Directional Outlyingness applied to distances between Genomic Words

Ana Helena M. P. Tavares[1]
ahtavares@ua.pt
Vera Afreixo[1]
vera@ua.pt
Paula Brito[2]
mpbrito@fep.up.pt
Peter Filzmoser[3]
peter.filzmoser@tuwien.ac.at

[1] Department of Mathematics & iBiMED - Institute of Biomedicine
Universty of Aveiro, PORTUGAL
[2] FEP & LIAAD - INESC TEC
University of Porto, PORTUGAL
[3] Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology, AUSTRIA

Abstract

The detection of outlier curves/images is crucial in many areas, such as environmental, meteorological, medical, or economic contexts. In the functional framework, outlying observations are not only those that contain atypically high or low values, but also curves that present a different shape or pattern from the rest of the curves in the sample. In this short paper, we mention some recent methods for outlier detection in functional data and apply a recently proposed [5] measure, the *directional outlyingness*, and the *functional outlier map* to detect words with outlying distance distribution in the human genome.

1 Introduction

In the functional framework, an outlying observation is not only one that contains atypically high or low values ("magnitude outliers"), but also a curve that presents a different shape or pattern than the rest of the curves of the sample ("shape outliers") [3]. While the first might be easily detected, the latter are often masked among the rest of the curves and thus more difficult to detect.

Different methods for outlier detection in functional data have been developed. Some of those rely on notions of functional depth ([1, 4, 6]. To visualize functional data and investigate the existence of possible outliers, Sun and Genton [6] proposed the functional boxplot and Arribas-Gil and Romo [1] introduced the outliergram. Based on robust principal component scores, Hyndman and Shang [3] proposed graphical tools for visualizing functional data and identifying functional outliers, *e.g.* the bagplot. A very recent approach to detect outlying functions was proposed by Rousseeuw *et al*. [5]. They introduced the directional outlyingness (DO) measure which assigns a robust value of outlyingness to each gridpoint of the function domain, and proposed a procedure that allows detecting outlying functions and outlying parts of a function.

In this work, we consider data arising from the human genome (reference assembly), more precisely, distances between consecutive occurrences of genomic words, and intend to detect words with atypical distance distribution. For fixed word length, the set of 4^k distance distributions can be seen as a sample of curves, which may be treated as functional data. We apply the DO measure to identify atypical distance distributions between genomic words.

1.1 Inter-word distance distribution

Consider the alphabet formed by the four nucleotides $\mathcal{A} = \{A, C, G, T\}$, and let s be a symbolic sequence of length N defined in $\mathcal{A}$. A genomic word, w, is a sequence of length k defined in $\mathcal{A}$. Assuming that the sequence is read through a sliding window of length k, the inter-word distances are the differences between the positions of the first symbol of consecutive occurrences of that word. For example, the inter-CG distances for the DNA sequence $s = ACGTCGATCCGTG$ are 3 and 5.

For each word w, we can define the inter-word distance distribution, f_w, associated with a genomic sequence. In sequences generated by a random process it is expect that distance distributions between genomic words are well fitted by some kind of exponencial law. However, in real genomic sequences we observe distances with peak frequencies and non-expected behaviours.

1.2 Directional Outlyingness

Rousseeuw *et al*. [5] proposed a procedure to detect outlying functions or outlying parts of a function, assigning a robust value of outlyingness to each gridpoint of the function domain. Based on the Stahel-Donoho outlyingness of a point $y \in \mathbb{R}$ relative to a univariate sample $Y = \{y_1, \ldots, y_m\}$, they introduced the notion of *directional outlyingness* (DO), which takes the possible skewness of the distributions into account. Quoting the authors, the main idea is "to split the sample into two halfsamples and then to apply a robust scale estimator to each of them" [5, *pag.3*],

$$DO(y;Y) = \begin{cases} \frac{y - med(Y)}{S_a(Y)} & \text{if } y \geq med(Y) \\ \frac{med(Y) - y}{S_b(Y)} & \text{if } y \leq med(Y) \end{cases}, \tag{1}$$

where S_a and S_b are robust scale estimates for the subsample of points above and for the subsample of points below the median, respectively[1].

The DO of a point $\mathbf{y} \in \mathbb{R}^n$ relative to a n-variate sample $\mathbf{Y} = \{\mathbf{Y_1}, \ldots, \mathbf{Y_1}$ is defined by means of univariate projections, applying the principle that a multivariate point is outlying with respect to a sampleset if it stands out in at least one dimension,

$$DO(\mathbf{y};\mathbf{Y}) = \sup_{v \in \mathbb{R}^n} DO(\mathbf{y}'v;\mathbf{Y}'v)\,, \tag{2}$$

Due to the impossibility of projecting on all directions, the computation of multivariate DO relies on approximate algorithms.

