Stat Methods Appl (2015) 24:245–251
DOI 10.1007/s10260-015-0321-z

DISCUSSION

Discussion of "Multivariate functional outlier detection" by Mia Hubert, Peter Rousseeuw and Pieter Segaert

Yuan Yan[1] · **Marc G. Genton**[1]

Accepted: 1 May 2015 / Published online: 16 May 2015

We would like to congratulate M. Hubert, P. Rousseeuw and P. Segaert (henceforth, HRS) on their publication of this innovative and inspiring paper on the topic of outlier detection for multivariate functional data. These authors have extended a robust statistical methodology for functional data from the univariate to the multivariate setting, an essential step due to the emergence of ever abundant and complex data sets.

In their paper, HRS presented a thorough literature review of existing notions of data depth and outlier detection techniques for multivariate, univariate functional and multivariate functional data, as well as of diverse topics related to the notion of halfspace depth. The authors proposed a distance-based measure of outlyingness ("bagdistance") and the use of heatmaps or centrality-stability plots as visualization tools to detect outliers in functional data. The ability to detect different kinds of multivariate functional outliers by applying these new techniques was demonstrated on several real data examples.

To begin our discussion of HRS's paper, we would like to point out that HRS only confirmed the expected behavior of heatmaps or centrality-stability plots based on what was previously known about the data, but they did not present a specific decision-making procedure. We therefore have only an approximate idea of the potential outlying curves by using these visualization methods. However, there are other clear-cut outlier identification methods for functional data and Hyndman and Shang (2010) compared the performance of several of them. Here, we discuss in detail the functional boxplot approach (Sun and Genton 2011) for detecting univariate

✉ Marc G. Genton
marc.genton@kaust.edu.sa

Yuan Yan
yuan.yan@kaust.edu.sa

[1] CEMSE Division, King Abdullah University of Science and Technology,
Thuwal 23955-6900, Saudi Arabia

functional data outliers and its extension to multivariate functional data (López-Pintado et al. 2014). We then describe a generalization of the methods proposed by HRS.

1 Outlier detection based on adjusted functional boxplots

In Section 5.2 of HRS's paper, the performance of the proposed outlier detection techniques was compared with the functional boxplot approach by applying them to the tablets data. HRS observed that many outlying curves were not detected by applying the functional boxplot to the three marginal curves. However, they used the default inflation factor 1.5 for fences in the functional boxplot, ignoring the fact that the factor needs to be modified due to possible "spatio-temporal" correlations. To address this issue, Sun and Genton (2012) proposed the adjusted functional boxplot, where the factor is selected based on simulation to account for the effect of spatio-temporal correlations.

We now explore the use of the adjusted functional boxplot as a univariate functional outlier detection tool on the tablets data. Since the tablets data are univariate, it is natural to apply the adjusted functional boxplot to the original tablets data set and the derivative function of this data set. The derivative function takes shape information into consideration, and by including it, we expect to be able to detect outliers differing either in magnitude or shape. We choose the inflation factor by using a simulation based method similar to that in Sun and Genton (2012), but with modification of the criteria: we select the smallest factor that makes the percentage of falsely detected outliers, p_f (the number of falsely detected outliers divided by the total number of non-outlying curves), lower than 0.7 %. In our simulation, we generate curves without contamination. Thus, p_f is simply the number of falsely detected outliers divided by the total number of curves generated.

For the tablets data, we can regard the correlations among wavenumbers as "temporal". We treat the 90 samples as independent curves, and explore how the factor will be affected by the temporal correlations. We first detrend the data from the median of each curve. Since the sample size 90 is too small compared to the dimension 404, we are not able to estimate the complete 404 × 404 temporal covariance matrix from the data. Instead, we select 46 wavenumbers out of 404 with equal spacing and including the last one, i.e., 1, 10, 19, …, 397, 404. Then we use data at the chosen 46 time points (wavenumbers) to estimate the temporal covariance matrix, with the 90 curves as replicates. A robust estimation of the 46 × 46 temporal covariance matrix is performed, where we use the orthogonalized quadrant correlation pairwise estimator with the function "covRob" in the R-package "robust". We generate 90 independent curves at 46 time points from a zero-mean Gaussian stochastic process with the estimated covariance matrix structure. Then, we apply the functional boxplot based on the modified band depth (López-Pintado and Romo 2009) to the simulated data. We repeat this 1000 times for each value of the inflation factor. Finally, we check p_f for different values of the inflation factor to select the smallest one that satisfies our aforementioned criteria. The selected factor is quite stable when the number of time points chosen ranges from 21 to 46. In the description above, we use 46 time points, which is the largest dimension for which "covRob" is effective.

