Working paper

2019-09

Statistics and Econometrics
ISSN 2387-0303

A Depth for Censured Functional Data

Antonio Elías, Raúl Jiménez, Anna M. Paganoni and Laura M. Sangalli

uc3m | Universidad Carlos III de Madrid

A Depth for Censured Functional Data

Antonio Elías[*], Raúl Jiménez[*], Anna M. Paganoni[†] and Laura M. Sangalli[†]

July 10, 2019

Abstract

Censured functional data are becoming more recurrent in applications. In those cases, the existing depth measure are useless. In this paper, an approach for measuring depths of censured functional data is presented. Its performance for finite samples is tested by simulation, showing that the new depth agrees with a integrated depth for uncensured functional data.

Keywords: Functional data, Partially observed data, Integrated depth.

[*]Deparment of Statistics, Universidad Carlos III de Madrid, Getafe, Madrid, Spain.

[†]Laboratory for Modeling and Scientific Computing MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy.

1 Introduction

For a random sample of functions $[a, b] \to \mathbb{R}$, partially observed functional data (POFD) refers to the case where the records of the functions are available only on subsets of $[a, b]$. We deal with the challenging case where partial observability could occur systematically on any observation of a functional data set, not only to isolated cases that affect to few observations.

In practise, partially observed data appear by different reasons and have been reported in different areas of research. Many medical data sets are recorded through periodical examinations and typical sources of censoring are patients missing revisions or devices failing to record. Some examples of such case studies are ambulatory blood pressure, health status of human immunodeficiency virus (HIV), growth curves and evolution of lung function (James et al., 2000; James and Hastie, 2001; Delaigle and Hall, 2013; Kraus, 2015; Delaigle and Hall, 2016). Also with medical purposes, Sangalli et al. (2009) present a case study about aneurysm (Sangalli et al., 2014b) where functions are partially observed at the extremes of the domain. Here the source of partial observability arises due to a prior reconstruction from three-dimensional arrays and a posterior processing to make them comparable across subjects. In Demography, it is common that age-specific mortality rates for older ages are not completely observed due to the decreasing number of survivors (University of California and for Demographic Research , Germany) and this cohort is the focus of actuarial science studies (D'Amato et al., 2011). Other examples involve electricity supply functions that may be not complete observed because suppliers and buyers typically agree prices and quantities depending on the market conditions (Kneip and Liebl, 2017; Liebl and Rameseder, 2019). Finally, partially observability could also emerge by-products of a functional data preprocessing method such as alignment (Marron et al., 2015; Sangalli et al., 2010).

The literature has tackled partial observability in several Functional Data Analysis (FDA) problems; including classification (Delaigle and Hall, 2013), discriminant analysis (James and Hastie, 2001), functional principal components (James et al., 2000; Yao et al., 2005) and linear prediction (Delaigle and Hall, 2016). Many authors propose the reconstruction or estimation of the missing parts, (Delaigle and Hall, 2016; Goldberg et al.,

2014; Kneip and Liebl, 2017; Kraus, 2015). The majority of the proposals, directly or indirectly, makes use of the so-called Missing-Completely-At-Random (MCAR) assumption and any violation of this assumption lead to undesirable results. Roughly, MCAR states an independent relationship between the censoring process and the functional process of interest. Remarkably, Liebl and Rameseder (2019) proposed a mean and covariance estimator applicable when MCAR is violated by allowing a particular dependency relationship.

Despite the important progress in the topic, there is a lack of depth notion for partially observed functional data and, therefore, a lack of all their potential uses and applications. Thus, the introduction of a suitable definition of depth measure is a contribution to the literature in two ways.

In this work, we introduce a general building-block depth definition suitable for POFD. The novelty here is that it takes into account the uncertainty related with the unobserved fragments in such a way that regions of high density are rewarded and regions of low density are penalized. The first component of the building-block is an integrated functional depth (Nagy et al., 2016) and the second one a weighting function that penalizes the domain where the sample is poorly observed.

