**Statistical methods and applications manuscript No.**
(will be inserted by the editor)

# Discussion of "Multivariate Functional Outlier Detection", by Mia Hubert, Peter Rousseeuw and Pieter Segaert

**Davy Paindaveine · Germain Van Bever**

Received: date / Accepted: date

**Abstract** We present a discussion of [2] — hereafter, HRS15 — that splits into two parts. In the first one, we argue that some structural properties of depth may, in some cases, limit its relevance for outlier detection. We also propose an alternative to bagdistances, which, while still based on depth, does not suffer from the same limitations. In the second part of the discussion, we investigate the possible uses of the weight functions that may enter the various integral functionals considered in HRS15.

**Keywords** Functional data analysis · Outlier detection · Statistical depth

## 1 Introduction

Building on a world renowned expertise in robustness and statistical depth, HRS15 proposes a systematic approach to outlier detection for univariate and multivariate functional data. We regard the proposed methodology as a major improvement over existing approaches and we would like to sincerely congratulate the authors for their work. As announced in the abstract, our discussion

Davy Paindaveine's research is supported by an A.R.C. contract from the Communauté Française de Belgique and by the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy). Germain Van Bever's research is supported by the British EPSRC grant EP/L010419/1.

D. Paindaveine
Université libre de Bruxelles, Av. F.D. Roosevelt, 50, ECARES - CP114/04, 1050 Brussels, Belgium
Tel.: +32-2-6503845
E-mail: dpaindav@ulb.ac.be

Germain Van Bever
The Open Univeristy, Walton Hall, MK7 6AA, Milton Keynes, Buckinghamshire, Milton Keynes
Tel.: +44-1908-655902
E-mail: germain.van-bever@open.ac.uk

splits into two parts. After commenting on some classical limitations of depth, Section 2 proposes an alternative to bagdistances overcoming some of those limitations, while Section 3 studies how useful weighting mechanisms may be for functional outlier detection.

## 2 Depth and outlier detection

In [5] and [6], a *Depth-Outlyingness-Quantile-Rank (D-O-Q-R)* paradigm is introduced and discussed. Under this paradigm, depth, outlyingness, quantile, and rank functions are seen as strictly equivalent objects; in particular, depth and outlyingness are inversely linked through, e.g., relations of the form $D = 1/(1+O)$ or $D = 1 - O$; see also [7]. This scheme allows to associate with *any* depth function a resulting outlyingness function, on the basis of which one may expect to perform (depth-based) outlier detection.

One of the merits of HRS15 is to clearly articulate that some depth concepts may, however, not be suitable for outlier detection. For instance, in a sample of observations $X_1, \ldots, X_n$ on the real line, many classical depths, including halfspace depth ([8]) and simplicial depth ([3]), will assign a common, small, depth value to the first order statistic $X_{(1)}$ and last order statistic $X_{(n)}$, despite the fact that one of these order statistics may be close to the other $n-2$ observations while the other may be arbitrarily far from them. This clearly shows that, for such depths, it is inadequate to perform outlier detection on the basis of depth only, and that depth should then be complemented with other functionals, typically of a *distance* nature.

The *bagdistance* considered in HRS15 goes in this direction. The bagdistance allows to detect outliers according to their distance to a location that is considered as most central in the data set. Depth plays a key role in the definition of the bagdistance, as it provides (a) the most central location (that is simply the deepest point, or more generally, the barycenter of the collection of deepest points) and (b) the bag, on the basis of which appropriate, directional, standardization can be achieved. The bagdistance is therefore strongly based on depth.

The axiomatic approach of [9] makes it clear, however, that depth is suitable only for distributions that are unimodal with a convex, or at least star-shaped, support. Therefore, for (i) non-convexly supported distributions or for (ii) multimodal distributions, depth may fail properly reflecting centrality[1]. We now show that this may affect outlier detection based on the bagdistance, by considering two bivariate examples that correspond to (i) and (ii), respectively :

(i) In the first example, we consider a sample of mutually independent observations $X_i = \binom{X_{i1}}{X_{i2}}$, $i = 1, \ldots, n = 205$. For the first 200 observations, $X_{i1} \sim \mathrm{Unif}(-1, 1)$ and $X_{i2}|[X_{i1} = u] \sim \mathrm{Unif}(1.5(1-u^2), 2(1-u^2))$ (moon-shaped support). The last five (outlying) observations are drawn from the bivariate

[1] This can be partly addressed by introducing local versions of depth; see [1] and [4].

normal distribution with mean vector $\binom{0}{0.5}$ and covariance matrix $I_2/25$, where $I_2$ is the $2 \times 2$ identity matrix.

