Applied Usage of the Minimum-Volume Ellipsoid

Kim van der Linde & David Houle, Department of Biological Science, Florida State University,
Tallahassee, FL 32306-1100, USA; telephone (850) 644-5091 (work), fax (850) 644-9829,
e-mail: kim@kimvdlinde.com.

Abstract:

The minimum volume ellipsoid (MVE) method is a powerful algorithm for detecting multivariate outliers. We report here extensions to the method that facilitate its use when variance-covariances matrices may be singular and when outliers can be checked to determine whether they are caused by measurement error or a truly anomalous observation. Before applying MVE, we perform a principal-components analysis and retain only those eigenvectors with positive eigenvalues. To facilitate the investigation of outliers, we rank them from the highest distance score to the lowest. In our application, the highest scores are almost inevitably erroneous measurements that should be corrected, whereas the lowest scores arise from slight departures from multivariate normality and are not removed. Elements of this approach are applicable to many other sets of multivariate data.

Key words: Minimum volume ellipsoid; Multivariate outlier detection; Principal-components analyses.

Introduction

Outlier detection is a crucial step in data analysis, because of the disproportion influence the outliers can have on the statistics; a single outlier can either obscure a real effect or introduce a nonexistent effect. Outlier detection in univariate samples is a common practice and can be carried out straightforwardly by visual inspection of the data or by statistical tests using order statistics. Outlier detection is less straightforward in two-dimensional spaces, because visual inspection is less effective and the order statistics are lacking. In higher-dimensional spaces, the problem is even more difficult. Traditional multivariate outlier-detection methods are based on the calculation of the generalized squared (Mahalanobis) distances for each data point. Mahalanobis distances are in essence weighted Euclidean distances; the distance of each point from the center of the distribution is weighted by the inverse of the sample variance-covariance matrix. Unfortunately, outliers greatly inflate the covariance matrix and can therefore effectively mask their own existence.

To counter this masking problem, Rousseeuw (1985) introduced the robust minimum volume ellipsoid (MVE) method for detection of outliers in multidimensional data. Subsets of approximately 50% of the observations are examined to find the subset that minimizes the volume occupied by the data. The best subset (smallest volume) is then used to calculate the covariance matrix and the Mahalanobis distances to all the data points. An appropriate cut-off value is then estimated, and the observations with distances that exceed that cut-off are declared to be outliers. To minimize computation time, Rousseeuw and Leroy (1987) proposed a resampling algorithm in which subsamples of p+1 observations (p is the number of variables), the minimum to determine an ellipsoid in p-dimensional space, are initially drawn. The algorithm is described in various publications (Rousseeuw & Leroy 1987; Rousseeuw & van Zomeren 1990a; Jackson & Chen 2004).

A serious problem is that both the traditional multivariate and the MVE approach require inversion of the covariance matrix. Therefore, neither method can be applied to samples with singular covariance matrices. We encountered this problem frequently; with our data, the Fortran program MINVOL (Rousseeuw 1990) as well as the SAS implementation of MVE (SAS Institute Inc. 1999) returned prematurely as a result.

In many applications, the original samples or some representation of them is still available. In such cases, inspection of outlier samples may be useful. For example, an anomalous individual may be useful or interesting in its own right. Perhaps more likely is that the measurements taken from the sample are simply in error and can be corrected. On the other hand, real samples are very unlikely to have precisely the multivariate normal distribution that these methods assume. Merely discarding each observation flagged as an outlier may force -the distribution into a near-normal state that is not justified. Reexamination of outlier material allows determination of which observations are caused by measurement errors and retention of those that are not.

Here, we describe our Java implementation of the MVE method, which deals with several of these issues. We provide a straightforward method for dealing with singular covariance matrices and facilitate reexamination of the original samples by sorting the outliers by their Mahalanobis distances.