This article is as follows: Next section presents the definition of the censured functional depth and its properties. In Section 3 we test by simulation the performance of the new measure. Some conclusions are presented at the end.

2 Depth Measures for POFD

Given a measurable space S equipped with a sample space Ω and some fixed σ-algebra, consider the collection of all probability measures on S, denoted here by $\mathcal{P}_\Omega$. A depth measure is a mapping that assigns to each $(\omega, P) \in \Omega \times \mathcal{P}_\Omega$ a value into $[0, 1]$ corresponding to the centrality of ω with respect to P. The larger the depth value, the more central will be ω with respect to P.

We consider the simplest functional data scenario in which $\Omega = \mathcal{C}([0, 1])$. In this context, the existing depth measures may be categorized as integrated or non-integrated (Nagy et al., 2016). In this paper, we are focused on the first, category, which includes many popular functional depths, such as the Fraiman and Muniz Depth (Fraiman and Muniz,

2001), the Modified Band Depth (MBD) of López-Pintado and Romo (2009), the Modified Half Region Depth (López-Pintado and Romo, 2011), all of them particular cases of the functional depths defined by Claeskens et al. (2014). For defining an integrated functional depth, we first consider an univariate depth $D : \mathbb{R} \times \mathcal{F} \to [0, 1]$, $\mathcal{F}$ being the collection of all probability distribution functions on $\mathbb{R}$. The literature always discuss the four key properties that such depth should satisfy (Liu et al., 1999). We remark two additional and necessary properties for the definition we propose in this article. Namely:

P1 (*Weak continuity*): For any $\{F_n\} \subset \mathcal{F}$ such that $\lim_{n\to\infty} F_n(x) = F(x)$, at each continuity point x of $F \in \mathcal{F}$, D satisfies

$$\sup_{x\in\mathbb{R}} |D(x, F_n) - D(x, F)| \xrightarrow[n\to\infty]{} 0.$$

P2 (*Measurability*): $(x, F) \mapsto D(x, F)$ is jointly Borel measurable and $D(\cdot, F) \not\equiv 0$ for all $F \in \mathcal{F}$.

P1 and P2 correspond to properties D_6 and D_7 discussed by Nagy et al. (2016). Following Claeskens et al. (2014), we define the functional depth in terms of D. Hereinafter, $P \in \mathcal{P}_{\mathcal{C}([0,1])}$ and $X \sim P$.

Definition 1. Let P_t be the marginal distribution of $X(t)$, this is $P_t(x) = \mathbb{P}(X(t) \leq x)$. The integrated functional depth of X with respect to P is

$$\mathrm{IFD}(X, P) = \int_0^1 D(X(t), P_t) w(t) dt, \tag{1}$$

w being a weight function that integrates to one.

Let $X_1, \ldots, X_n$ be n independent copies of X. As we previously commented, we consider that the realizations $X_1, \ldots, X_n$ are partially observed. Following Delaigle and Hall (2013), we formalize this sampling framework by introducing a random censoring process with distribution Q that generates compact subsets of $[0, 1]$. From now on, $V_1, \ldots, V_n$ are independent copies of $V \sim Q$ and X_i is only observed on V_i. The relationship between P and Q plays here an important role and, as the majority of the literature, we assume MCAR. This is, $(X_1, V_1), \ldots, (X_n, V_n)$ are i.i.d from $P \times Q$.

Now, we define the censored functional depth as an IFD restricted to the compact set V where X is observed.

Definition 2. The *censored functional depth* of (X, V) with respect to $P \times Q$ is defined by

$$\mathrm{CFD}((X,V), P \times Q) = \int_V D(X(t), P_t) w(t|V) dt, \tag{2}$$

$w(t|V)$ being the weight function defined by

$$w(t|V) = \frac{Q(t)}{\int_V Q(t)dt}, \tag{3}$$

with $Q(t) = \mathbb{P}(t \in V)$.

