ENVIRONMETRICS
Environmetrics 2004; **15**: 129–139 (DOI: 10.1002/env.628)

Robust principal component analysis and outlier detection with ecological data

Donald A. Jackson[1,*,†] and Yong Chen[2]

[1]*Department of Zoology, University of Toronto, Toronto, Ontario, Canada*
[2]*School of Marine Sciences, University of Maine, Orono, ME 04469, U.S.A.*

SUMMARY

Ecological studies frequently involve large numbers of variables and observations, and these are often subject to various errors. If some data are not representative of the study population, they tend to bias the interpretation and conclusion of an ecological study. Because of the multivariate nature of ecological data, it is very difficult to identify atypical observations using approaches such as univariate or bivariate plots. This difficulty calls for the application of robust statistical methods in identifying atypical observations. Our study provides a comparison of a standard method, based on the Mahalanobis distance, used in multivariate approaches to a robust method based on the minimum volume ellipsoid as a means of determining whether data sets contain outliers or not. We evaluate both methods using simulations varying conditions of the data, and show that the minimum volume ellipsoid approach is superior in detecting outliers where present. We show that, as the sample size parameter, *h*, used in the robust approach increases in value, there is a decrease in the accuracy and precision of the associated estimate of the number of outliers present, in particular as the number of outliers increases. Conversely, where no outliers are present, large values for the parameter provide the most accurate results. In addition to the simulation results, we demonstrate the use of the robust principal component analysis with a data set of lake-water chemistry variables to illustrate the additional insight available. We suggest that ecologists consider that their data may contain atypical points. Following checks associated with normality, bivariate linearity and other traditional aspects, we advocate that ecologists examine their data sets using robust multivariate methods. Points identified as being atypical should be carefully evaluated based on background information to determine their suitability for inclusion in further multivariate analyses and whether additional factors explain their unusual characteristics.

KEY WORDS: principal component analysis; outliers; robust statistics; water chemistry; environmental data; multivariate analysis

1. INTRODUCTION

Species and their composite, ecological communities, are the result of a combination of biotic and abiotic conditions. The interaction of various environmental factors has long been recognized to influence the species combinations. As a result of this complexity, ecologists often use multivariate techniques to determine the relationship between the various taxa, the various environmental

*Correspondence to: Donald A. Jackson, Department of Zoology, University of Toronto, Toronto, Ontario, Canada.
†E-mail: jackson@zoo.utoronto.ca

Received 15 December 2001
Accepted 12 April 2003

conditions and the association between these two sets of variables. This necessitated that ecologists be at the forefront in the development and application of many multivariate applications to summarize the complexities of the patterns encountered. When examining environmental data (but typically not species abundance data that show non-linear relationships), it is frequently assumed that the data follow well-behaved statistical distributions and relationships. The data are often assumed to have (or are transformed to approximate) normal distributions and linear relationships between variables. Such assumptions may not be critical where the goal is pattern exploration, but various characteristics of the data may complicate or confound both exploration and hypothesis testing. A principal factor contributing to such problems is the influence of atypical points or outliers. Although it is likely that outliers may exist in many ecological studies, ecologists have directed little attention at the detection of, and dealing with, multivariate outliers relative to the efforts from many other fields (e.g. statistics: Rousseeuw, 1985a,b; geology: Barcelo *et al.*, 1996; chemistry: Egan and Morgan, 1998).

There is a rich literature related to detecting univariate and bivariate outliers or influential points, and many ecologists are aware of these studies and methods as they are covered in standard and specialized texts (e.g. Rousseeuw and Leroy, 1987; Barnett and Lewis, 1993). Many of these approaches rely on graphical techniques (univariate or bivariate plots) for a visual assessment of unusual points or some form of quantitative description. In the simplest case of univariate data, people have employed methods such as identifying points exceeding 3 or more standard deviations from the mean and then consider those points to be 'outliers'. However, moments describing the data, e.g. the mean and variance, are influenced by outliers. This influence may hide or mask true outliers, but also incorrectly lead to the identification of points as being outliers that are representative of the sampled population. Approaches based on medians and quartiles have provided one solution but are not without problems (see Mosteller and Tukey, 1977; or Chen and Jackson, 1995, for discussions). Once one begins to work with bivariate relationships, the complexity and variety of approaches increases greatly. Atypical points may lead to inflated variances as with univariate data, and also alter the covariance or correlation structure (see Figure 1 for an example). With bivariate data, atypical points may follow the same general trend, but simply be more extreme in their location. This simply gives them greater