Procedure Description

Our data sets were generated in a large project on the evolution of shape and vein patterns in *Drosophila* wings (Houle *et al.* 2003). We use a semi-automated system for detecting vein intersections. A digital image of a *Drosophila* wing is obtained; two starting points are supplied by the user, and a program then fits B-splines to the vein pattern. Twelve intersections of these veins are routinely used for our data analysis (Figure 1, top image). The result is 12 × 2 dimensions (x and y coordinates) = 24 measurements for each wing. We analyze these data in the framework of geometric morphometrics (Dryden & Mardia 1998), where the spatial positions of landmarks are of primary interest, rather than derived measures such as lengths or areas. The first step in a geometric morphometric analysis is to align the sets of coordinates. As a result, four degrees of freedom are lost, two from translating the set of coordinates to a common centroid location, and one each from rescaling the coordinates to a common size and rotating the array to maximize the fit. We use a modified generalized Procrustes least squares algorithm for alignment (Rohlf & Slice 1990; Rohlf 2002b).

Although our automated system is generally fairly accurate, the splining program sometimes fails to produce a good estimate of wing form. In some cases, the problem is operator error, but in others, subtle properties of the image mislead the fitting algorithm. Consequently, all the images must be checked for accuracy. We initially examined the fit to each image, but we have obtained images of over 300,000 wings in the past few years, so the time burden is considerable. We have therefore replaced exhaustive checking with automated outlier detection.

Outliers in the data are detected with the MVE module (van der Linde 2004b). The loss of degrees of freedom due to the alignment of landmarks means that the variance-covariance matrices of even the largest data sets will be singular. We therefore perform a principal-components analysis before outlier detection and score the observations on the eigenvectors with positive eigenvalues. These variates are then subjected to the MVE algorithm. The Mahalanobis distance of each detected outliers is retained, and these are sorted from largest to smallest. The outlier observations are then inspected by a human observer (Figure 1, lower three images) and corrected if necessary with the digitizing program tpsDig (Rohlf 2002a). Checking proceeds from the largest outliers to the smallest. Commonly a large group of observations are correctly splined but fall slightly farther from the distribution than expected under normality. Once the observer checking finds that the overwhelming majority of remaining images are in this category, checking is suspended, so additional time is saved in large data sets. Observations that remain anomalous after errors are removed can either be removed from the data set or retained at the discretion of the user.

For our geometric morphemetric data, the whole outlier-detection procedure must usually be repeated two or three times before no new correctable outliers are detected, because