Now denote by P_n the distribution that assigns mass $1/n$ to each sample curve $X_1, \ldots, X_n$. Similarly, let Q_n be the distribution that assigns mass $1/n$ to each sample compact set $V_1, \ldots, V_n$. Thus, $\mathrm{CFD}((X,V), P_n \times Q_n)$ is the plug-in estimator of $\mathrm{CFD}((X,V), P \times Q)$. When the weight function $w(t|V)$ is defined as in (3), the plug-in estimator may be written as

$$\mathrm{CFD}((X,V), P_n \times Q_n) = \int_V D(X(t), P_{t,n}) q_n(t) dt \Big/ \int_V q_n(t) dt, \tag{4}$$

$P_{t,n}$ being the empirical distribution function of univariate sample $\{X_i(t), i \in I(t)\}$ and

$$q_n(t) = \#\{1 \leq i \leq n : t \in V_i\}/n. \tag{5}$$

As it occurs in practise, suppose now that the time points on which the curves may be observed are $t_1 < \cdots < t_T$. So, V and V_i, $1 \leq i \leq n$, are all subsets of $\{t_1, \ldots, t_T\}$. For simplicity and without loss of generality, we will assume that $0 = t_0 < t_1 < \cdots < t_T < t_{T+1} = 1$ are equidistant. Then, we consider the sample version of the depth, by using a standard Riemann approximation. This is,

$$\mathrm{CFD}_T((X,V), P_n \times Q_n) = \sum_{t \in V} D(X(t), P_{t,n}) \frac{q_n(t)}{\sum_{t \in V} q_n(t)}, \tag{6}$$

3 Simulation study

The aim of this section is to study sample properties of our proposal when considering different population features for P and Q. Hence, we compare $\mathrm{CFD}_T((X_i, V_i), (P_n \times Q_n))$ values versus the corresponding discretization of IFD, namely

$$\mathrm{IFD}_T(X_i, P_n) = \sum_{k=1}^{T} D\big(X_i(t_k), P_{t_k,n}\big) \frac{q_n(t_k)}{\sum_{j=1}^{T} q_n(t_j)}, \tag{7}$$

For generating samples $(X_1, V_1), \ldots, (X_n, V_n)$ from $P \times Q$, first we generated Gaussian processes with periodic mean function and exponential covariance function. A detailed description of theses processes is in the Appendix. Second, we considered the following two censoring processes:

Model A: We censored data on m intervals uniformly spread on the unit interval, controlling the total proportion p of censored data. For this, for each trajectory X_i, we generated a random sample of size $(m-p)/p$ from an uniform distribution. Then, we considered the intervals $(u_{(i-1)}, u_{(i)})$, $u_{(i)}$ being the i-th order statistics of the sample, and censored data belong to m intervals chosen at random.

Model B: We censored data at the extremes of the observation domain. We simulated this situation also controlling the total proportion p. For this, we censored data on $[0, a) \cup (b, 1]$, being $a \sim U[0, p/2]$ and $b \sim Unif[1-p/2, 1]$. This setting mimics the censoring that arises when a functional sample is aligned by affine methods (Sangalli et al., 2014a; Marron et al., 2015).

Then, we computed the Willmott index (Duveiller et al., 2016; Willmott, 1981) between the uncensured and censured depths. Given two vectors of depths $\tilde{Y}$ and Y, this statistic provides a number between 0 and 1, being one a perfect agreement between $\tilde{Y}$ and the reference Y. Notice that, Pearson correlation would indicate a strong agreement for any two vectors with a linear relationship, however, Willmott only would provide an agreement of 1 to relationships in the 45° line.

