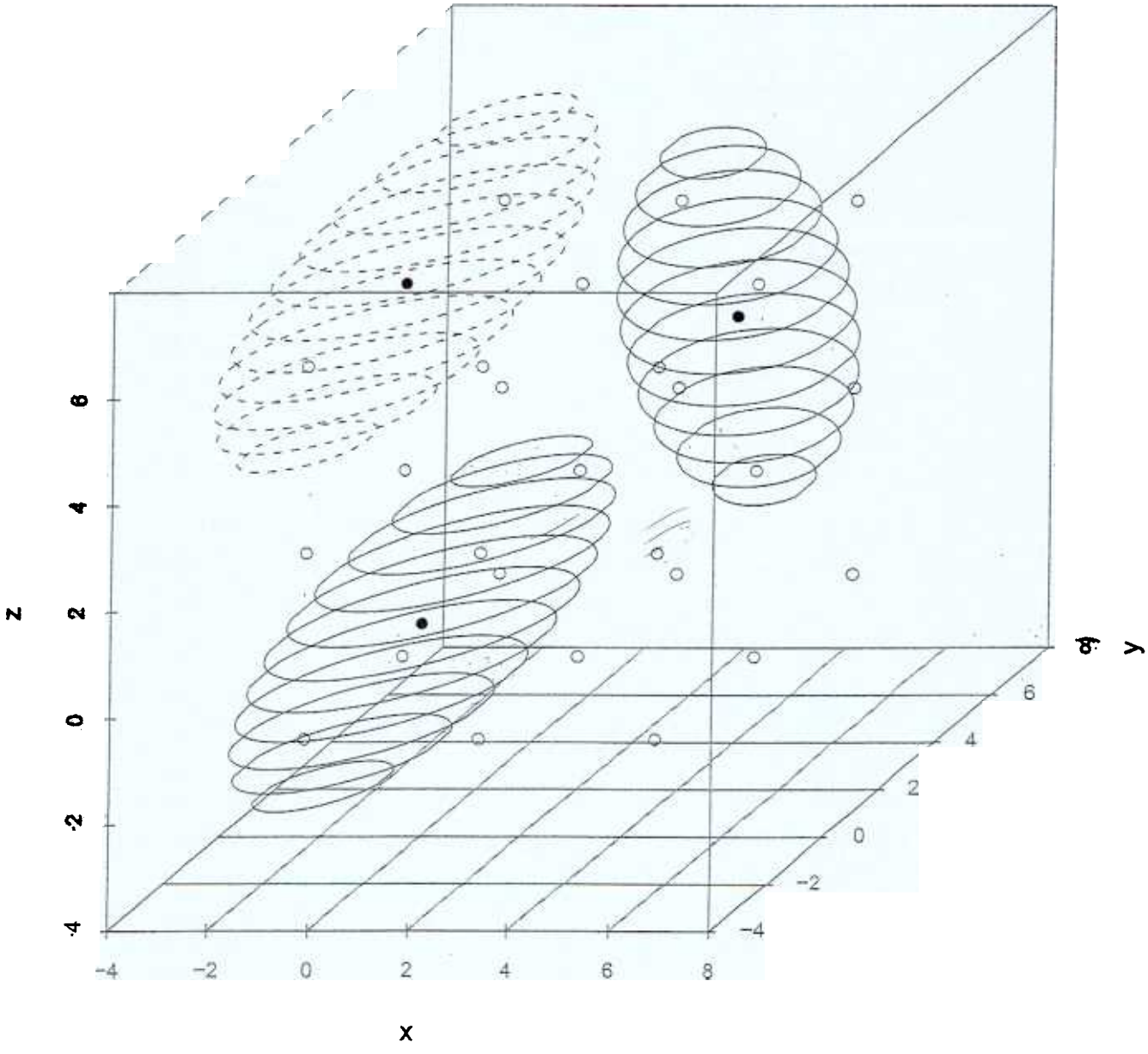

Figure 3. Mixture Distribution and Outlier Means for Trivariate Case in Example *3b*

