A notion of depth for curve data

Pierre Lafaye De Micheaux [1] & Pavlo Mozharovskyi [2] & Myriam Vimond [3]

[1] *School of Mathematics and Statistics, University of New South Wales;* lafaye@unsw.edu.au
[2] *CREST (Ensai, Université Bretagne Loire);* pavlo.mozharovskyi@ensai.fr
[3] *CREST (Ensai, Université Bretagne Loire);* myriam.vimond.@ensai.fr

Résumé. Initialement introduite par John W. Tukey (1975), la profondeur statistique des données est une fonction qui détermine la centralité d'un point de l'espace par rapport à un nuage de points ou à une distribution de probabilité. Au cours des dernières décennies, la profondeur des données a rapidement évolué vers un mécanisme puissant qui s'avère utile dans divers domaines de la science. Dernièrement, l'extension de profondeur des données dans le cadre fonctionnel a attiré beaucoup d'attention. Nous suggérons une notion basée sur la profondeur de données de Tukey appropriée pour des données représentées par des trajectoires ou des courbes non-paramétrées. Cette profondeur basée sur la longueur des trajectoires ou des courbes hérite à la fois de la géométrie euclidienne et des propriétés fonctionnelles, tout en surmontant certaines limitations des approches précédentes. Les applications de cette profondeur de courbe comprennent l'imagerie cérébrale et la reconnaissance de motifs écrits.

Mots-clés. Courbes non-paramétrées, profondeur des données, statistique non-paramétrique, imagerie cérébrale, apprentissage supervisé.

Abstract. Following the seminal idea of John W. Tukey (1975), statistical data depth is a function that determines centrality of an arbitrary point w.r.t. a data cloud or a probability measure. During the last decades, data depth rapidly developed to a powerful machinery proving to be useful in various fields of science. Recently, implementing the idea of depth in the functional setting attracted a lot of attention among theoreticians and applicants. We suggest a Tukey-based notion of data depth suitable for data represented as curves, or trajectories, which inherits both Euclidean-geometry and functional properties but overcomes certain limitations of the previous approaches. It can be shown that the Tukey curve depth satisfies the requirements posed on the general depth function, which are meaningful for trajectories. Applications of the Tukey curve depth include brain imaging and written patterns recognition.

Keywords. Curve data, data depth, nonparametric statistics, brain imaging, supervised learning.

1 Introduction

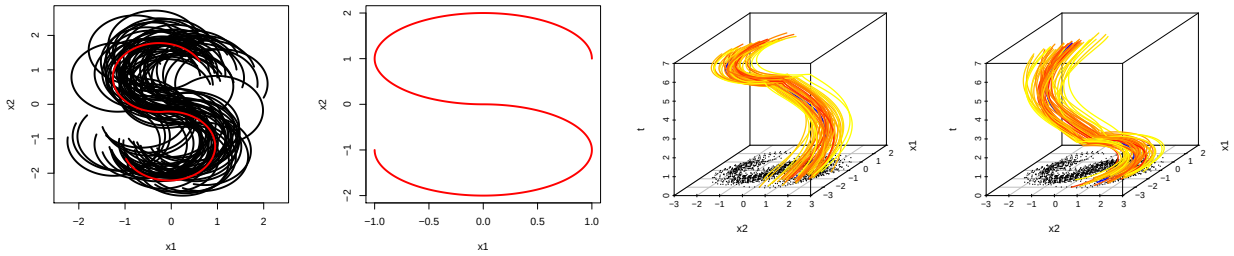
Suggested as a descriptive statistics by Tukey (1975) the idea of data depth rapidly developed to a powerful machinery finding applications in various areas. It consists in defining a function $D(\boldsymbol{x}|\boldsymbol{X})$ that determines centrality of a point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. a data cloud $\boldsymbol{X}$ in $\mathbb{R}^d$ (or a probability measure). Roughly speaking a depth function provides a center-outward ordering of points $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. the empirical distribution of sample $\boldsymbol{X}$. According to the literature (Dyckerhoff, 2004; Mosler, 2013; also Zuo and Serfling, 2000), a statistical depth function has to satisfy desirable properties: affine invariance, vanishing at infinity, monotonocity w.r.t. the deepest point, and upper semicontinuity. Numerous depth functions have been developed, which fulfill the above postulates to different extent, see for instance Zuo and Serfling (2000) and Mosler (2013) for a survey.