We apply the above factor selection procedure to both the original tablets data set and the corresponding derivative function. The selected factors are 1.1 for both cases. For the original data set, observations 1–7, 9, 10, 71–83, 85, 87, 88, 90 were detected as outliers by the adjusted functional boxplot with the factor 1.1. For the derivative function, observations 1, 2, 4–6, 71–90 were detected, because their shapes differ from the shapes of the other observations. The combined results detect all the outliers except for curve 8, which matches the left most red dot in the centrality-stability plot (Fig. 28 in HRS) and is not an obvious outlier even in that plot.

These results demonstrate that applying the adjusted functional boxplot to univariate functional data and the corresponding derivative function yields the desired results in the tablets data set and gives explicit identification of outliers, something that is lacking in the methods proposed by HRS.

2 Outlier detection based on bivariate modified simplicial band depth

We can also use the modified simplicial band depth (MSBD) of López-Pintado et al. (2014) to compute bivariate functional depth as applied to the tablet curves and their derivatives. We then use the functional boxplot based on this bivariate functional depth to determine the three-dimensional central region and thus to detect outliers. Because there is no significant correlation between these two variables for the tablets data set, applying a functional boxplot to the two marginal functions is sufficient.

To select the inflation factor, we generate two uncorrelated data sets from a zero-mean Gaussian stochastic process using the two 46×46 matrices estimated before for the tablets data and for the derivatives, respectively. Combining the two data sets to a bivariate data set, we can compute the modified simplicial band depth for each curve. Then, we apply the functional boxplot to the two simulated data sets separately based on the same bivariate modified simplicial band depth. In this case, the number of falsely detected outliers is the number of distinct outliers detected from the two simulated data sets. We again repeat this 1000 times for each value of the inflation factor. Finally, we check p_f for different values of the inflation factor to select the smallest one that satisfies our aforementioned criteria. By using the factor selection procedure described before, we find that the proper factor is 1.0. Based on this inflation factor value, all 30 outliers were detected by the adjusted functional boxplot applied to the two marginal functions.

Thus, multivariate functional outlier detection based on the adjusted functional boxplot in conjunction with the modified simplicial band depth performs competitively.

3 Generalization of the methods in HRS

The outlier detection methods demonstrated in Sects. 3 and 4 of HRS's paper share some similarities and can thus be generalized.

The bagdistance in Sect. 3 and the skew-adjusted outlyingness (AO) in Sect. 4 are both distance-based rankings of outlyingness of multivariate data. The integration of the bagdistance or AO over time can be regarded as the multivariate functional outlyingness. Multivariate functional depth can also be defined as the integral of the local multivariate depth incorporating a weight (Claeskens et al. 2014). There is a one-to-one correspondence between depth (D) and outlyingness (O) in the multivariate setting given that $\mathrm{O}(\mathbf{x}; P_Y) = \frac{1}{\mathrm{D}(\mathbf{x}; P_Y)} - 1$ or $\mathrm{D}(\mathbf{x}; P_Y) = \frac{1}{1+\mathrm{O}(\mathbf{x}; P_Y)}$ (Mosler 2013). Multivariate functional depth and outlyingness can thus be constructed based on any existing notions of depth or outlyingness for multivariate data. Depth and outlyingness heatmaps can be very useful visualization tools to compare the behaviors of different definitions of multivariate functional depth or outlyingness based on various building blocks in the multivariate setting. HRS showed through heatmaps that the distance-based measure of outlyingness and its corresponding depth are able to detect outliers that are not persistent.

Because it is not necessary for shape or isolated outliers to have low functional depth or high functional outlyingness, it can be difficult to detect them based solely on the functional depth or outlyingness. However, these outliers tend to show greater variability in the degree of outlyingness over time. The centrality-stability plot proposed by HRS is in essence based on the relationship between the arithmetic and harmonic means, revealing and quantifying the variability of the local outlyingness at each time point as well as the overall functional depth. We can thus use the centrality-stability plot to visualize different kinds of outliers.