Consider a function x and a functional dataset $X = \{X_1, \ldots, X_m\}$, formed by n-variate functions with univariate domain. At each domain point, t, it is possible to compute the DO of $x(t)$ with respect to the set of values taken by the other functions in the same domain point. Computing a kind of average of those values, a global outlyingness measure of x with respect to X may be achieved. The *functional directional outlyingness* (fDO) of a function x with respect to the functional dataset X, proposed by [5], is defined as

$$fDO(x;X) = \sum_{j=1}^{T} DO(x(t_j);X(t_j))W(t_j)\,, \tag{3}$$

where $W(.)$ is a weight function, which sums one, and $\{t_1, \ldots, t_T\}$ is a discrete set of points of the domain where the functions are observed. The variability of the DO values of a function x is measured by

$$vDO(x;X) = \frac{stdev_j(DO(x(t_j);X(t_j)))}{1 + fDO(x;X)}\,. \tag{4}$$

To visualize the outliers the *functional outlier map* (FOM) is used, a graphical tool firstly proposed in [2] and extended to the DO measure by [5]. The FOM shows a scatter plot of the pairs (fDO, vDO) associated with each curve Y_i, and a fence, drawn from a cutoff rule discussed in [5], which allows puting outliers in evidence. Points in the lower left part of the FOM represent regular functions, holding central positions in the data set. Points in the upper left have low fDO and high vDO, which may be associated with functions with local outliers. Points in the upper right part of the FOM have high fDO and vDO, corresponding to functions which deviate strongly from the majority of the sample.

The method may be applied to multivariate functional data, from univariate curves to images and video data.

[1] The authors used a one-step M-estimator with Huber ρ-function, among many available robust estimators, due to its fast computation and favorable properties [5, *pag.*4].

2 Experimental Results

2.1 Data set

In this study, we used the complete DNA sequences of reference assembly for human genome (GRCh38.p2) downloaded from the website of the National Center for Biotechnology Information. We processed the assembled chromosomes available as separate sequences and studied every word formed by k consecutive nucleotides, with $1 < k \leq 5$.

We computed the inter-word distance distribution of each word, f_w. The dataset contains functions with irregular behaviour revealing several unexpected strong peaks, as the word length increases. The rates of change of the curves may comprise important features on the shape of the data. The inter-word distance distributions were treated as functional data and the dynamic behaviour of the curves was incorporated, by numerically computing their first derivative. To resume, for each word length k, we have a functional dataset formed by 4^k bivariate functions, which response is f_w and its derivative. Since the domains of the curves may be different, we define a cutoff distance, $d^k_{\max}$, associated with each word length.

The computations were performed using the R language. For computing DO and fOM we used R-code provided by Rousseeuw *et al*. at *http://wis.kuleuven.be/stat/robust/software*.

2.2 Detection of outlying inter-word distance distributions

In the present context, the detection of outlier functions obviously depends on the cutoff of the function domain, $d^k_{\max}$. In this first exploratory study, we perform the analysis considering several cutoff distances.

For the dinucleotide case, $k = 2$, the dataset consists of 16 functions defined over a discrete interval (figure 1, top left). We observe that, for short distances, the f_{CG} curve (red) deviates from the other curves. For word length 3, the dataset comprises oscillating functions, but with no evidence of strong peaks. As the word length increases, several distributions have a more expressive oscillating behaviour revealing strong and unexpected peaks. Figure 1 (bottom) shows the f_w for all words of length $k = 5$, where we observe the existence of peaks along a substantial part of the domain.

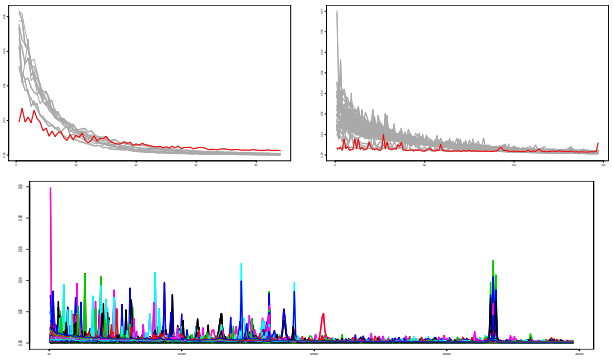

Figure 1: Inter-word distance distributions for different word lengths: $k = 2$, formed by 16 curves, $d^2_{\max} = 90$ (top left); $k = 3$, formed by 64 curves, $d^3_{\max} = 150$ (top right); $k = 5$, formed by 1024 curves, $d^5_{\max} = 400$ (bottom).

The FOMs in figure 2, for the $k = 2$ dataset, show that *CG* data have both high fDO and high vDO. Indeed, for short distances, the *CG* curve deviates from the other curves. However the identification of this curve as outlier depends on the $d^2_{\max}$ value. For $k = 3$, the procedure allows identifying the existence of distributions with both high fDO and high vDO (figure 3, left), which correspond to "flat" distributions, *i.e* distributions with under represented short distances. For $d^3_{\max} = 150$, the TCG curve is identified as outlier (figure 1, middle, in red). Increasing $d^3_{\max}$, other curves with the same behaviour are flagged as outlying functions.

