Stat Methods Appl (2015) 24:257–261
DOI 10.1007/s10260-015-0324-9

DISCUSSION

Hubert, Rousseeuw and Segaert: multivariate functional outlier detection

Aldo Corbellini[1] · Marco Riani[1] ·
Anthony C. Atkinson[2]

Accepted: 8 May 2015 / Published online: 27 May 2015

As would be expected from these authors, this is an interesting and well-written paper. It provides a stimulating introduction to robustness problems in the analysis of functional data. We have four general comments, none extensive.

1 Robust bivariate boxplots

The "bag-plot" of Rousseeuw et al. (1990), which you exemplify in your Figure 8, provides a polygonal approximation to the unknown distribution. The robust bivariate boxplot introduced by Zani et al. (1998) provides a smoother approximation to the unknown distribution. We briefly recall some of its properties.

Zani et al. (1998) use the peeling of convex hulls to, in the tradition of very robust statistics, find a region that contains approximately 50 % of the data. Peeling of hulls continues until the first one is obtained which includes not more than 50 % of the data (and asymptotically half the data as the sample size increases). The "50 % hull" so found is smoothed using B-splines, with cubic pieces which use the vertices of the 50 % hull to provide information about the location of the knots.

The robust centre of the data is found as the bivariate mean of the observations lying with the 50 % contour. Given the centre and the contour, the distance of any point from

✉ Marco Riani
mriani@unipr.it

Aldo Corbellini
aldo.corbellini@unipr.it

Anthony C. Atkinson
a.c.atkinson@lse.ac.uk

[1] Dipartimento di Economia, Università di Parma, Parma, Italy

[2] Department of Statistics, London School of Economics, London WC2A 2AE, UK

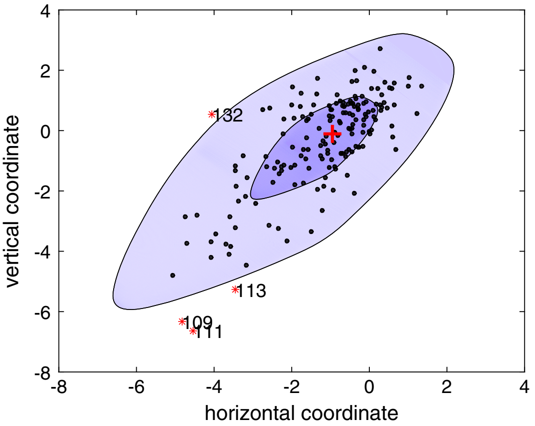

Fig. 1 Robust bivariate boxplot of the writing data at time $t = 5$. The greater spread of the data in the lower left corner is evident

the centroid can be defined. Apart from the method of constructing the reference contour, the method is the same as that in your section 3.2. Riani and Zani (1998) comment that the distance they find is non-parametric, robust, takes into account, as you stress, the differing spread of the data in different directions and reduces to the customary squared Mahalanobis distance when the uncontaminated data are elliptically distributed.

The assumption for other contours is that they will have the same shape as the 50 % contour. For diagnostic purposes, Riani and Zani (1998) find expansion coefficients leading to contours which contain a prespecified portion of the data when they are normally distributed.

Figure 1 shows the resulting robust bivariate boxplot for the data at time $t = 5$, with a 99 % contour, to be compared with your Figure 8. Both plots reveal the outlying nature of observation 132. However, our plot suggests that the three observations for low values of both variables are also outlying. In both bivariate boxplots the robust centroids are not at the centre of the 50 % contour but are located closer to the upper right inner contour. Indeed, both boxplots indicate that the spread of the data in the lower left part of the plot is greater than that in the upper right corner. This is an indication that the data may need to be transformed. In our next section we consider transformations of the data to normality.

2 Transformations of data

Atkinson and Riani (1997) illustrate the use of scatterplot matrices incorporating the robust bivariate boxplot in the analysis of a five-dimensional example on soil properties. Some of the boxplots are plausibly elliptical, others are not. They extend their analysis to a robust form of the Box-Cox transformation to normality, which we now further extend.

Developments of our work on the robust analysis of multivariate data are described in Atkinson et al. (2004). We have found, there and elsewhere, that use of a robust form of the Box-Cox transformation leads to data for which Mahalanobis distances

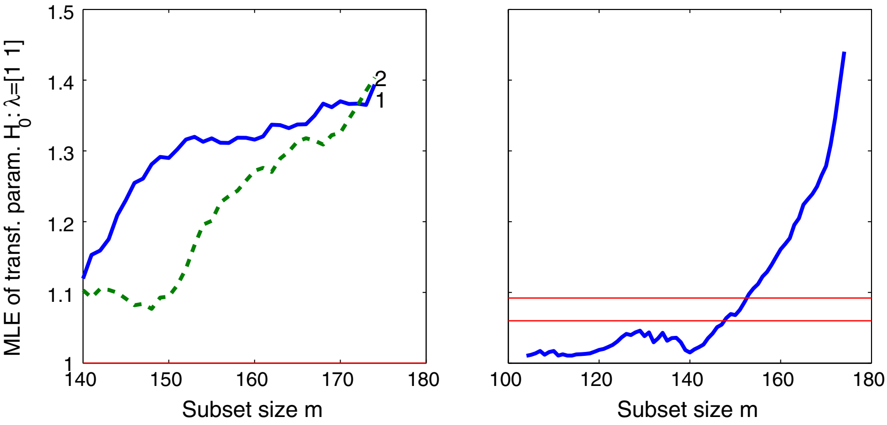

Fig. 2 *Left-hand panel*: monitoring the maximum likelihood estimates of the two transformation parameters during the forward search. *Right-hand panel*: monitoring the likelihood ratio test of no transformation. The two horizontal bands respectively correspond to 95 and 99 % asymptotic pointwise confidence limits

are useful indicators of outlyingness and incorrectly specified structure. Mahalanobis distances are, of course, easy to calculate, robustly or not, whatever the dimension. A discussion of the use of this transformation in principal components analysis is in Maadooliat et al. (2015). Given that in this case the data contain both positive and negative values, we instead use the transformation of Yeo and Johnson (2000). This is a Box-Cox transformation of $y + 1$ for nonnegative values of y, and of $|y| + 1$ with parameter $2 - \lambda$ for y negative.