(ii) In the second example, a sample of size $n = 205$ made of mutually independent observations is still considered. Here, the first 100 observations (resp., 100 next observations) are drawn from the bivariate normal distribution with mean vector $-\binom{8}{0}$ (resp., mean vector $\binom{8}{0}$) and covariance matrix $I_2$. The last five (outlying) observations are drawn from the bivariate normal distribution with mean vector $-\binom{0}{1.5}$ and covariance matrix $I_2/2$.

Scatter plots of the resulting data points are provided in Figures 1(i-a) and (ii-a), respectively. In each case, the Tukey median is marked as a blue dot and the (halfspace) bag is drawn. To make visualization of the bagdistances easier, the segment from the Tukey median to each data point is plotted in grey and in red, for the 200 regular observations and for the five outlying observations, respectively. Figures 1(i-b) and (ii-b) report histograms of the corresponding bagdistances, with the bagdistances of the five outliers marked with vertical red lines. For convenience, the outliers were renumbered from one to five, according to their bagdistance. In both cases (i)-(ii), bagdistances of outlying observations are not particularly large nor small[2], hence do not reflect outlyingness properly. For the multimodal distribution in (ii), distances to a unique center are not adequate to measure outlyingness, while, for the non-convexly supported distribution in (ii), halfspace depth provides a center that is outside the moon-shaped support. In both cases, the structural properties of depth affect the efficiency of the outlier detection procedure considered.

As example (ii) shows, measuring distances to a unique center may hurt. Instead, one may think of (a) *measuring "distances" to a subset of "closest" observations*. Also, detecting outliers through bagdistances is based on the assumption that outlying observations are far from the bulk of the data (or at least from the center). Outliers, however, can be found at an arbitrary distance from the center, e.g. under the form of a few isolated points. To detect such outliers, (b) *density* ideas should be used. Interestingly, (a)-(b) can be addressed through depth, by replacing bagdistances with the $\beta$-distances we now define.

Fix a sample of observations $X_1, \ldots, X_n$ and a given location $x$ in $\mathbb{R}^d$. First consider the symmetrized sample $X_1, \ldots, X_n, 2x-X_1, \ldots, 2x-X_n$, where $2x-X_i$ is the reflection of $X_i$ about $x$. For any depth satisfying the axioms of [9], the depth regions associated with this symmetrized sample provide nested neighborhoods of $x$. For any $\beta \in (0, 1)$, the $\beta$-distance of $x$ to $X_1, \ldots, X_n$ is then defined as the Lebesgue measure of the smallest such depth region that contains at least a proportion $\beta$ of the original observations $X_1, \ldots, X_n$. If $\beta$ is not too small, the $\beta$-distances of a few isolated outliers are expected to be large, since the corresponding depth region will need to expand to include observations that are far away from the outliers.

To briefly illustrate the use of these alternative distances, we provide histograms of $\beta$-distances for examples (i)-(ii) above, both for $\beta = 0.25$; see

[2] A few, isolated, small bagdistances arguably also may be a sign of "outlyingness".