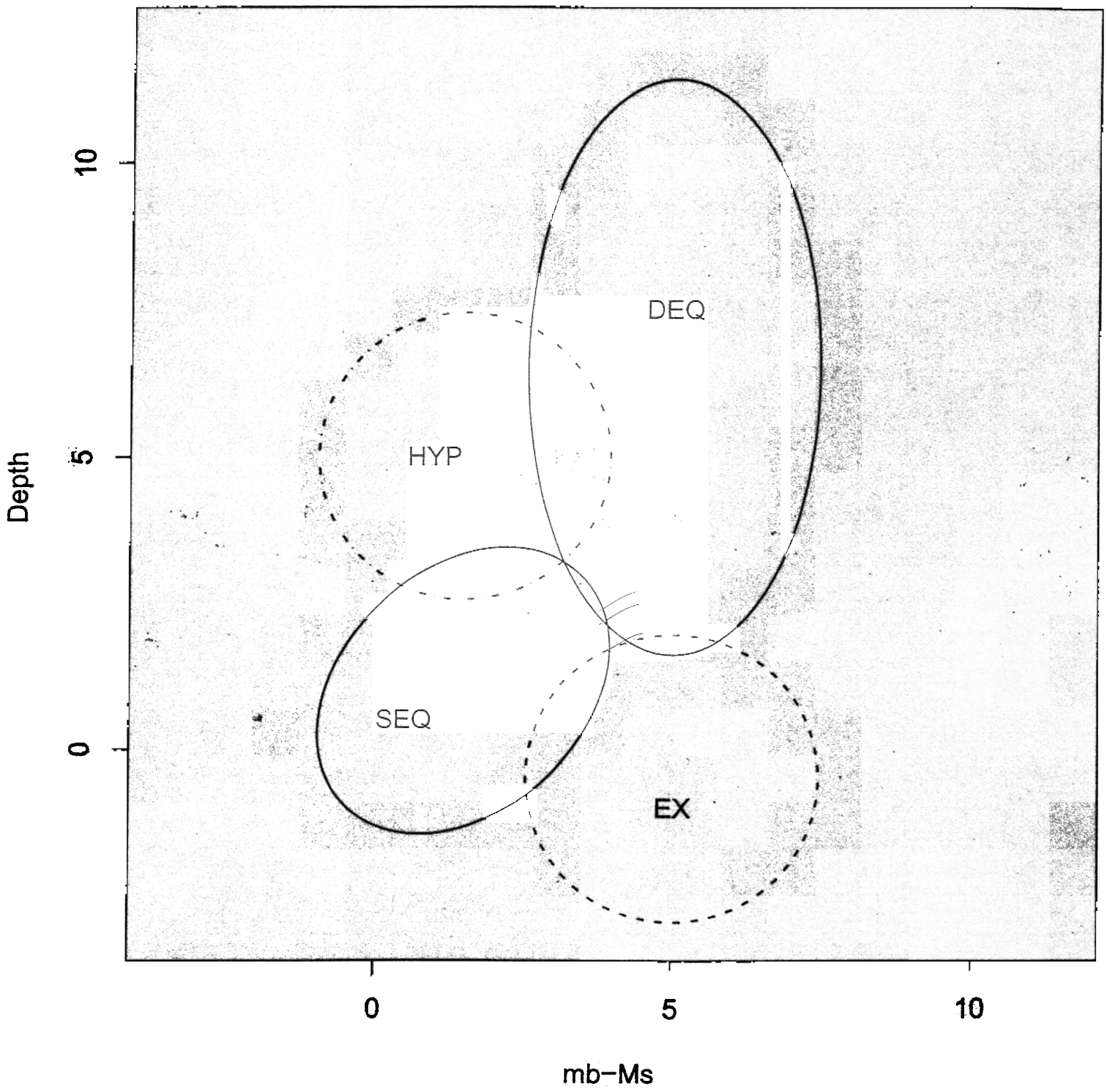

Figure 4. Solid Contours Showing the Components of the Mixture Distribution of Shallow and Deep Earthquakes along with Dashed Contours for the Explosion and Hypothetical Outlier Populations