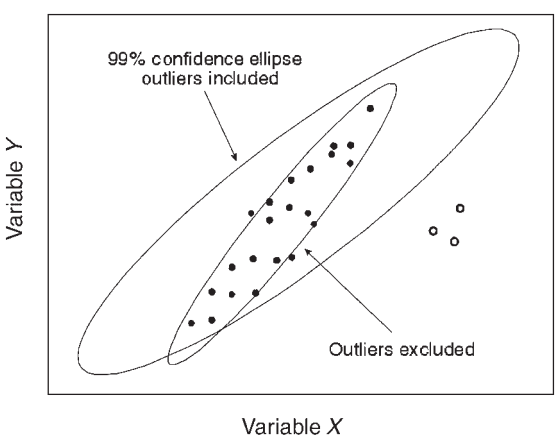

Figure 1. Bivariate scatterplot of simulated data illustrating the effect outliers have on the covariance structure and the associated confidence limits. The outliers (open circles) were identified using a least-median of squares approach. The solid ellipse represents the 99% confidence limits where the outliers are retained and the dashed line is the 99% confidence ellipse where they are removed

influence in determining the direction of the slope and statistical significance in ordinary least-squares (OLS) regression. However, atypical points falling off the general trend may greatly influence the estimates of the slope, intercept and statistical significance. Various measures of influence associated with bivariate data have been developed (e.g. simple residual plots, Cook's D). Once identified as being atypical, a decision can be made about how to weight these data points (i.e. fully weighted, down-weighted or deleted) and recalculate the model using the remainder of the data. Alternatively one can employ different regression models assuming different error structures or weight the data points in a manner different from the OLS. One of the methods shown to have the greatest promise is the least median of squares (Rousseeuw, 1985a; Chen *et al.*, 1994).

An important parameter in dealing with outliers is the breakdown point. This measures the ability of an estimation method to identify the unbiased estimates for parameters with data having outliers. It is defined as the smallest fraction of contamination (e.g. data points not belonging to the general pattern of the population) that can cause the estimator to take on biased values far away from the true estimates (Rousseeuw and Leroy, 1987; Barnett and Lewis, 1993). Clearly the maximum value that can be achieved for a breakdown point is 50% as, when more than 50% of the data are contaminated points (i.e. 'outliers'), it is impossible to distinguish the 'good' from the 'bad' parts of the data. Methods based on OLS estimators typically have a breakdown point near zero and are, therefore, subject to giving poor model estimates with even a few outliers in the data. In contrast, methods based on robust estimators have been shown to be effective, with large proportions of the data being atypical points. The least median of squares has been shown to be effective with nearly 50% contamination (see Chen *et al.*, 1994, for details and comparisons of methods).

In making the transition to multivariate data, one can no longer view the full set of data directly. Plotting of data is restricted to two- or three-dimensional plots of either the original variables or axes resulting from summarization methods (e.g. principal component analysis, PCA). The multivariate nature of the data and the covariation of variables may mask some outliers or suggest other points as atypical (Becker and Gather, 1999; Caroni, 2000; Pell, 2000). Different covariance patterns for the outliers may not be visible in simple univariate or bivariate plots of the original environmental variables. Due to masking from biased estimates of covariance or correlation structure, outliers may not be visible in plots from multivariate analyses. For these reasons, even diligent researchers carefully checking their data prior to analysis may not recognize contaminated data. Alternatively they may discard observations that they believe are 'bad' observations and have been erroneously classified as such due to biased measures. Given the large numbers of observations and variables used in community and environmental studies, it is likely that some 'unusual' data points may appear in the observations. Therefore there is a need for more robust measures to identify these atypical points and allow researchers to decide whether to retain them, or classify them as true outliers (see Chen and Jackson, 1995, for a discussion of what are true outliers) and remove them. So ecologists must be prepared to identify such points and determine how they will be treated subsequently (e.g. see Filzmoser, 1999).