The left-hand panel of Fig. 2 displays the monitoring of the maximum likelihood estimates of the two transformation parameters during the latter part of the forward search. This exhibits an upward trend departing from one, as the more remote observations are included in the subset. The right-hand panel of Fig. 2, which shows the monitoring of the likelihood ratio test for the hypothesis of no transformation, indicates that the need to transform the data does not depend on the presence of few observations. The value of the test statistic begins to exceed the 99 % confidence band from a subset size $m = 155$.

Figure 3 shows the bivariate boxplot of the transformed data using the transformation $\lambda_1 = \lambda_2 = 1.3$. In the untransformed scale of Fig. 2 there was a high concentration of data points near the upper-right corner of the 50 % hull. This feature is no longer evident in Fig. 3. The contours seem more nearly elliptical and the points are appreciably closer to having an elliptically symmetrical distribution.

Our form of boxplot produces smooth contours; further, no data point is forced to lie on the fence. However, as Rousseeuw et al. (1990) comment, there may in general be some loss of robustness in small samples due to the use of peeling. However, in this particular example, the 50 % hull includes exactly $n/2 - 1$ (86) data.

Like the bagplot, our robust bivariate boxplot becomes computationally much more challenging in three or more dimensions. Scatterplot matrices are then very helpful. Of course, the analysis which has been presented here is for one time point. For the transformation model to be useful in the analysis of these functional data, one would

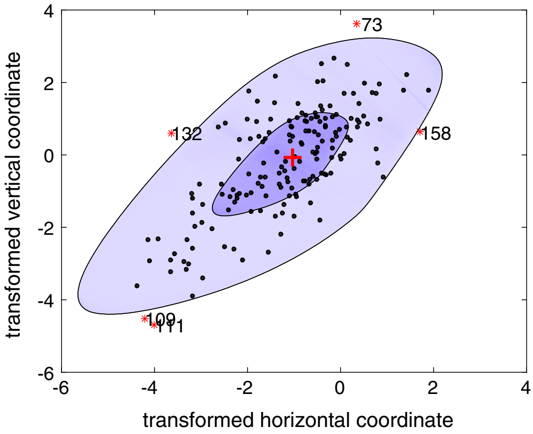

Fig. 3 Robust bivariate boxplot of the writing data at time $t = 5$ transformed with $\lambda_1 = \lambda_2 = 1.3$. The contours and distribution of data points are more elliptical than in Fig. 1

hope that the same transformation would hold at all times. Failing that, the transformation parameters might be expected to vary in a smooth, and perhaps informative, way. Outlying periods will be revealed by a jump to a different set of parameter values.

3 Time series and Kriging

We wonder about the relationship between functional data analysis and the analysis of time series. In your abstract you refer to “each time point” and it is tempting to wonder whether such Kalman-filter based methods as those of West and Harrison (1989) and Harvey (1989), suitably robustified, might be appropriate. In your work the distances are calculated independently at each “time” point and then summed (your equation 3) over time. Self evidently, in functional data, there is no sense of moving from “left” to “right”. But one might expect some stochastic relationship between adjacent observations, such as is incorporated in Kriging models.

4 Graphics

All the new plots which are presented are highly informative and are strongly interrelated. It will be very important to have brushing and linking tools which enable the user to connect the information on a particular subgroup of units in different plots. Such procedures are possible in Matlab (e.g. Riani et al. 2014), but not routinely in R. We think it will be very helpful to have your new routines produced for both languages.

We enjoyed reading your paper and look forward to your responses in the light of your experience of the analysis of functional data.

References

Atkinson AC, Riani M (1997) Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis: the 1997 Hunter lecture. Environmetrics 8:583–602

Atkinson AC, Riani M, Cerioli A (2004) Exploring multivariate data with the forward search. Springer, New York

Harvey AC (1989) Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge

Maadooliat M, Huang JZ, Hu J (2015) Integrating data transformation in principal components analysis. J Comput Gr Stat 24:84–103

Riani M, Zani S (1998) Generalized distance measures for asymmetric multivariate distributions. In: Rizzi A, Vichi M, Bock H-H (eds) Advances in data science and classification. Springer, Berlin, pp 503–508

Riani M, Cerioli A, Atkinson AC, Perrotta D (2014) Monitoring robust regression. Electr J Stat 8:642–673

Rousseeuw PJ, Ruts I, Tukey JW (1990) The bagplot: a bivariate boxplot. Am Stat 53:87–88

West M, Harrison PJ (1989) Bayesian forecasting and dynamic models. Springer, New York

Yeo I-K, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87:954–959

Zani S, Riani M, Corbellini A (1998) Robust bivariate boxplots and multiple outlier detection. Comput Stat Data Anal 28:257–270