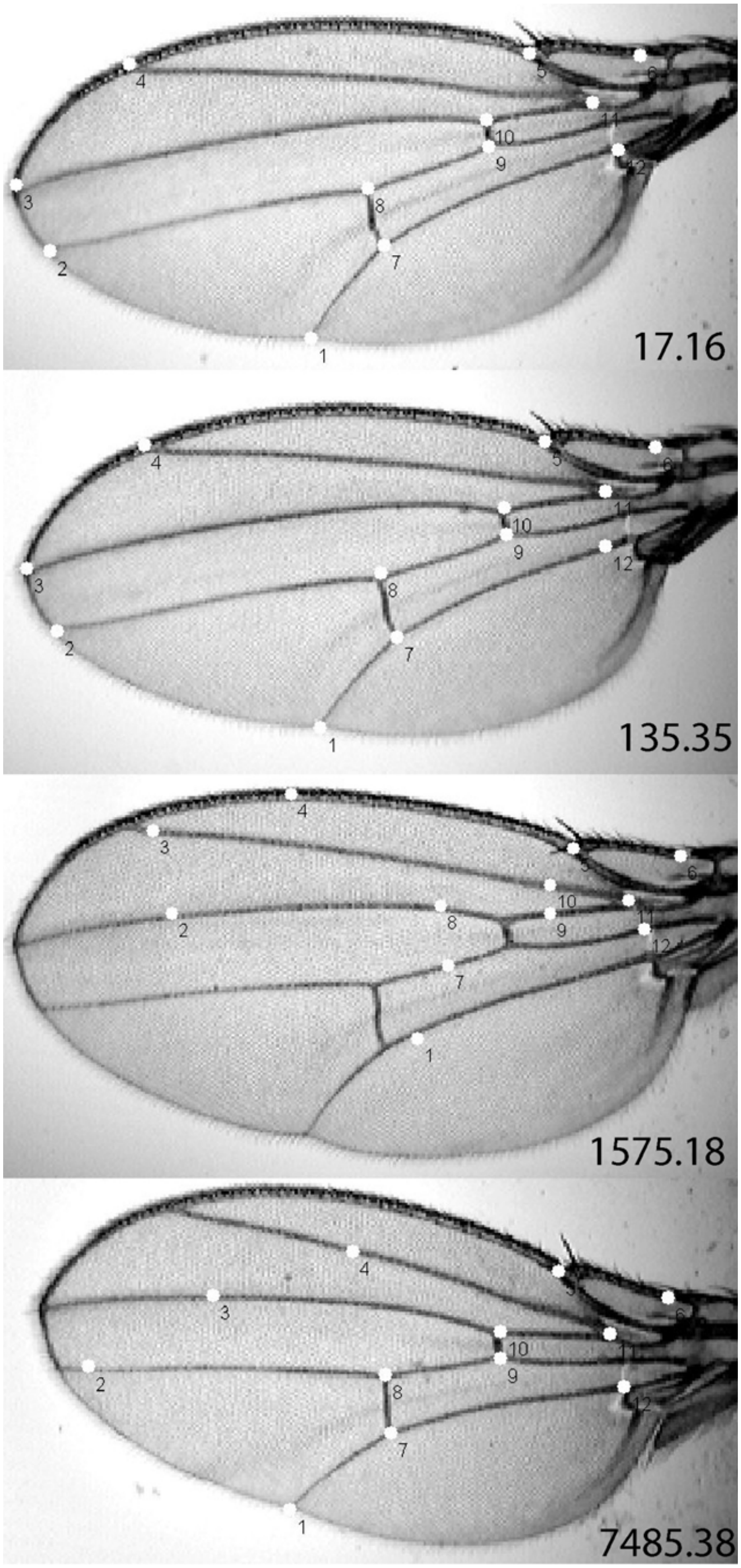

Figure 1: Wings with superimposed landmark points (white dots). The number at the bottom right in each image is the Mahalanobis distance for that wing. All four images were considered outliers by the PCA-MVE module, but the top wing shows the landmarks at their correct locations. The Mahalanobis distance is just slightly larger than the cut-off threshold. We interpret this individual as a "biological outlier." The lower three wings all contain misplaced landmarks. The second wing only has one misplaced landmark (12). Almost every landmark is misplaced on the third wing, yet the Mahalanobis distance is less than that of the fourth wing because the relative locations of the landmarks are roughly correct. All wings are from *Drosophila simulans*.

the alignment of wings by a least-squares criterion can itself mask outliers. A robust alignment algorithm that might solve this problem does exist (Rohlf and Slice 1990), but currently no software is available to implement it for large data sets (D. Slice, pers. comm.). In general, the robust nature of the MVE algorithm should obviate the need for iterative outlier detection for most data sets.

Our approach is implemented in several Java classes based on Sun's Java™ 2 SDK, Standard Edition, version 1.4.2 (Sun Microsystems Inc. 1992--2004). The classes are available as executable jar files from the website of the first author (van der Linde 2004b), as is a stand-alone version of the MVE procedure (van der Linde 2004a). This combined PCA-MVE method requires approximately 15 seconds on an Intel® 2.8 GHz processor with 448 MB of RAM for a typical dataset of about 100 individuals, but processing time increases linearly with increasing number of individuals.