Our study provides a comparison of a standard method used in multivariate approaches to a robust method in order to determine whether data sets contain outliers or not. The standard method is based on the squared Mahalanobis distance calculated using the covariance matrix. Our robust method is based on the minimum volume ellipsoid (Rousseeuw, 1985a,b) that is a multivariate extension of the least median of squares. We have selected these two methods as they represent what is the most common or standard method based on the Mahalanobis approach with what has been proposed as the most robust method available. We examine the robust, minimum volume ellipsoid in greater detail to determine the impact that the choice in the size of the subset sampled has on the reliability and sensitivity of the approach. We evaluate both methods using simulations varying conditions of the

data. We also demonstrate the use of the robust method with a data set of lake water chemistry to illustrate the additional insight available.

2. METHODS

2.1. Outlier detection methods

The commonly used method for identifying outliers in multivariate analysis is based on the squared Mahalanobis distance. For a data matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{ij} & \cdots & x_{1p} \\ . & . & . & . & . \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ . \\ \mathbf{X}_i \\ . \\ . \\ \mathbf{X}_n \end{pmatrix}$$

the squared Mahalanobis distance is calculated as

$$MD^2(\mathbf{x}_i, \mathbf{X}) = (\mathbf{x}_i - T(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - T(\mathbf{X}))^{\mathrm{t}}$$

for each observation, where $T(\mathbf{X})$ is a multivariate location estimator (in this case it is the arithmetic mean) and $\mathbf{C}(\mathbf{X})$ is the classical covariance estimate, with the denominator being $n-1$ rather than n, and $\mathbf{x}_i$ being the vector for sample i, and they are calculated as

$$T(\mathbf{X}) = \bar{\mathbf{x}}$$

$$\mathbf{C}(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - T(\mathbf{X}))^{\mathrm{t}}(\mathbf{x}_i - T(\mathbf{X}))$$

Points for which $\mathrm{MD}^2(\mathbf{x}_i, \mathbf{X})$ is large are identified as atypical points or outliers and evaluated using the χ^2 distribution with the appropriate degrees of freedom.

A robust method, the minimum volume ellipsoid (MVE), was proposed to identify outliers in estimating means and covariance for multivariate data by Rousseeuw (1985b). The algorithm for the MVE can be summarized as follows:

1. For a multivariate data matrix $\mathbf{X}$ (as described before), draw a subsample of $p+1$ (p is the number of variables in the $\mathbf{X}$ matrix) different observations, indexed by $\mathbf{J} = (j_1, \ldots, j_{p+1})$, and calculate the arithmetic mean and the corresponding covariance matrix as

$$\bar{\mathbf{x}}_{\mathbf{J}} = \frac{1}{p+1}\sum_{i\in\mathbf{J}}\mathbf{x}_i$$

$$\mathbf{C}_{\mathbf{J}} = \frac{1}{p}\sum_{i\in\mathbf{J}}(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathbf{J}})^{\mathrm{t}}(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathbf{J}})$$

   where $\mathbf{C}_{\mathbf{J}}$ is non-singular.

2. Calculate $m_{\mathbf{J}}^2$ as

$$m_{\mathbf{J}}^2 = \text{med}\left[(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathbf{J}})\mathbf{C}_{\mathbf{J}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathbf{J}})^{\text{t}}\right]_{h:n}$$

where med is the median for $i = 1$ to n, $h = (n + p + 1)/2$; the above computation corresponds such that the ellipsoid is inflated or deflated to contain exactly h points (out of n points).
3. Calculate $P_{\mathbf{J}} = (\det(m_{\mathbf{J}}^2\mathbf{C}_{\mathbf{J}}))^{0.5}$, which is proportional to the resulting ellipsoid.
4. Repeat the above procedure for many subsamples $\mathbf{J}$, and the one with the lowest $P_{\mathbf{J}}$ is retained.
5. Then compute $T(\mathbf{X}) = \bar{\mathbf{x}}_{\mathbf{J}}$, and $\mathbf{C}(\boldsymbol{X}) = c^2(n,p)(\chi^2_{p,0.50})^{-1}m_{\mathbf{J}}^2\mathbf{C}_{\mathbf{J}}$, where $c^2(n, p)$ is a small-sample correction term calculated as $[1 + 15/(n\text{-}p)]^2$ and $\chi^2_{p,0.50}$ is the median of the χ^2 distribution with p degrees of freedom. It is apparent that intensive sampling and computation are required to find the solution in the MVE analysis. The total amount of subsampling depends on the n and p. Rousseeuw and Leroy (1987) identified 5000 subsamples as being sufficient. However, based on our preliminary results, we increased this to 50 000 to enhance the stability of our results similar to the findings of Jackson and Somers (1989). Based on the MVE-estimated mean $\mathbf{T}(\mathbf{X})$ and covariance $\mathbf{C}(\mathbf{X})$, the following statistic, similar to MD^2, can be calculated:

$$W_i = (\mathbf{x}_i - \mathbf{T}(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))^{\text{t}}$$

for an observation i, where if $W_i > \chi_{p,0.975}$ it is defined as an outlier; otherwise it is considered to be a 'normal' observation following the approach outlined in Rousseeuw and Leroy (1987). The PCA was conducted with MVE-defined outliers and normal data having weights of 0 and 1, respectively, that we identify as a reweighted PCA.

2.2. *Data simulation*

We examined the influence of various characteristics of the data and the choice of h in the MVE through a series of simulations. We chose the following approach:

(i) Each data set comprised 50 observations by 4 variables and was simulated following a normal distribution for each variable for each of the two covariance matrices listed below:

Covariance matrix 1				Covariance matrix 2			
4	2	3	3	4	4.24	4.5	3
2	8	2	3	4.24	8	6.36	4.24
3	2	9	1	4.5	6.36	9	4.5
3	3	1	4	3	4.24	4.5	4

(ii) A level of contamination for each data set was set at one of the following levels: 0, 5, 15 or 25 of the 50 observations. This provided a range of contaminated data between 0 and 50%.
(iii) The mean for each variable in the sample was set at 20, with the mean for the contaminated values set at one of 12, 20 or 40.
(iv) The variance of the contaminated values was set at either 0.1 or 0.5.
(v) The values of h were set at 20, 25, 27, 30 or 35, where 27 would be the level based on Rousseeuw's formula listed above.

This provided a total of 240 scenarios based on the combination of all parameter settings. For each scenario, 100 data sets were simulated. These were then analyzed to assess the number of outliers

present using the squared Mahalanobis distance approach and the robust statistical approach from the MVE. Each of the 100 data sets for each scenario was subsampled 50 000 times to estimate the MVE.

An example was based on the water chemistry data for 34 lakes from the Black River watershed in south-central Ontario (Jackson, 1988). Environmental variables including pH, sodium, potassium, chloride and conductivity were used in a PCA based on the correlation matrix.

3. RESULTS AND DISCUSSION

When the data were free of contamination, i.e. no design outliers were present, the Mahalanobis distance method proved to be a more reliable measure of the number of outliers present, compared with the MVE. This assessment provided a consistent estimate of a single outlier in the data set (Figure 2; these graphical results are based on using covariance matrix 1, but those from covariance matrix 2 show similar effects). In contrast the MVE measure provided estimates ranging from 0.38 to 8 outliers as being present where none were simulated. This variation was due to the setting of the parameter h in the MVE, and the accuracy of the estimator increased as the value for h was increased. The average number of outliers was estimated at 1.54 when $h = 27$, providing an estimate slightly greater than that obtained using the Mahalanobis distance method. In MVE-based scenarios having $h \geq 27$, the results were comparable to those based on the Mahalanobis distance approach.

Introducing a low number of outliers into the data set led to a separation of the two approaches based of their performance, and this differentiation increased as the degree of contamination was increased. With the number of outliers set at 5 (i.e. 10%), all of the Mahalanobis-based results

