Journal of Multivariate Analysis 146 (2016) 46–62

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Weak convergence of discretely observed functional data with applications

Stanislav Nagy [a,b,c,*], Irène Gijbels [a,b], Daniel Hlubinka [c]

[a] KU Leuven, Department of Mathematics, Statistics Section, Celestijnenlaan 200b - box 2400, 3001 Leuven, Belgium

[b] KU Leuven, Leuven Statistics Research Centre (LStat), Celestijnenlaan 200b - box 5307, 3001 Leuven, Belgium

[c] Charles University in Prague, Department of Probability and Math. Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic

ARTICLE INFO

Article history:
Received 17 December 2014
Available online 16 June 2015

AMS 2000 subject classifications:
62G20
62G05
62H12

Keywords:
Consistency
Data depth
Functional data
Functional moments
Weak convergence

ABSTRACT

A general result on weak convergence of the empirical measure of discretely observed functional data is shown. It is applied to the problem of estimation of functional mean value, and the problem of consistency of various types of depth for functional data. Counterexamples illustrating the fact that the assumptions as stated cannot be dropped easily are given.

1. Introduction: complete and discrete design of functional data

By functional data we usually understand an outcome of an experiment that can be represented as a set of continuous functions, whose domain is a common compact interval. The domain can be referred to as time. The value of the outcoming function can then be interpreted as the value of the measured characteristic, evolving within the given time frame. Depending on the nature of the experiment, two distinct approaches towards the observation design of functional data can be found in the literature.

The first, more traditional approach assumes that all the functions involved are observed completely in time. In other words, the functional values of all the functions resulting from the experiment are known at each time point of their domain. We call this setup **complete design**, for brevity. Under this complete design assumption there exists a wide variety of statistical procedures enabling the analysis of functional data. For exposition see, for example, [25,26,10,15].

The second, perhaps more realistic, approach to the observation design of functions arises when assuming that each random function is observed only at a finite grid of time points. These points can be either deterministic and preset by the experimenter, or occurring at random. Here, we focus on nonparametric statistical analysis of such partially observed data.

Typically, to make valid nonparametric inference exploiting the functional nature of data, it should be assumed that the number of points at which random functions are observed tends to be larger when the sample size increases, and that the

* Corresponding author at: KU Leuven, Department of Mathematics, Statistics Section, Celestijnenlaan 200b - box 2400, 3001 Leuven, Belgium.
E-mail addresses: stanislav.nagy@wis.kuleuven.be (S. Nagy), irene.gijbels@wis.kuleuven.be (I. Gijbels), hlubinka@karlin.mff.cuni.cz (D. Hlubinka).

http://dx.doi.org/10.1016/j.jmva.2015.06.006
0047-259X/

largest span between two adjacent points vanishes as the sampling process continues to infinity. This setup will be called **discrete design**.

The majority of papers in functional data analysis consider the data as being observed at each point of the domain. See [7, for example Section 4.2] for some discussion on the nature and treatment of functional data. In a variable selection problem Aneiros and Vieu [2] consider the functions to be observed through a fine grid, and look into asymptotics when the distances between the grid points diminish to zero. Other recent contributions taking data as discretized can be found in a book edited by Bongiorno et al. [5].

To accomplish statistical analysis for functions observed within the discrete design setting, it is necessary to perform an initial step of representing the discretely observed functions by elements of the space of complete functions. This is usually done by approximating, or interpolating, the discretely observed values of functions. After doing this, the resulting approximating (complete) functions can be utilized as if they were the original, continuously unobservable, set of curves.

Such preprocessing of functional data is customarily considered to be imperative when encountering a discretely observed functional data set [25]. Though often discussed, few theoretical results investigate the effect of this fundamental data imputation on the statistics involved.

The essential contribution of the present paper lies in a theoretical result facilitating the understanding of this phenomenon. In Section 2, we propose a natural and straightforward method of approximating functions from a random sample whose values are observed only at a finite number of points in the domain. Within the discrete design of observations, we show that the probability distribution of these approximations converges weakly towards the original sampling distribution. This enables us to state a Varadarajan type of result [28] for such discretely observed functional data dealing with the weak convergence of empirical measures based on these approximations.