We generated 100 samples of size 100 and considered different IFDs. They were, the Fraiman and Muniz Depth (FM) (Fraiman and Muniz, 2001), the Modified Band Depth (MBD) (López-Pintado and Romo, 2009) and the Modified Half Region Depth (MHRD) (López-Pintado and Romo, 2011). In addition, we considered the Modified Epigraph Index (MEPI) (López-Pintado and Romo, 2011) because, in conjunction with MBD, the Outliergram is built (Arribas-Gil and Romo, 2014), an outlier detection tool. Table 1 shows mean values of the Willmott index for different settings. Notably, the means are always close to 1 indicating a high agreement even for high levels of censoring (up to fifty percent missing). In addition, from this simulation study, we could say that a high number of missing intervals is preferable than just one for a given level of censoring. Figure 1 shows

			Mean Willmott		
		m	10%	25%	50%
FM	Model A	1	0.9969	0.9804	0.9219
		2	0.9984	0.9891	0.9496
		4	0.9992	0.9938	0.9676
	Model B	2	0.9974	0.9878	0.9568
MBD	Model A	1	0.9973	0.9826	0.9287
		2	0.9986	0.9907	0.954
		4	0.9993	0.9946	0.97
	Model B	2	0.9978	0.9896	0.9615
MEPI	Model A	1	0.9964	0.9771	0.911
		2	0.9982	0.9874	0.9421
		4	0.9991	0.9928	0.9642
	Model B	2	0.9969	0.986	0.9526
MHRD	Model A	1	0.9989	0.9927	0.9701
		2	0.9995	0.9962	0.9804
		4	0.9997	0.9979	0.989
	Model B	2	0.9991	0.996	0.9848

Table 1: Mean values of Willmott index based on one hundred replicates of models A and B and different levels of censoring and number of intervals.

the scatter plots between both depths, revealing a symmetry along the 45° line. Notice that some particular features of the scatter plots are inherit to the used depth. For example, MBD is more concentrated around high values of depth (0.4) meanwhile MHRD assigns values in a more uniform way. In contrast, the censoring has an important effect on the dispersion as can be noticed by the way the points are more spread as the censoring proportion is increased. Furthermore, the dispersion is not uniform along the line. This is due to Gaussian processes are more dense along the center. Hence, a missing interval is more determinant for a function in the center than for an observation outside the mass.

In contrast with other simulation settings, the censoring process above is applied to