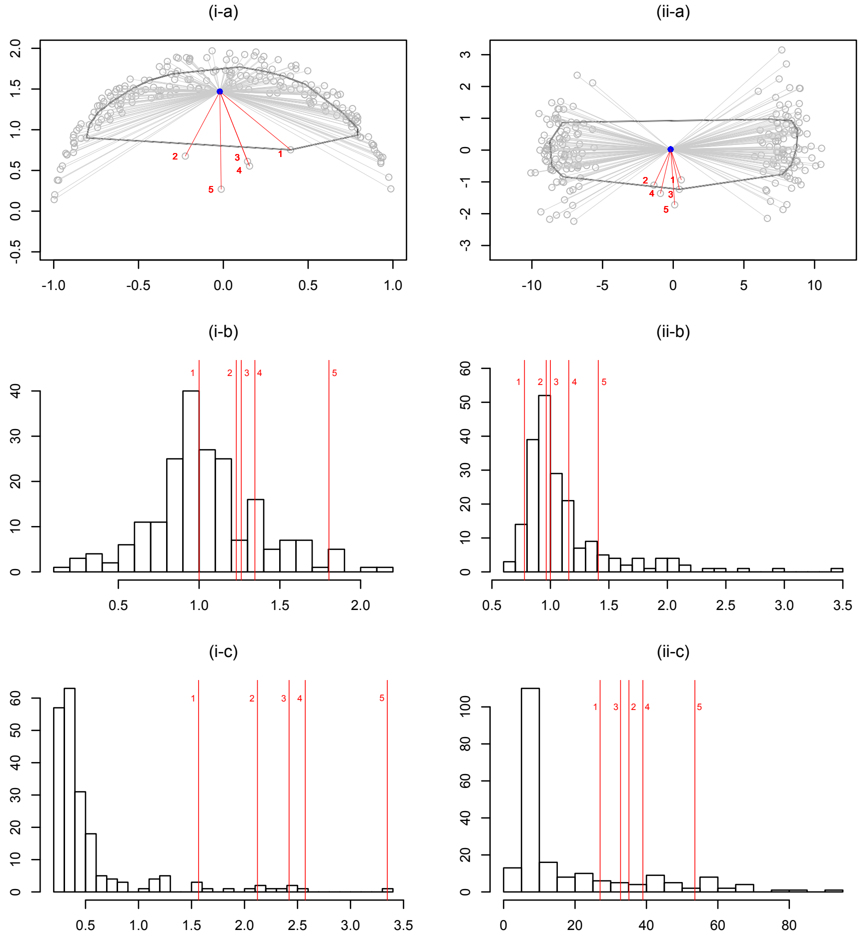


**Fig. 1** Panels (i-a) and (ii-a) report scatter plots of the 205 data points generated in setups (i) and (ii), respectively (see Page 2). In each case, the Tukey median is marked as a blue dot, the bag is plotted, and the five outliers are numbered according to their bagdistance. Panels (i-b) and (ii-b) provide histograms of the resulting 205 bagdistances; bagdistances of the five outliers are marked in red. Panels (i-c) and (ii-c) provide the corresponding histograms of $\beta$-distances, for $\beta = 0.25$.

Figures 2(i-c) and (ii-c), respectively. It is seen that the rankings of the five outlying observations, within those five observations only, are essentially the same for $\beta$-distances as for bagdistances (only outliers 2 and 3 are exchanged in the multimodal example). When ranking the five outliers among the full sample, however, $\beta$-distances clearly tend to flag outliers better than bagdistances : not only $\beta$-distances of outlying observations are given very high (overall) ranks, but $\beta$-distances of non-outlying observations are much less spread than bagdistances, so that outliers more clearly stand out in the collection of

$\beta$-distances than in the collection of bagdistances. At least for cases (i)-(ii), these alternative (affine-equivariant) distances are tools that nicely complement bagdistances.

Of course, this is a single numerical exercise only, and a proper comparison between both types of distances should be conducted. One could of course object that an appropriate $\beta$ has to be selected when computing $\beta$-distances. But actually, considering $\beta$-distances for a whole range of $\beta$-values may also be of interest — for a group of isolated outliers, one may indeed expect a "break point" at $\beta = \beta_0$, where $\beta_0$ is the relative size of the group within the full sample. Finally, note that $\beta$-distances can be extended to the (possibly multivariate) functional case, by just integrating $\beta$-distances over "time" $t$, in the same way as bagdistances in HRS15.

## 3 Weight functions

On several occasions, HRS15 mentions the possibility to use — at least for depth-based procedures — a weight function $w(t)$ to emphasize or downweight particular regions of the time interval $U$. Surprisingly, no non-trivial (i.e., non-constant) weight functions are considered in the paper and the potential benefits of such weight functions are not discussed. In this second part of our discussion, we therefore comment on this point.

While such weighted integrated depth will not change drastically for persistent outliers (i.e., for observations whose marginal depth is low everywhere on $U$), it will prove useful for isolated outliers, provided, of course, that the weight function correctly focuses on the region supporting the abnormal behavior. Typically, outliers will contribute — when deviating from the bulk of the data — to a higher local variability[3]. It therefore seems reasonable to base $w(t)$ on (marginal) *dispersion* measures. Focusing on regions with high dispersion might decrease the isolated outliers' depths that would otherwise be averaged up due to regions in which the observations are not abnormal.