During the last decades and especially recently, a number of works related to depth and its applications have been published, constantly opening new domains: multivariate data analysis (Liu *et al.*, 1999), statistical quality control (Liu and Singh, 1993), classification (Jöornsten, 2004; Lange *et al.*, 2014), tests for multivariate location and scale (Liu, 1992; Dyckerhoff, 2002), multivariate risk measurement (Cascos and Molchanov, 2007), robust linear programming (Bazovkin and Mosler, 2015), *etc.* A natural direction of this process is adjusting existing and developing new definitions applicable to further ways of data registration, and eventually other types of data. For instance, the notion of data depth has been extended to the functional setting. Posing different further restrictions on the space of functions, a number of notions of depth for functional data have been and are being developed. These can be roughly categorized into two groups: directly employing the multivariate depth in the infinite dimensional space, see Chakraborty and Chaudhuri (2014), Mosler and Polyakova (2012); and averaging a particular (univariate) depth over the time interval according to the suggestion of Fraiman and Muniz (2001), see Lòpez-Pintado and Romo (2009) for simplicial depth, Lòpez-Pintado and Romo (2011) for univariate and Claeskens *et al.* (2014) for multivariate halfspace depth. Recently Nieto-Reyes and Battey (2016) proposed an axiomatization of depth function for functional data accounting for topological features.

In this paper, we present a notion of data depth for curve data. Examples of such data can be any stochastic process representable as paths in a multivariate space, *e.g.* trajectories of animals' displacement, highly non-synchronized in time at the beginning and at the end economic processes, or — the application in the center of our attention — axons connecting brainstem with the cortex. First, we show that the functional data depth is not accurate for curve data. Then we define a data depth for curve data.

2 Motivation

Consider a set of paths, or trajectories, eventually curves in $\mathbb{R}^d$. For the first view, functional data depth can be naturally adapted to provide a center outward ordering. One



(a) (b) (c) (d)

Figure 1: A set of curves with one curve in red (a), an "ideal" curve for the parametrization example (b), depth-colored curves for parametrizations A (c) and B (d). The depth of each curve is calculated w.r.t. to the same sample. The used depth notion is the multivariate functional halfspace depth by Claeskens *et al.* (2014). The depth increases from yellow to red, the deepest curve has blue color.

possibility would be to choose a direction mostly aligned with the majority of the curves and use this as the argument axis while intersections of the curves with the orthogonal $(d-1)$-dimensional affine subspaces would serve as functions' evaluations; after the functional data depth can be employed. First immediate problem connected with this approach comes from the fact that the obtained dependency may not have functional character for some such directions. Further, depth values may depend on the chosen direction.

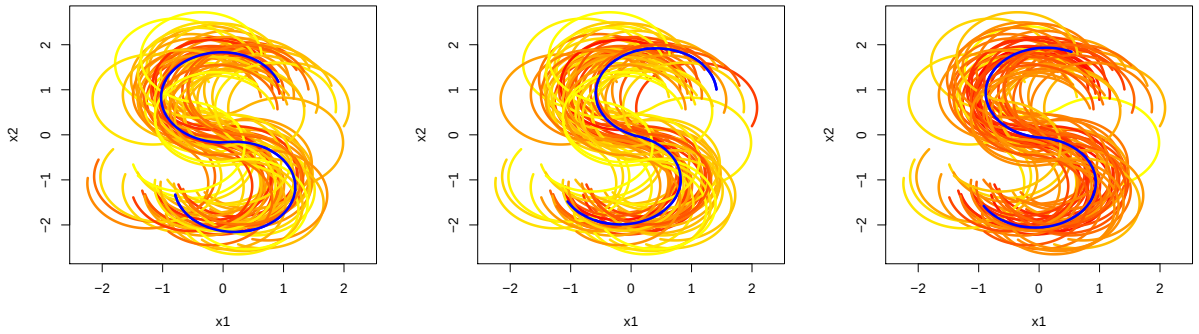
More appropriate way to proceed is to parametrize curves in a certain manner. But then depth of a curve will be dependent on its parametrization and parametrizations of the other curves in the set. This parametrization may be not obvious, especially when curves are observed as sets of points after a certain process is finished and no information about its development is available. We demonstrate this last issue on an artificial example.