The centrality-stability plot can be generalized by using any other notion of functional depth (FD), which is the integral of the local depth (D) incorporating a weight (W). We therefore define the generalized centrality-stability plot with the scatterplot of

$$\left(1 - \mathrm{FD}_n(Y_i; P_n); \mathrm{ave}_j \left[\frac{1}{\mathrm{D}\{Y_i(t_j); P_n(t_j)\}W_j}\right] - \frac{T}{\mathrm{FD}_n(Y_i; P_n)}\right) \tag{1}$$

for all $i = 1, \ldots, n$.

In the generalized centrality-stability plot, the horizontal axis measures a curve's overall deviation from centrality by using $1 - \mathrm{FD}$. The vertical axis measures the stability of the local outlyingness by taking the difference between the arithmetic mean and harmonic mean of the reciprocal of weighted local depths. There can be alternative ways for measuring the centrality or stability. For example, we can use the functional outlyingess (FO) for the horizontal axis and the vertical distance can be the difference between the arithmetic mean and harmonic mean of the weighted local depths. We therefore define an alternative centrality-stability plot with the scatterplot of

$$\left(\mathrm{FO}_n(Y_i; P_n); \mathrm{ave}_j \left[\mathrm{D}\{Y_i(t_j); P_n(t_j)\}W_j\right] - \frac{T}{\sum_{j=1}^{T} \left[\mathrm{D}\{Y_i(t_j); P_n(t_j)\}W_j\right]^{-1}}\right) \tag{2}$$

for all $i = 1, \ldots, n$, where $\mathrm{FO}_n(Y_i; P_n) = \sum_{j=1}^{T} \mathrm{O}\{Y_i(t_j); P_n(t_j)\}W_j$.

The bivariate MSBD for functional data used in the above section can be expressed as the average of the bivariate simplicial depths (Liu 1990) at each time point. The

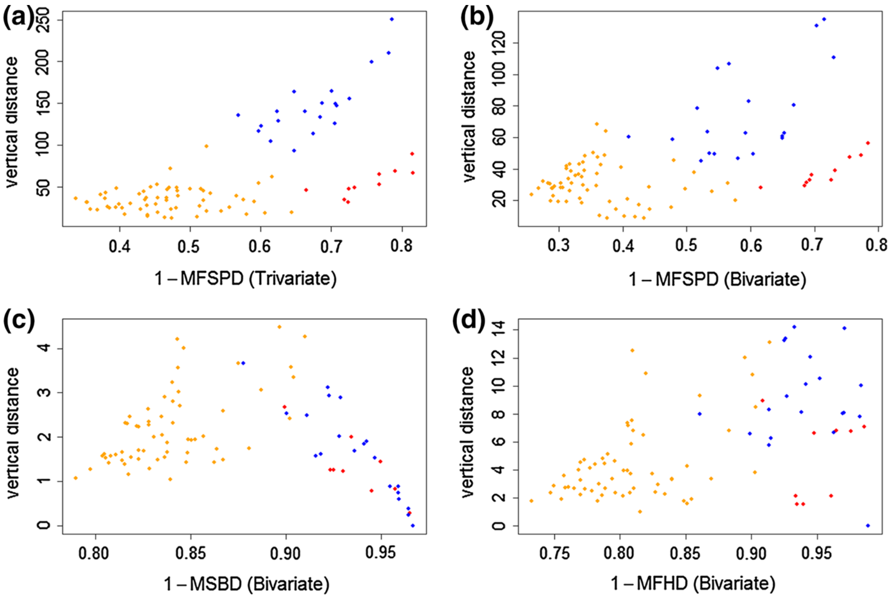

Fig. 1 **a**, **b** Centrality-stability plot for trivariate and bivariate tablets data; **c**, **d** Generalized centrality-stability plot (1) based on MSBD and MFHD for bivariate tablets data

multivariate functional halfspace depth (MFHD) can also be expressed as the average of the bivariate halfspace depths at each time point. We can thus make generalized and alternative centrality-stability plots based on these different bivariate functional depths for the tablets data. In Fig. 1, the original centrality-stability plot based on MFSPD applied to trivariate and bivariate tablets data and two generalized centrality-stability plots based on MSBD and MFHD for the bivariate tablets data are shown for comparison. In Fig. 2, four alternative centrality-stability plots are shown. We use the same color scheme as in Fig. 28 of the HRS paper to distinguish the three different groups of curves.