Even if one restricts to weights based on marginal dispersion measures, the possible weight functions are still numerous, and a rationale needs to be provided to favour one over the other. For functional halfspace depth, the integrand is bounded and, from affine-invariance, is insensitive to the variability within the data. Basing weight functions on robust dispersion measures (such as, e.g., the volume of the bag, which is the natural multivariate extension of the interquartile range) would provide a weighted depth that will not be influenced by most extreme data. It is therefore preferable to adopt a dispersion measure which, on the contrary, is sensitive to outliers. In the univariate functional case, we then propose using weight functions of the form

$$w_\alpha(t) \sim (\mathrm{Var}[X(t)])^\alpha, \quad \alpha \geq 0, \tag{1}$$

[3] This is not only the case for shift outliers, but also for shape outliers, provided that the multivariate functional data considered contains derivatives of the original curves.

where "∼" indicates that $w_\alpha(t)$ is normalized to integrate to one over $U$. Clearly, the boundary case $\alpha = 0$ corresponds to the unweighted functional depth, while increasing values of $\alpha$ will put more and more emphasis on regions with high marginal variances.

We briefly illustrate the use of the weights in (1) on the Octane data set. Figure 2 plots the $w_\alpha(t)$-weighted integrated (halfspace) depths of the $n = 39$ observations of this data set as a function of $\alpha \in [0, 3]$. The six shape outliers considered in HRS15 are plotted in red. The weight function $w_\alpha(t) = w_1(t)$ is displayed as a black curve in the left panel (since the "time" span is very large in this example, the actual values of $w_1(t)$ are small, and this weight function was thus rescaled for illustration purposes so as to have its maximal value equal to 0.3). Clearly, emphasis is put on the end of the time region. Doing so clearly decreases the depth of the outliers but also their ranks within the depth values (going from 16, 3, 12, 10, 5 and 15 at $\alpha = 0$ to 11, 1, 7, 6, 4 and 5 at $\alpha = 3$, respectively).

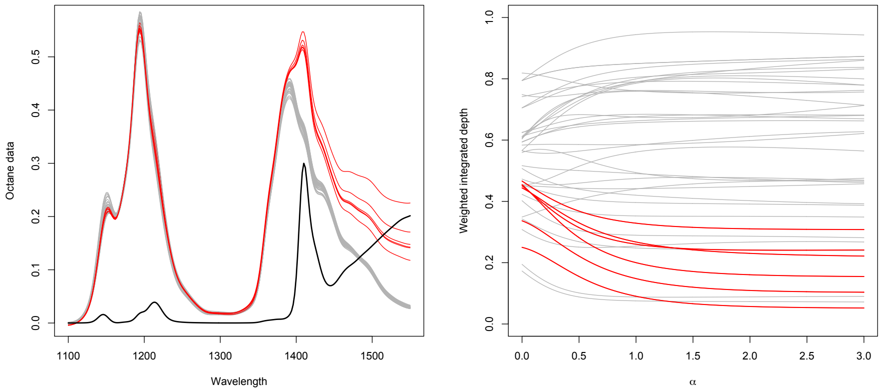


**Fig. 2** The Octane data set. (Left:) Data, with the six shape outliers plotted in red. The black curve is, up to a positive multiplicative constant, the weight function associated with $\alpha = 1$; see Section 3 for details. (Right:) Weighted functional halfspace depths, as a function of $\alpha \in [0, 3]$.

One of the other advantages of varying the weights is the possibility to discover possible masking effects. Varying the weights (which would then concentrate on areas in which few outliers are masking some others) might allow to recover some part of the hidden behavior, as we now illustrate on the basis of a simulated data example. We generated $n = 100$ curves $X_i(t)$, $t \in U = [0, 1]$, according to $X_i(\frac{j}{100}) = f(\frac{j}{100}) + Z_{ij}$, where $j = 0, 1, \ldots, 100$ and the $Z_{ij}$'s are mutually independent standard normal variables. For 80 curves (standard observations), the function $f$ is the zero function. For the remaining 20 curves, 10 were based on $f(t) = 1/5 \exp(-100(t - 0.3)^2)$ (mild outliers) and 10 were based on $f(t) = 1/2 \exp(-200(t-0.3)^2)$ (severe outliers). Figure 3 shows the resulting data, along with the ranks, as functions of $\alpha \in [0,1]$, of the 100 corresponding $w_\alpha(t)$-weighted integrated (halfspace) depths — only 30 (randomly selected) of the 80 non-outlying curves are shown in order to avoid overcrowding the pictures. For $\alpha = 0$, severe outliers (in red) clearly mask mild outliers (in orange), since the rank of most mild outliers exceeds 20. As $\alpha$ increases, this masking effect is lifted (the depths (resp. the ranks) of the orange curves decrease).