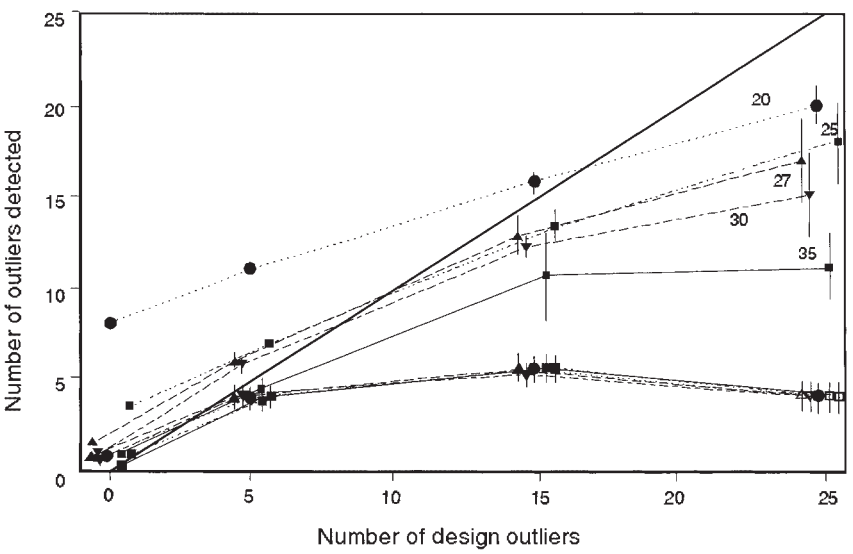

Figure 2. Plot showing the average number of outliers detected as a function of the number of design outliers in the simulations and the size of the sample (h) used in the minimum volume ellipsoid approach. The diagonal line has a slope of one, providing a reference for the number of outliers that should be predicted under optimal results. The five top lines show the results from the five levels of h in the MVE-based method, and the five lower lines are based on the Mahalanobis distance approach. The points represent the means, and vertical bars are the standard deviations associated with estimates from all simulations for each level of design outliers

underestimated the number of outliers present within the data sets (Figure 2). In contrast, all the MVE-based methods overestimated the number of outliers, with the exception of when $h = 35$. At this level, the MVE underestimated the number of outliers and provided results similar to those obtained using the Mahalanobis-based method. At the opposite extreme was the result for $h = 20$, which indicated that the average data set contained slightly more than 10 outliers (i.e. 20%).

Stepping up the number of design outliers included in the simulations to 15 (i.e. 30%) now clearly separated the performance of the two approaches. The number of outliers identified by the Mahalanobis-based method was comparable to that found where only 5 outliers were included (i.e. approximately 5, versus 4 for the mean levels detected). The MVE-based approach showed that when $h = 20$ the results were very close to those designed into the simulation, and all other levels of h underestimated the true level of contamination. The results varied from approximately 10 outliers detected where $h = 35$ to 14 outliers detected where $h = 25$. There were considerable differences in the mean number of outliers detected for each level of h depending on the settings of other parameters in the model. This was shown clearly by the increased standard deviations associated with the estimates, particularly those when $h = 35$, in contrast to the stable results when $h = 20$.

The most extreme level of contamination was with 25 design outliers being included. This represented 50% of the observations being considered as unrepresentative of the population. At this level, the Mahalanobis distance measure indicated an average of 4 outliers present. This is comparable to the results obtained when only 5 design outliers were included and a decrease from the number detected when only 15 design outliers were included. This shows that the Mahalanobis based method becomes increasingly unreliable at higher levels of contamination. The MVE-based results for $h = 35$ showed an identical mean value, but a smaller standard deviation to those obtained when only 15 design outliers were included. The other levels of h also showed increased numbers of outliers detected and increased standard deviations associated with each level of h where the number of design outliers was 25 rather than 15. There was a decrease in the overall performance of the MVE-based estimates relative to the design as the value assigned to h was increased from 20 to 35, and the standard deviation also increased along this trend.

The principal factors responsible for differences in the number of outliers detected are the number of design outliers included and how distant these outliers are positioned relative to the population (Tables 1 and 2). There was a significant interaction between the mean of the design outliers and the number of outliers for both covariance matrices used when evaluated using the MVE approach (Table 2). However, results varied for the two different covariance matrices when using the Mahalanobis distance approach (Table 1). A significant interaction was found for covariance matrix 1 but not for covariance matrix 2, which had the stronger correlation structure between all variables. In this latter case, only the mean value for the design outliers was a significant factor and the number of design outliers was not a significant explanatory factor of the number of outliers detected. This matched with the results shown in Figure 2. The models explained a greater amount of the total variance for the MVE approach relative to the Mahalanobis method (i.e. approximately 80% versus 40% of the variance explained, respectively). Within the MVE results, the number of design variables also summarized much more of the variation associated with the detected outliers than any other factor.