The generalized centrality-stability plot based on MSBD exhibits different behavior compared to the behavior of the plot made by using MFSPD: there is no significant difference between the three groups in terms of the variability of the local outlyingness. The alternative centrality-stability plot based on MFSPD shows that non-outlying curves can also have great variability of the local depths. And in the alternative centrality-stability plot based on MSBD, we can see that outlying curves tend to have stable local depths over time. These different behaviors can be explained by the fact that the outlyingness induced from simplicial depth is not distance based. On the other hand, we can also see from the plots based on MSBD that almost all outlying curves have the lowest overall functional depths or highest overall functional outlyingness. That is, almost all of the outlying curves are on the right-most side of the plots. Though identifying outliers remains ambiguous in some regions, this ambiguity can be overcome by using adjusted functional boxplots as an outlier detection technique, as we have described above.

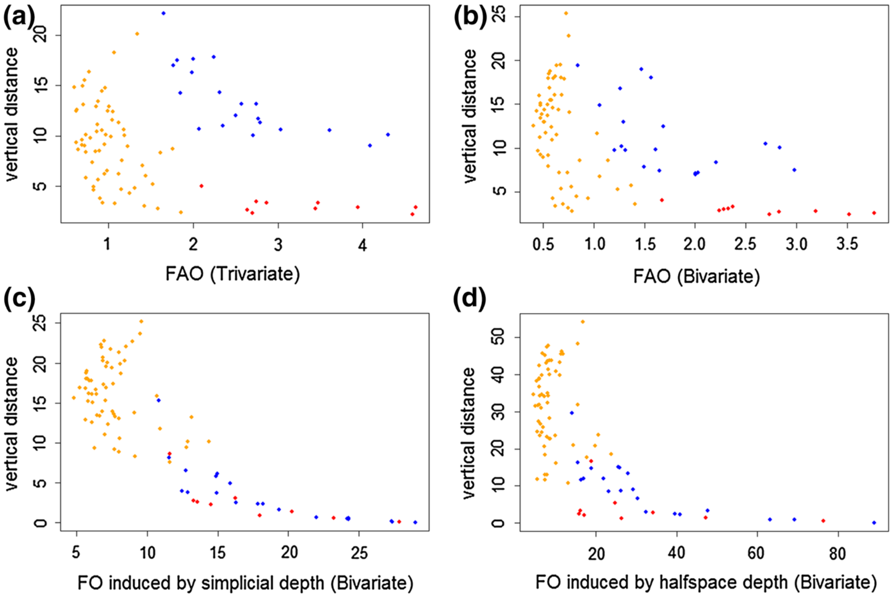

Fig. 2 **a**, **b** Alternative centrality-stability plot (2) for trivariate and bivariate tablets data based on MFSPD; **c**, **d** Alternative centrality-stability plot based on MSBD and MFHD for bivariate tablets data

4 Final remarks

The main difficulty in outlier detection in multivariate functional data is to identify the isolated and shape outliers, which may not result in a low functional depth. To overcome this drawback, distance-based ranking was advocated by HRS. The weight function for the depth of multivariate functional data defined by Claeskens et al. (2014) may also play an important role as demonstrated on the industrial data. Although HRS used uniform weight functions throughout their paper, it might be interesting to compare the performance of different weight functions to that of the proposed distance-based ranking methods. Another direction worth exploring is the case of surfaces/images (Genton et al. 2014) where the index is space rather than time. The detection of outliers in that setting is challenging as well.

References

Claeskens G, Hubert M, Slaets L, Vakili K (2014) Multivariate functional halfspace depth. J Am Stat Assoc 109(505):411–423

Genton MG, Johnson C, Potter K, Stenchikov G, Sun Y (2014) Surface boxplots. Stat 3:1–11

Hyndman R, Shang H (2010) Rainbow plots, bagplots, and boxplots for functional data. J Comput Graph Stat 19(1):29–45

Liu R (1990) On a notion of data depth based on random simplices. Ann Stat 18(1):405–414

López-Pintado S, Romo J (2009) On the concept of depth for functional data. J Am Stat Assoc 104:718–734

López-Pintado S, Sun Y, Lin JK, Genton MG (2014) Simplicial band depth for multivariate functional data. Adv Data Anal Classif 8:321–338

Mosler K (2013) Depth statistics. In: Becker C, Fried R, Kuhnt S (eds) Robustness and complex data structures. Festschrift in honour of Ursula Gather. Springer, Berlin, pp 17–34

Sun Y, Genton MG (2011) Functional boxplots. J Comput Graph Stat 20:316–334

Sun Y, Genton MG (2012) Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics 23:54–64