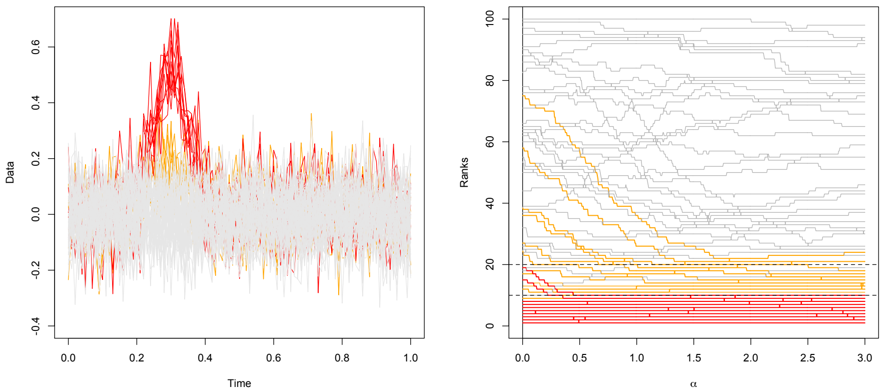


**Fig. 3** (Left:) Simulated data, where the severe (resp., mild) outliers are plotted in red (resp., orange). (Right:) Ranks, as a function of $\alpha \in [0,3]$, of the $w_\alpha(t)$-weighted integrated (halfspace) depths; see Section 3 for details.

Note that, in this simulated example, the non-outlying curves will not be assigned a small depth value, since the randomness at each time points prevents these curves to lie persistently on the edge of the data. This is in contrast with what occurs in the Octane data (Figure 2). There, while the depth of the outliers decreases, some non-outlying curves keep very low integrated depth even for large values of $\alpha$. Indeed, albeit not as drastic as before, weighted integrated depth still suffers from some limitations. For example, a severe persistent (shift) outlier and a curve lying, *uniformly in* $U$, on the opposite border of the data will be assigned a common weighted halfspace depth value of $1/n$, irrespective of their very different marginal distances to the $n-2$ remaining observations.

As already mentioned in Section 2, HRS15 argued, convincingly, that depth alone is not enough to perform outlier detection and, as a result, used distance measures to complement the purely depth-based methods. Although the integrated quantities (bagdistance or adjusted outlyingness) in the resulting outlyingness measures are not bounded anymore, the same argumentation as above remains valid and a *weighted functional bagdistance* (resp., weighted functional adjusted outlyingness) might provide extra insight on the data. We will illustrate this in the multivariate case, where weight functions can still be chosen to reflect the marginal dispersion while being sensitive tooutliers. Natural weight functions are then

$$w_\alpha(t) \sim (\det(\mathrm{Var}[X(t)]))^\alpha, \quad \alpha \geq 0,$$

where $\mathrm{Var}[X(t)]$ denotes the variance-covariance matrix of $X(t)$ and "$\sim$" still refers to normalization of the weight function considered.

As an example, we consider the tablets data set, consisting of $n = 90$ curves measured at 404 different wavenumbers (inverse of wavelengths). The first 70 curves (90 mg pills) contain 10 shift outliers lying on the lower edge of the data (in orange) while the remaining 20 (randomly selected amongst 80 250mg pills) are shape outliers (in red). We treat this functional data set as a bivariate one, where the absorbance curves are coupled with their first derivatives (baseline corrections, that typically allow to detect shift outliers, were not considered since such outliers are detected anyway). As can be seen from Figure 4, both orange and red outliers have large integrated bagdistances at $\alpha = 0$. Larger values of $\alpha$, however, allow to discriminate between both groups. Indeed, while the shift outliers keep roughly the same bagdistances for all $\alpha$, the weighted functional bagdistances increase for the shape outliers, revealing the fact that their behavior exhibits variations in time. This is further illustrated on the rank plot, where all (but one) red outliers increase in rank as $\alpha$ grows.

The examples provided in this section suggest that weight functions may be useful for depth-based functional outlier detection. Of course, much remains to be investigated and further comparison between the various methods should be conducted. Natural questions include, but are not limited to : (i) what other dispersion measures could be used in the weight function? (ii) Are dispersion measures the best bases for weight functions? (iii) How sensitive to outliers should weight functions be?