The standard PCA of the water chemistry data set shows that the first axis strongly summarized the pattern in water conductivity and sodium whereas axis 2 differentiated those lakes having high pH and calcium values from those with high chloride concentrations (Figure 3a). Analyzing this data set with an MVE-based PCA identified six lakes as potential outliers (shown as open circles). Closer examination of these six lakes showed them to have unusually high levels of sodium, and each of these lakes border on major roads that receive winter salting to remove ice. As such, they have been

Table 1. Analysis of variance of the number of outliers identified using the Mahalanobis-distance criterion for the two covariance matrices. The top set of results relates to covariance matrix 1 and the second set to covariance matrix 2. Tables 1 and 2 show only reduced models following the removal of non-significant interactions and main effects

Source	Degrees of freedom	Sum of squares	*F*-value	Associated *P*-value	Model *R*-squared
Model	4	32.53	8.13	0.036	0.402
Error	19	48.32	2.54		
Corrected total	23	80.85			
C(*X*)	1	0.246	0.10	0.759	
Mean	1	1.464	0.58	0.457	
# Outliers	1	4.098	1.61	0.220	
Mean*#outliers	1	26.729	10.51	0.004	
Model	4	36.45	3.00	0.045	0.387
Error	19	57.78			
Corrected total	23	94.23			
C(*X*)	1	0.130	0.04	0.838	
Mean	1	2.109	0.69	0.006	
# Outliers	1	5.404	1.78	0.415	
Mean*#outliers	1	28.805	9.47	0.198	

Table 2. Analysis of variance of the number of outliers identified using the minimum volume ellipsoid criterion for the two covariance matrices. The top set of results relates to covariance matrix 1 and the second set to covariance matrix 2

Source	Degrees of freedom	Sum of squares	*F*-value	Associated *P*-value	Model *R*-squared
Model	4	3360.57	840.14	0.0001	0.793
Error	115	874.90	7.608		
Corrected total	119	4235.47			
C(*X*)	1	2.45	0.32	0.572	
Mean	1	0.17	0.02	0.882	
# Outliers	1	3282.31	431.44	0.0001	
Mean*#outliers	1	75.64	9.94	0.0021	
Model	4	3748.23	134.20	0.0001	0.824
Error	115	803.00			
Corrected total	119	4551.24			
C(*X*)	1	2.307	0.33	0.567	
Mean	1	0.069	0.01	0.921	
# Outliers	1	3709.23	531.21	0.0001	
Mean*#outliers	1	36.64	5.25	0.0238	

altered from the standard population of lakes in the region. Excluding these points and re-running the standard PCA provided a different pattern of lakes and association of the variables with the components (Figure 3b). The first component is now strongly correlated with pH, conductivity, calcium and sodium. This is a pattern of positive association between pH, cations and conductivity that is commonly seen in low conductivity PreCambrian Shield lakes. The second axis differentiates the lakes on the basis of their relative chloride concentrations.