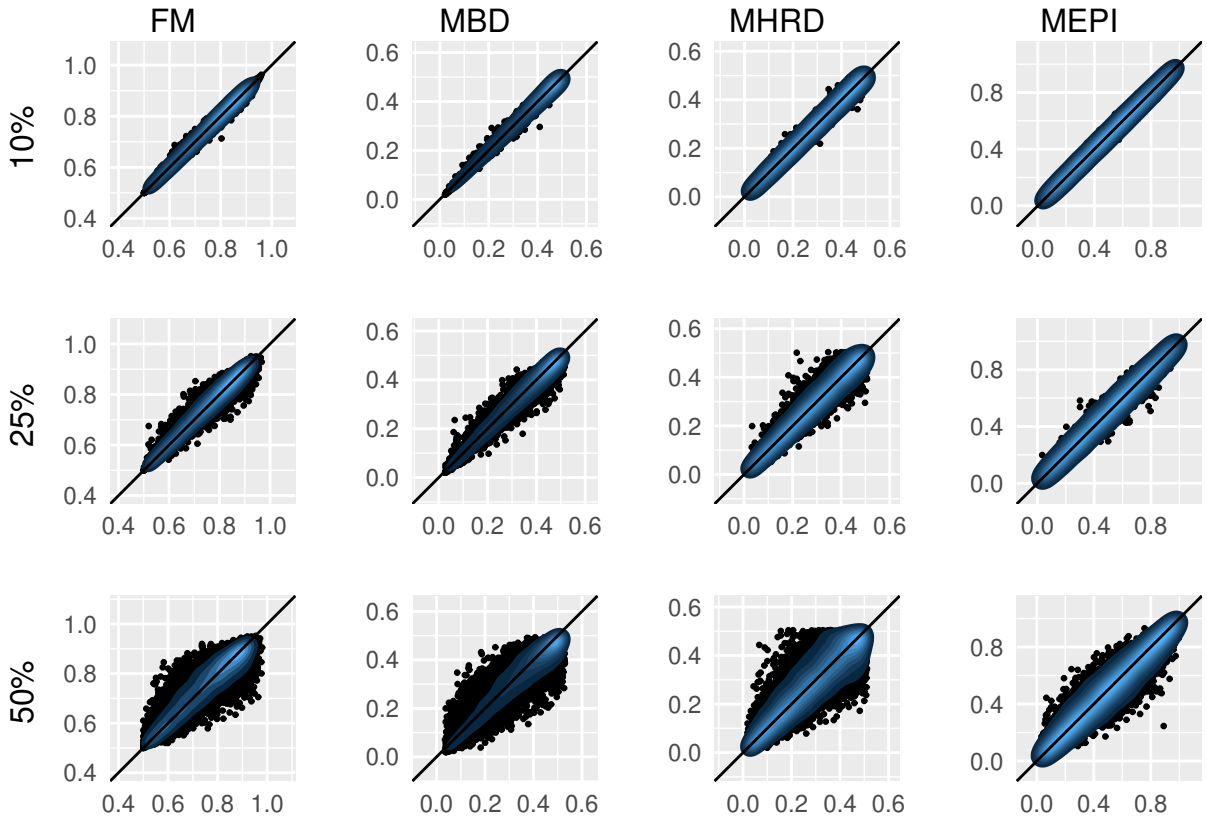

Figure 1: Cloud of (CFD_T, IFD_T)-points, based on FM, MBD, MHRD and MEPI, over one hundred simulations of size $n = 100$. Joint density estimators are represented in a blues scale, from light blues (highest density values) to dark ones (lowest values) . The censoring processes correspond to Model A with $m = 2$ missing intervals and $p = 0.10, 0.25, 50\%$ levels of censoring.

any single observation. Other articles, Kraus (2015) and Kneip and Liebl (2017) censor a function with a probability smaller than 1; in this setting our proposal would achieve better performance due to the inclusion of completely observed functions to the sample.

4 Conclusions

We introduce a censored functional depth for partially observed functional data. Our simulations show that the empirical version of the censored depth nearly matches the corresponding empirical version of the uncensored depth for finite samples even for large proportions of missing data. We believe that our approach can be an articulation to spread the application of existing depth-based tools for partially observed functional data, a situation more a more frequent today.

Acknowledgement

Antonio Elías is supported by the Spanish Ministerio de Educación, Cultura y Deporte under grant FPU15/00625. Antonio Elías and Raúl Jiménez are partially supported by the Spanish Ministerio de Economía y Competitividad under grant ECO2015-66593-P.

Appendix

The simulated functional sample follows the model

$$X(t) = \mu(t) + \epsilon(t), \qquad t \in [0, 1],$$

where $\mu(t) = \sin(2\pi t)$ and $\epsilon(t)$ is a zero mean Gaussian process with covariance function

$$\mathbb{E}[\epsilon(s)\epsilon(t)] = \alpha e^{-\beta|s-t|}, \quad s, t \in [0, 1].$$

In particular, we set $\alpha = 0.3$ and $\beta = 0.5$. Figure 2 shows a generated sample.