The most interesting case in our analysis is the $k = 5$ dataset. This functional dataset comprises a large proportion of distributions with strong and unexpected peaks, which occur at short and long distances. Furthermore, it reveals clusters of distances where different functions reach unexpected strong peaks. Figure 3 (right) shows the resulting FOM, which reveals the presence of 17 outlying cases, though 10 of them are relatively

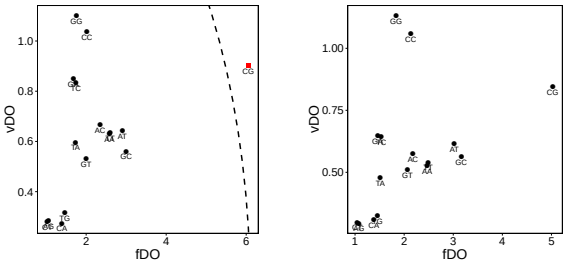

Figure 2: FOM of the $k = 2$ dataset. The detection of outliers depends on the function domain cutoff: for $d^2_{\max} = 90$, one point is flagged as outlier (left); for $d^2_{\max} = 80$ there are no outliers (right).

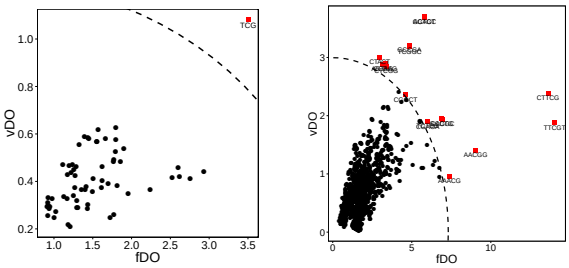

Figure 3: FOM of the inter-word distance distributions data: $k = 3$ data reveal one outlier (left); $k = 5$ data reveal 17 outliers (right).

close to the fence. Analysing the flagged cases one by one, we conclude that the method captures curves with peaks at subdomains where no other peak occurs, as well as curves whose pattern strongly differs from the majority. The two points in the middle right - $CTTCG$ and $TTCGT$ - correspond to functions that deviate strongly from the majority of the curves, they are "shape outliers". Points in the upper left - $AGTGC$, $GCACT$, $GCCGA$, $TCGGC$ - correspond to functions with low fDO but highest vDO values, with outlying behaviour in a small part of the domain. Indeed, the AGTGC curve shows a peak frequency around distance 210. Despite the low peak magnitude, it is located in a interval of the domain with absence of peaks (figure 4, right). Figure 4 (left) confronts the TTCGT curve with the complete set of functions, exposing an unusual curve pattern.

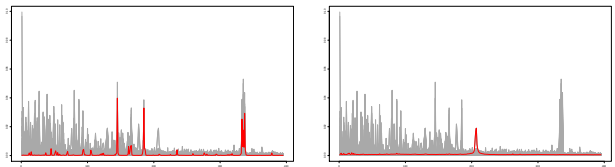

Figure 4: Outlying distributions:; a shape outlier for $w = TTCGT$ (left); local outlier for $w = AGTGC$ (right).

3 Conclusions

The preliminary results indicate that the DO procedure is promising for our problem, puting in evidence outlying inter-word distance distributions masked among the rest of the curves. In the case where the functional dataset comprises a large proportion of functions with strong peaks, spreading over a large part of the domain (*e.g*. $k = 5$ dataset), it is difficult to detect outlying behaviours. The method was able to capture outlying functions distinct from magnitude outliers, highlighting curves whose shape strongly differs from the majority. In particular, it allowed detecting functions with peaks at subdomains where no other peaks occur, as well as functions with several strong peaks. Further analysis will be performed for longer words; future work will investigate the relation between the cutoff in the functions domain and cutoff values for outlier detection.

4 Acknowledgements

This work was supported by Portuguese funds through the iBiMED-Institute of Biomedicine and the Portuguese Foundation for Science and Technology (FCT) within projects UID/BIM/04501/2013 and UID/EEA/50014/2013. AT is supported by FCT PhD fellowship PD/BD/105729/2014. PB is also financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961.

References

[1] Ana Arribas-Gil and Juan Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4): 603–619, 2014.

[2] Mia Hubert, Peter J. Rousseeuw, and Pieter Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24 (2):177–246, 2015. (with discussion).

[3] Rob J. Hyndman and Han Lin Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010.

[4] Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104 (486):718–734, 2009.

[5] Peter J. Rousseeuw, Jakob Raymaekers, and Mia Hubert. A measure of directional outlyingness with applications to image data and video. *arXiv preprint arXiv:1608.05012*, 2016.

[6] Ying Sun and Marc G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.