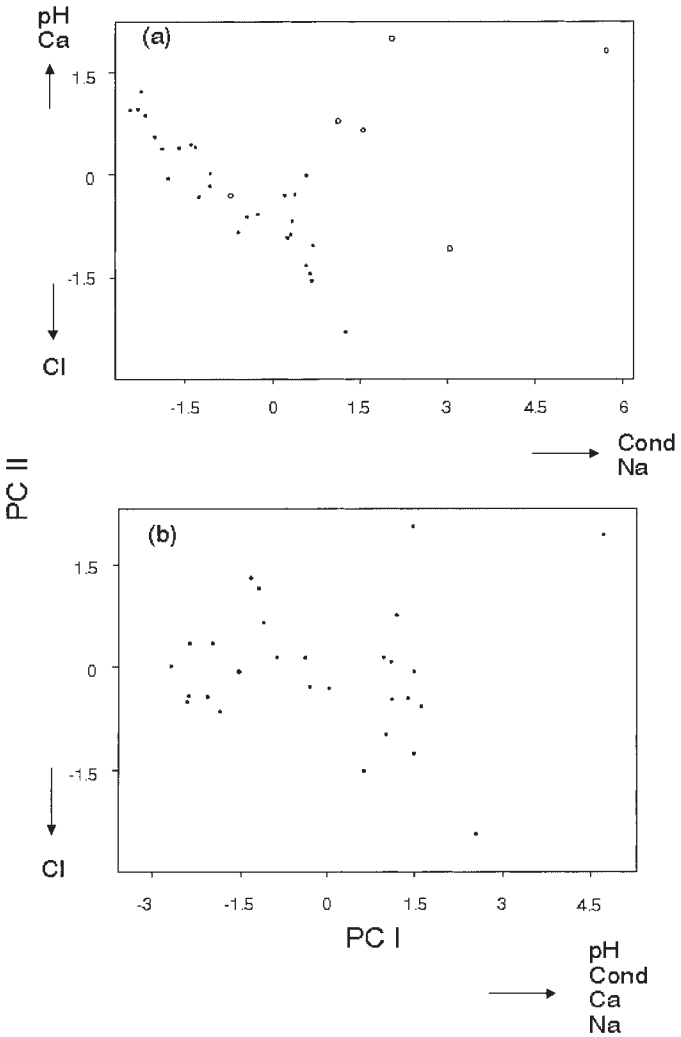

Figure 3. (a) Scatterplot of axes 1 and 2 from a standard principal component analysis of water chemistry variables from 34 lakes. Variables having strong loadings on either axis 1 or 2 are shown associated with them. Points shown with an open circle are those identified as atypical using the MVE-based PCA. These points were examined, removed and the standard PCA recalculated on the remaining 28 lakes and shown in (b)

Although several of these six lakes may have been noted as being somewhat atypical in a standard PCA, at least one of these points would have escaped detection. One point specifically falls within the general cloud on the first three axes and may not have been noticed at all. Even had some of the points been recognized and considered outliers, the explanation of the common cause, i.e. road salt, may not have been determined. As a result one would not have known whether these lakes accurately represented: part of the natural range of variation; samples contaminated in the field; inadequately analyzed samples in the laboratory; or whether these lakes actually represented a different population. The MVE-based PCA suggests the latter explanation as being the most likely.

In general the performance of the Mahalanobis-distance-based approach showed a low breakpoint. It underperformed the MVE-based approach whenever outliers were present and it rapidly failed in detecting outliers as the degree of contamination increased. In fact, when contamination levels continued to increase, there was actually a decrease in the number of outliers detected. In contrast, the MVE-based approach was much more successful in detecting outliers, but still tended to underestimate the number present when they represented more than about 10% of the data values. These results about the higher breakpoint for MVE-based methods are consistent with others' findings (e.g. Chen *et al.*, 1994; Seaver and Triantis, 1995: Kosinski, 1998; Marden, 1999).

When there are no atypical points in the data sets, the Mahalanobis distance measure provides the best method of assessment. However, this technique quickly underestimates the number of atypical points when they are present. This effect becomes most pronounced as the number of outliers is increased. The sensitivity of this method in detecting outliers depends on the conditions of how different the mean of the outliers is relative to the remainder of the sample and the different covariance matrix structures. The minimum volume ellipsoid (MVE) approach proved to be superior to the Mahalanobis distance approach in all cases where outliers were present. However, where outliers were absent in the data, the MVE method tended to identify some observations as being atypical. In most instances the examination of the plots from a PCA based on the MVE provide additional insight as to whether these atypical points really appear different or not. In these cases where no design outliers were present, one can typically recognize this condition from the MVE-generated PCA plots. Closer examination of these points and their underlying variables provides aids in determining whether these points should be considered true outliers and removed from the analysis. This approach was taken with the water chemistry example. We are not advocating the automatic removal of any points considered to be atypical without careful examination of them (see Chen *et al.*, 1994; Chen and Jackson, 1995, for further discussion). By carefully examining the identified observations, and considering independent measures or information, one may be able to determine whether the points in question are true outliers or simply extreme observations (see Chen and Jackson, 1995, for a discussion), thereby allowing analysts to better determine how such problematic observations should be treated. Although the minimum volume ellipsoid approach overestimates the number of atypical observations when none or few are present, proper consideration and evaluation of these identified points should allow one to exclude those that are not correct and retain those that should be included. However, the failure of the Mahalanobis distance method to identify true outliers is of greater concern as no further consideration of these observations is possible, nor is any corrective action taken.