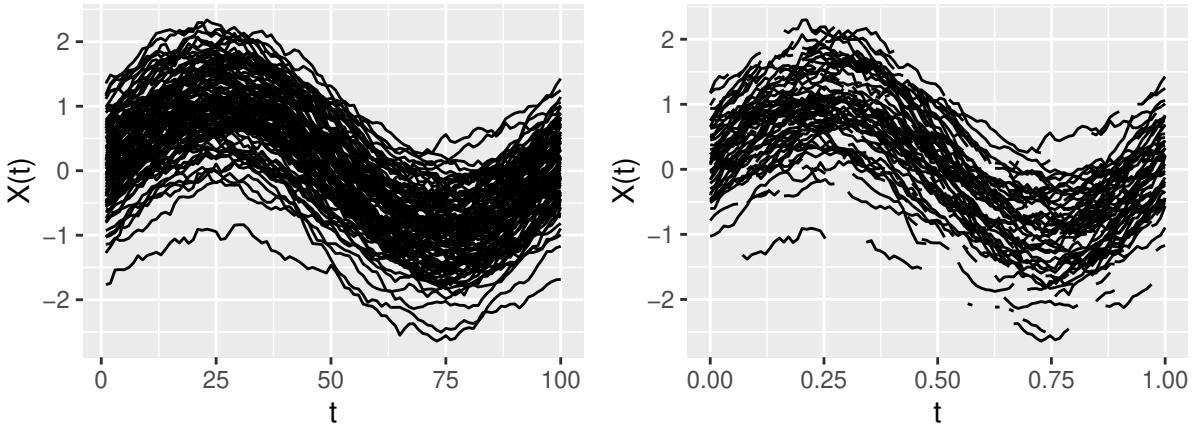

Figure 2: Left panel: 100 Gaussian processes with sinusoidal mean and exponential covariance function. Right panel: the same sample but censored with the Model A, $m = 10$ and $p = 0.5$.

References

Arribas-Gil, A. and Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619.

Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109(505):411–423.

D'Amato, V., Piscopo, G., and Russolillo, M. (2011). The mortality of the italian population: Smoothing techniques on the lee-carter model. *Ann. Appl. Stat.*, 5(2A):705–724.

Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association*, 108(504):1269–1283.

Delaigle, A. and Hall, P. (2016). Approximating fragmented functional data by segments of markov chains. *biomet*, 103(4):779–799.

Duveiller, G., Fasbender, D., and Meroni, M. (2016). Revisiting the concept of a symmetric index of agreement for continuous datasets. *Scientific Reports*, 6:19401.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

Goldberg, Y., Ritov, Y., and Mandelbaum, A. (2014). Predicting the continuation of a function with applications to call center data. *Journal of Statistical Planning and Inference*, 147:53–65.

James, G., Hastie, T., and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550.

Kneip, A. and Liebl, D. (2017). On the optimal reconstruction of partially observed functional data. *arXiv e-prints*, page arXiv:1710.10099.

Kraus, D. (2015). Components and completion of partially observed functional data. *J. R. Stat. Soc. B*, 77(4):777–801.

Liebl, D. and Rameseder, S. (2019). Partially observed functional data: The case of systematically missing parts. *Computational Statistics & Data Analysis*, 131:104 – 115. High-dimensional and functional data analysis.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by liu and singh). *Ann. Statist.*, 27(3):783–858.

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

López-Pintado, S. and Romo, J. (2011). A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55(4):1679–1695.

Marron, J. S., Ramsay, J. O., Sangalli, L. M., and Srivastava, A. (2015). Functional data analysis of amplitude and phase variation. *Statist. Sci.*, 30(4):468–484.

Nagy, S., Gijbels, I., Omelka, M., and Hlubinka, D. (2016). Integrated depth for functional data: statistical properties and consistency. *ESAIM. Probability and Statistics*, 20.

Sangalli, L. M., Secchi, P., and Vantini, S. (2014a). Analysis of aneurisk65 data: k-mean alignment. *Electron. J. Statist.*, 8(2):1891–1904.

Sangalli, L. M., Secchi, P., and Vantini, S. (2014b). Aneurisk65: A dataset of three-dimensional cerebral vascular geometries. *Electron. J. Statist.*, 8(2):1879–1890.

Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2009). A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, 104(485):37–48.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233.

University of California, B. U. and for Demographic Research (Germany), M. P. I. (2019). Human mortality database. Available at www.mortality.org.

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, 2(2):184–194.

Yao, F., Muller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.