Regard a set of two-dimensional curves, see Figure 1 (a), which we demonstrate on the "ideal" curve shown in Figure 1 (b). We consider two parametrizations of spatial coordinates x_1 and x_2, (A) and (B) defined as follows,

$$\begin{cases} x_1(t) = -\big(\cos(t)+1\big)\mathbb{1}\{t < \dfrac{3\pi}{2}\} \; - \; \big(\cos(3t-3\pi) \;\; +1\big)\mathbb{1}\{t \geq \dfrac{3\pi}{2}\} + 1, \\ x_2(t) = \quad \big(\sin(t)+1\big)\mathbb{1}\{t < \dfrac{3\pi}{2}\} - \big(\sin(3t-3\pi)+1\big)\mathbb{1}\{t \geq \dfrac{3\pi}{2}\}; \end{cases} \tag{A}$$

$$\begin{cases} x_1(t) = -\big(\cos(3t)+1\big)\mathbb{1}\{t < \dfrac{\pi}{2}\} - \big(\cos(t+\pi)+1\big)\mathbb{1}\{t \geq \dfrac{\pi}{2}\} + 1, \\ x_2(t) = \quad \big(\sin(3t) \;\; +1\big)\mathbb{1}\{t < \dfrac{\pi}{2}\} \; - \big(\sin(t+\pi) \;\; +1\big)\mathbb{1}\{t \geq \dfrac{\pi}{2}\}. \end{cases} \tag{B}$$

After having employed the multivariate functional hafspace depth by Claeskens *et al.* (2014) to calculate depth of each curve w.r.t. the sample, the respective orderings and



(a) (b) (c)

Figure 2: Depth-colored curves for parametrization A (a); depth-colored curves for parametrization B (b). Both calculated using the multivariate functional halfspace depth by Claeskens *et al.* (2014). The proposed depth notion (c) according to (1). The depth of each curve is calculated w.r.t. to the same sample. The depth increases from yellow to red, the deepest curve has blue color.

the deepest curves are pictured in Figure 2 (a) and (b) for parametrizations A and B, respectively. One observes that the depth-induced order differs. In addition one can see that some of rather outlying curves have relatively high depth (orange or even close to red color in Figure 2 (a) and (b)) while those closer to center can have lower depth values (yellow color in Figure 2 (a) and (b)).

3 A notion of data depth for unparametrized curves

Let $(\mathbb{R}^d, |\cdot|_2)$ be the Euclidean space. A *path* in $\mathbb{R}^d$ is a continuous map from $[0,1]$ to $\mathbb{R}^d$. Under a path neither a map nor its image is meant: two maps visiting the same collection of points (the same *curve*) are equivalent. In the current section, we define a measure of centrality for curves sticking to the seminal philosophy of John W. Tukey.

The space of unparametrized curves Γ is the quotient space of the space of continuous functions defined on the interval $[0,1]$, $\mathcal{C}([0,1], \mathbb{R}^d)$, and an equivalence relation is defined as follows: two continuous functions γ_1 and γ_2 describe the same curve if and only if there exist two monotone continuous functions $\phi_i : [0,1] \to [0,1]$, $i = 1, 2$, such that $f_1 \circ \phi_1 = f_2 \circ \phi_2$. We denote $[\gamma]$ the associated equivalence class of $\gamma \in \Gamma$. The space Γ endowed with the following metric:

$$d([\gamma_1], [\gamma_2]) = \inf\{\|\gamma_1 - \gamma_2\|_\infty, \, \gamma_1 \in [\gamma_1], \, \gamma_2 \in [\gamma_2]\} \, .$$

is a separable, metric, complete (polish) space.

The *length* of a path $\gamma : [0,1] \to \mathbb{R}^d$ *w.r.t. the Euclidean distance* is defined as

$$L(\gamma) := \sup\Big\{\sum_{i=1}^{n} |\gamma(t_i) - \gamma(t_{i-1})|_2 \,:\, (t_i)_{i=0,\dots,n} \text{ is a partition of } [0,1]\Big\}\,.$$

We consider only *rectifiable* curves, that means the length of curves are assumed to be finite.