The choice in the value of h in the MVE appears to have a considerable effect on its error rate in detecting outliers. At low values of h, the technique has a high Type I error rate, but also has more power in detecting outliers when they are present, particularly when they represent large proportions of the data set. High values of h provide low Type I error rates, but have more limited power in detecting outliers where present. However all levels of h provide superior outlier detection than the Mahalanobis distance method when outliers were present. Rousseeuw and Leroy (1987) advocated that the value of h should be just slightly above half the total for the number of observations and variables (i.e. $h = [n + p + 1]/2$). Based on our simulations this provides a reasonable compromise between the error rates associated with either higher or lower values of h.

4. CONCLUSIONS

We emphasize that researchers should consider that their data may contain atypical points, and our experience with many ecological and environmental data sets suggests that they occur frequently. Such

statements of concern regarding outliers have been made previously (e.g. Hinch and Somers, 1987), although the tools and methods available to detect outliers have been enhanced greatly. Following checks associated with normality, bivariate linearity and other traditional aspects, we advocate that researchers examine their data sets using robust methods, in particular the minimum volume ellipsoid approach. Points identified as being atypical should be carefully evaluated to determine their suitability for inclusion and whether additional factors explain their unusual characteristics. Following these assessments, the PCA can be carried out using traditional approaches with any designated outliers removed or downweighted. Such points can be projected onto the solution a posteriori if their inclusion is desired. The more robust measure of correlation/covariance may often lead to more insightful interpretations of the data pattern rather than simply contrasting a few atypical observations with the general sample.

ACKNOWLEDGEMENTS

We wish to acknowledge funding support for this project from the Natural Sciences and Engineering Research Council to both authors. We thank two anonymous reviewers for their helpful comments. Software for running the robust principal component analysis is available on request from the second author.

REFERENCES

Barcelo C, Pawlowsky V, Grunsky E. 1996. Some aspects of transformations of compositional-based data and the identification of outliers. *Journal of Mathematical Geology* **28**: 501–518.

Barnett V, Lewis T. 1993. *Outliers in Statistical Data* (3rd edn). John Wiley & Sons Ltd.: Chichester, U.K.

Becker C, Gather U. 1999. The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* **94**: 947–955.

Caroni C. 2000. Outlier detection by robust principal component analysis. *Communications in Statistics—Simulation and Computation* **29**: 139–151.

Chen Y, Jackson DA. 1995. Robust estimation of mean and variance in fisheries. *Transactions of the American Fisheries Society* **123**: 401–412.

Chen Y, Jackson DA, Paloheimo JE. 1994. Robust regression approach to analyzing fisheries data. *Canadian Journal of Fisheries Aquatic and Sciences* **51**: 1420–1429.

Egan W, Morgan SL. 1998. Outlier detection in multivariate analytical chemical data. *Analytical Chemistry* **70**: 2372–2379.

Filzmoser P. 1999. Robust principal component and factor analysis in the geostatistical treatment of environmental data. *Environmetrics* **10**: 363–375.

Hinch SG, Somers KM. 1987. An experimental evaluation of the effect of data centering, data standardization, and outlying observations on principal component analysis. *Coenoses* **2**: 19–23.

Jackson DA. 1988. Fish communities of the Black and Hollow River basins. *M.Sc. Thesis*, University of Toronto, Toronto, Ontario.

Jackson DA, Somers KM. 1989. Are probability estimates from the permutation model of Mantel's test stable? *Canadian Journal of Zoology* **67**: 766–769.

Kosinski AS. 1998. A procedure for the detection of multivariate outliers. *Computational Statistics and Data Analysis* **29**: 145–161.

Marden JI. 1999. Same robust estimates of principal components. *Statistics and Probability Letters* **43**: 349–359.

Mosteller F, Tukey JW. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley: Reading, MA, U.S.A.

Pell RJ. 2000. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems* **52**: 87–104.

Rousseeuw PJ. 1985a. Least median of squares regression. *Journal of the American Statistical Association* **79**: 871–880.

Rousseeuw PJ. 1985b. Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, Vol. B, Grossman W, Pflug G, Vincze I, Wertz W (eds). Reidel Publishing; 283–297.

Rousseeuw PJ, Leroy AM. 1987. *Robust Regression and Outlier Detection*. Wiley: New York.

Seaver BL, Triantis KP. 1995. The impact of outliers and leverage points for technical efficiency measurement using high breakdown procedures. *Management Science* **41**: 937–956.