## References


1. Agostinelli, C., Romanazzi, M.: Local depth. J. Statist. Plann. Inference **141**(2), 817–830 (2011)
2. Hubert, M., Rousseeuw, P., Segaert, P.: Multivariate functional outlier detection. Statistical Methods and Applications, p. to appear (2015)
3. Liu, R.Y.: On a notion of data depth based on random simplices. Ann. Statist. **18**(1), 405–414 (1990)
4. Paindaveine, D., Van Bever, G.: From depth to local depth : a focus on centrality. J. Amer. Statist. Assoc. **105**, 1105–1119 (2013)
5. Serfling, R.: Depth functions in nonparametric multivariate inference. In: Data depth: robust multivariate analysis, computational geometry and applications, *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, vol. 72, pp. 1–16. Amer. Math. Soc. (2006)
6. Serfling, R.: Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. J. Nonparametr. Stat. **22**, 915–926 (2010)
7. Serfling, R.J., Zuo, Y.: Discussion of "multivariate quantiles and multiple-output regression quantiles: From $L_1$ optimization to halfspace depth", by M. Hallin, D. Paindaveine, and M. Šiman. Ann. Statist. **38**, 676–684 (2010)

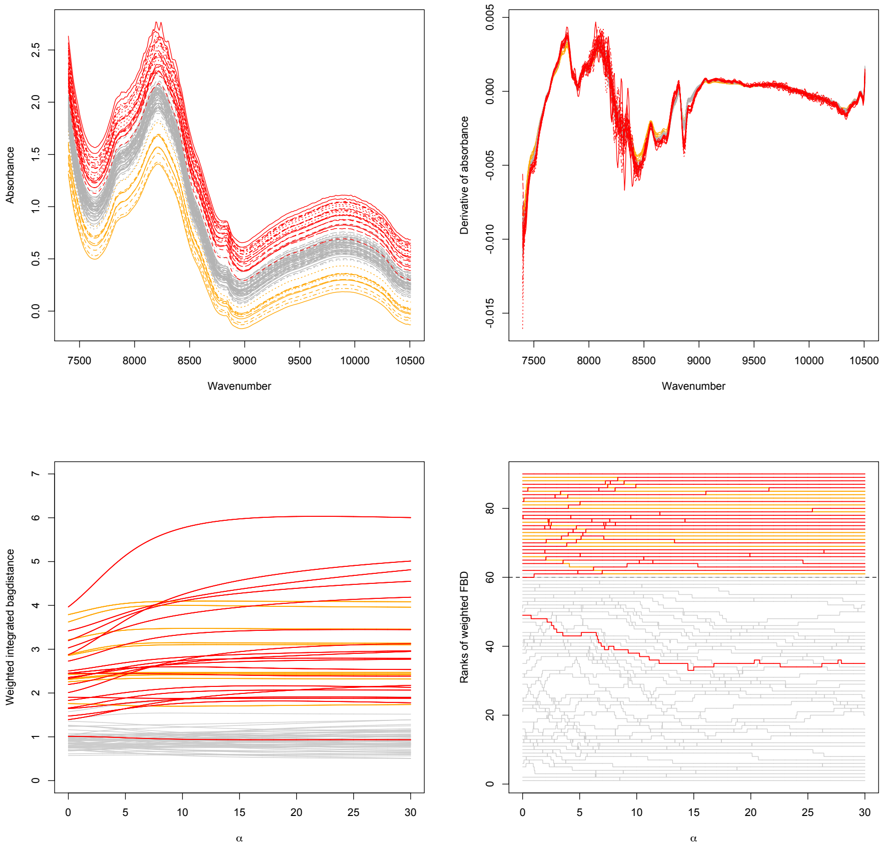


**Fig. 4** The tablets data set. (Top left:) Absorbance curves. Shift outliers (resp., shape outliers) are shown in orange (resp., in red). (Top right:) The corresponding derivatives. (Bottom left:) $w_\alpha(t)$-weighted integrated (halfspace) bagdistances, as a function of $\alpha \in [0, 30]$. (Bottom right:) Ranks of these bagdistances, still as a function of $\alpha \in [0, 30]$.

8. Tukey, J.W.: Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2, pp. 523–531. Canad. Math. Congress, Montreal, Que. (1975)
9. Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Statist. **28**(2), 461–482 (2000)