Let $\mathcal{Y} = \{[\gamma_1], ..., [\gamma_n]\}$ be a sample of curves in Γ. For a curve $[\gamma] \in \Gamma$, we define a depth of its arbitrary point $\boldsymbol{x} \in [\gamma]$ w.r.t. the sample $\mathcal{Y}$ as

$$D(\boldsymbol{x}|[\gamma], \mathcal{Y}) = \inf\left\{ v\left(\frac{1}{n}\sum_{i=1}^{n} \frac{L([\gamma_i] \cap \mathcal{H})}{L([\gamma_i])}, \frac{L([\gamma] \cap \mathcal{H})}{L([\gamma])}\right) \,:\, \mathcal{H} \text{ closed halfspace, } \boldsymbol{x} \in \mathcal{H}\right\},$$

where for $a, b \in \mathbb{R}$

$$v(a,b) = \begin{cases} 0 & \text{if } a = 0 \text{ and } b = 0\,, \\ \frac{a}{b} & \text{otherwise.} \end{cases}$$

Now, we define the *Tukey curve depth* of $[\gamma]$ w.r.t. $\mathcal{Y}$ in Γ as:

$$D_\Gamma([\gamma]|\mathcal{Y}) = \int_0^1 D(\boldsymbol{x}|[\gamma], \mathcal{Y}) \frac{d\gamma}{L([\gamma])}\,. \tag{1}$$

4 Discussion

This work introduces a novel notion of data depth operating on the space of nonparametrized curves. Being invariant to the manner of traversing the curve the Tukey curve depth delivers coherent results exploiting purely the geometry of the data. For the above example, *e.g.*, the depth-colored curves are plotted in Figure 2 (c). One can observe the proper position of the deepest curve, and that more centrally lying curves are rather deep (red), with smaller depth for those having somewhat outlying shape (orange) and low depth for those seriously differing in shape and outlying in location (yellow). Further, the developed depth notion adapts properties of the statistical depth function transferable to the space of curves, such as Euclidean invariance or vanishing at infinity. This measure of centrality is useful as descriptive statistics, when testing for homogeneity, or in classification and finds applications, *e.g.*, in description of brain imaging or written pattern recognition.

References

[1] Bazovkin, P. and Mosler, K. (2015). A general solution for robust linear programs with distortion risk constraints. *Annals of Operations Research*, 229, 103–120.

[2] Cascos, I., and Molchanov, I. (2007). Multivariate risks and depth-trimmed regions. *Finance and Stochastics*, 11, 373–397.
[3] Chakraborty, A. and Chaudhuri, P. (2014). On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66, 303–324.
[4] Claeskens, G., Hubert, M., Slaets, L. and Vakili K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109, 411–423.
[5] Dyckerhoff, R. (2002). Inference based on data depth. Chapter 5, in: Mosler, K., *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach*, Springer, New York.
[6] Dyckerhoff, R. (2004). Data depths satisfying the projection property. *AStA – Advances in Statistical Analysis*, 88, 163–190.
[7] Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *TEST*, 10, 419–440.
[8] Jörnsten, R. (2004). Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*, 90, 67–89.
[9] Lange, T., Mosler, K., and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers*, 55, 49–69.
[10] Liu, R. Y. (1992). Data depth and multivariate rank tests. In Y. Dodge (ed.), L_1-*Statistical Analysis and Related Methods*, North-Holland, Amsterdam, 279–294.
[11] Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27, 783–858. With discussion.
[12] Liu, R. Y., and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88, 252–260.
[13] López-Pintado, S. and Romo, J. (2009). On the concepth of depth for functional data. *Journal of the American Statistical Association*, 104, 718–734.
[14] López-Pintado, S. and Romo, J. (2011). A half-region depth for functional data. *Computational Statistics and Data Analysis*, 55, 1679–1695.
[15] Mosler, K. (2013). Depth statistics. In: Becker, C., Fried, R., and Kuhnt, S. (eds.), *Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather*, Springer, Berlin, Heidelberg, 17–34.
[16] Mosler, K. and Polyakova, Y. (2012). General notions of depth for functional data. `arXiv:1208.1981v1 [stat.ME]`.
[17] Nieto-Reyes, A. and Battey, H. (2016). A topologically valid definition of depth for functional data. *Statistical Science*, 31, 61–79.
[18] Tukey, J. W. (1975). Mathematics and the picturing of data. In: James, R. D. (ed.) *Proceedings of the International Congress of Mathematicians (Volume 2)*, Canadian Mathematical Congress, 523–531.
[19] Zuo, Y., and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28, 461–482.