Discussion

Our procedure for detecting outliers in multivariate data sets by the MVE method effectively eliminates the problems associated with singular variance-covariance matrices. In addition, the return of a ranked list of outliers facilitates error checking and helps to emphasize that outlier detection is a somewhat arbitrary process. When the underlying data are still available, as in our case, the data can themselves be checked and an informed decision made about the disposition of each flagged observation. The validity of the outliers cannot always be checked against the raw data, but our implementation always indicates which cases should be viewed with the most caution.

In MVE the best ellipsoid could be missed because of the random sampling of the data set, so some outliers might be missed (Cook & Hawkins 1990) or some valid points labeled as outliers. In practice, the number of subsamples examined ensures that the best subset is always close to the actual best MVE, so the discrepancies will be small and only individuals very close to the optimal cutoff value will be missed. The implementation of MVE in SAS (SAS Institute Inc. 1999) masks this random effect by seeding the pseudo-random number generator with the same seed every time. The random aspect in the method will always remain a point of concern, and outliers close to the edge of the ellipsoid should be treated cautiously.

A more serious issue is the assumption of multivariate normality, which may often be false. Slight departures from normality may either increase or decrease the proportion of observations declared outliers. Reinterpretations of familiar data sets, such as the infamous stack-loss data (Cook & Hawkins 1990; Rousseeuw & van Zomeren 1990a, b) or the example used by Croux and Haesbroeck (2002) turn on this issue.

Acknowledgements

This work was supported by U.S. National Science Foundation grant DEB-0129219.

References

Cook, D.R. & Hawkins, D.M. 1990. Unmasking multivariate outliers and leverage points—comment. *J. Am. Stat. Assoc.* 85: 640-644.

Croux, C. & Haesbroeck, G. 2002. A note on finite-sample efficiencies of estimators for the minimum volume ellipsoid. *J. Stat. Comput. Simul.* 72: 585-596.

Dryden, I.L. & Mardia, K.V. 1998. Statistical Shape Analysis. John Wiley and Sons, Chichester, U.K.

Houle, D., Mezey, J., Galpern, P. & Carter, A. 2003. Automated measurement of *Drosophila* wings. *BMC Evol. Biol.* 3:25.

Jackson, D.A. & Chen, Y. 2004. Robust principal component analysis and outlier detection with ecological data. *Environmetrics* 15: 129-139.

Rohlf, F.J. 2002a. tpsDig. http://life.bio.sunysb.edu/morph.

Rohlf, F.J. 2002b. tpsRegr. http://life.bio.sunysb.edu/morph.

Rohlf, F.J. & Slice, D. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Syst. Zool.* 39: 40-59.

Rousseeuw, P.J. 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I.&Wertz, W. (Eds.), Mathematical Statistics and Applications, vol.B. Riedel Publishing, Dordrecht, the Netherlands, pp. 283-297.

Rousseeuw, P.J. 1990. MINVOL http://www.agoras.ua.ac.be/Robustn.htm.

Rousseeuw, P.J. & Leroy, A.M. 1987. Robust Regression and Outlier Detection. Wiley-Interscience, New York, U.S.A.

Rousseeuw, P.J. & van Zomeren, B.C. 1990a. Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* 85: 633-639.

Rousseeuw, P.J. & van Zomeren, B.C. 1990b. Unmasking multivariate outliers and leverage points—rejoinder. *J. Am. Stat. Assoc.* 85: 648-651.

SAS Institute Inc. 1999. The SAS System for Windows.

Sun Microsystems Inc. 1992--2004. Java(tm) Development Kit. http://java.sun.com.

van der Linde, K. 2004a. MVE: Minimum Volume Ellipsoid Estimation for Robust Outlier Detection in Multivariate Space. http://www.kimvdlinde.com/professional/mve.html.

van der Linde, K. 2004b. PCA-MVE: Robust Minimum Volume Ellipsoid Estimation for Robust Outlier Detection in Multivariate Space. http://www.kimvdlinde.com/professional/pcamve.html.