The second part of the paper concerns two applications of the main result. In Section 3 we apply it to the problem of estimation of mean of functional data. We show that by an average based solely on a random sample of discretely observed curves, it is possible to estimate the mean value of a probability distribution in a functional space in an asymptotically unbiased and consistent manner. Finally, in Section 4, the developed theory is applied to a nonparametric tool suitable for functional data—data depth (cf. [30]). There, conditions under which consistency is preserved for functional data depth when functions are observed discretely are explored. Both application sections are completed with a number of examples. These aim to illustrate that the conditions of the theoretical results cannot be dropped generally, and depict pitfalls to keep in mind when replacing a set of discretely observed functional data by complete curves.

The proofs of the theoretical results are provided in a Supplementary material (see Appendix A) part accompanying this paper. That part also contains an additional example providing information on the necessity of the conditions.

2. Weak convergence of discretely observed functions

In this section, we state a rather technical, but useful, theorem concerning general weak convergence in the space of continuous functions on a compact interval. We start by introducing the notation.

For $K \in \mathbb{N} = \{1, 2, \ldots\}$ let $\mathcal{C}^K([0,1])$ be the Banach space of continuous $\mathbb{R}^K$-valued functions on $[0,1]$

$$\mathcal{C}^K([0,1]) = \left\{x\colon [0,1] \to \mathbb{R}^K\colon \ x \text{ is continuous on } [0,1]\right\}$$

equipped with the uniform norm $\sup_{t\in[0,1]} \|x(t)\|$. Here, $\|\cdot\|$ denotes the Euclidean norm on the space $\mathbb{R}^K$.

For an arbitrary metric space $\mathcal{M}$ with the σ-algebra of its Borel sets, $\mathcal{P}(\mathcal{M})$ stands for the collection of all probability measures defined on $\mathcal{M}$.

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space on which all random variables will be defined. For $\mathrm{P} \in \mathcal{P}\left(\mathcal{C}^K([0,1])\right)$, let $\{X_n\}_{n=1}^{\infty} \subset \mathcal{C}^K([0,1])$ denote an infinite sequence of independent random functions distributed as P. For a fixed random element $\omega \in \Omega$ we denote the empirical measure defined by the first n functions from this sequence by $\mathrm{P}_n(\omega) \in \mathcal{P}\left(\mathcal{C}^K([0,1])\right)$. To put it precisely, if δ_x is a Dirac measure on $\mathcal{C}^K([0,1])$ concentrated at $x \in \mathcal{C}^K([0,1])$, then

$$\mathrm{P}_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)} \quad \text{for } \omega \in \Omega. \tag{1}$$

If no confusion about the underlying random element can arise, the argument ω will be dropped.

For a fixed time point $t \in [0,1]$ and $X \sim \mathrm{P}$ we use $\mathrm{P}_t \in \mathcal{P}\left(\mathbb{R}^K\right)$ to denote the marginal distribution of $X(t)$. Likewise, $\mathrm{P}_{n,t}$ stands for the marginal distribution of P_n defined in (1) at t.

The symbol $\mathbb{I}[A]$ stands for the indicator function of A, i.e. equals 1 if A holds true, and 0 if not. For a set S and a sequence of (possibly random) positive integers $\{m_n\}_{n=1}^{\infty} \subset \mathbb{N}$, a triangular array of elements of S is an arbitrary doubly indexed sequence of elements of the set S

$$\left\{s_{j,n}\right\}_{j=1}^{m_n} = \left\{s_{j,n} \in S : j = 1, \ldots, m_n \text{ and } n \in \mathbb{N}\right\}.$$

In the present section, all random functions are understood as observed within the discrete design. To put this rigorously, let $\left\{T_{j,n}\right\}_{j=1}^{m_n} \subset [0,1]$ be an arbitrary triangular array of points in $[0,1]$. This array is referred to as the array of observation