Computational Statistics and Data Analysis 105 (2017) 24–45

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Depth-based nonparametric description of functional data, with emphasis on use of spatial depth

Robert Serfling[a,*], Uditha Wijesuriya[b]

[a] *Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, USA*

[b] *Department of Mathematics, University of Southern Indiana, Evansville, IN 47712, USA*

HIGHLIGHTS

- Methods for displaying sample quantile curves for a functional data set.
- Methods for computing and displaying confidence bands for population quantile curves for a functional data model.
- Formulation of double-fence functional boxplot to better identify shape outliers.
- Validation of the spatial depth approach for nonparametric description of functional data.

ARTICLE INFO

Article history:
Received 2 October 2015
Received in revised form 11 July 2016
Accepted 13 July 2016
Available online 25 July 2016

Keywords:
Functional data
Nonparametric
Quantiles
Outlier detection
Boxplot

ABSTRACT

Statistical depth and related quantile functions, originally introduced for nonparametric description and analysis of multivariate data in a way sensitive to inherent geometry, are in active development for functional data and in this setting offer special options since the data may be visualized regardless of dimension. This paper provides depth-based methods for revealing the structure of a functional data set in terms of relevant sample quantile curves displayed at selected levels, and for constructing and displaying confidence bands for corresponding "population" versions. Also, the usual functional boxplot is enhanced, by adding inner fences to flag possible shape outliers, along with the outer fences that flag location outliers. This enables the boxplot to serve as a stand-alone tool for functional data, as with univariate and multivariate data. Further, the spatial depth approach, well-established for multivariate data, is investigated for nonparametric description of functional data along these lines. In comparison with four other commonly used depth approaches for functional data, over a range of actual and simulated data sets, the spatial depth approach is seen to offer a very competitive combination of robustness, efficiency, computational ease, simplicity, and versatility.

1. Introduction and preliminaries

Functional data, where each point is a curve, arises in all areas of modern science, engineering, and industry. Nonparametric description of such data sets, in terms of a median curve, quantile curves at various levels, the middle 50% region of curves, and identification of outliers, is fundamental to understanding structure. A natural approach is provided by depth functions, which order data points according to some given notion of centrality or, equivalently, of outlyingness, and

[*] Corresponding author.
E-mail address: serfling@utdallas.edu (R. Serfling).
URL: http://www.utdallas.edu/~serfling (R. Serfling).

http://dx.doi.org/10.1016/j.csda.2016.07.007
0167-9473/© 2016 Elsevier B.V. All rights reserved.

which generate related quantile functions. With functional data, which may be visualized regardless of dimension, depth methods offer special options beyond their standard role with multivariate data.

Here we present three depth-based methodological tools for nonparametric description of functional data. The first concerns definition and display of sample quantile curves at arbitrary levels. The second concerns construction of confidence bands for the corresponding "population" quantile curves being estimated by the sample versions, under the technical condition that asymptotic normality holds for the discretized sample quantile curves based on the depth function in use. The third concerns nonparametric identification of outliers, for which we introduce a "double-fence functional boxplot", enhancing the Sun and Genton (2011) functional boxplot by incorporating additional fences in order to better flag "shape" outliers, which are not envisioned in the classical boxplot design oriented only to location outliers.

Further, we focus on a particular depth function – the spatial depth – and establish its effectiveness and special appeal in the functional data setting. Despite its success with multivariate data, this approach has received but limited treatment with functional data, even though Chaudhuri (1996) indicated its potential for natural extension from Euclidean spaces to infinite-dimensional Hilbert and Banach spaces. One attractive feature is that it satisfies an asymptotic normality condition required for our confidence band procedure. Also, for outlier identification using the double-fence functional boxplot, we compare the spatial depth approach with four leading depth-based approaches for functional data, over a range of actual and simulated data sets and a variety of outlier types, and find that the spatial approach exhibits superior overall performance.

Our setting consists of a data set of curves $X_k(t)$, $t \in [a, b]$, for $1 \leq k \leq n$, defined over a finite interval $[a, b]$ as elements of a Hilbert space $\mathcal{H}$. (Confining to an interval $[a, b]$ is arbitrary, for convenience.) We can work entirely within $\mathcal{H}$, but it is technically convenient and computationally efficient to make use also of associated discretized versions of the curves as vectors in $\mathbb{R}^d$ for some suitable d. Indeed, in many typical applications, the curves are actually secondary constructions from data that has arisen initially in discretized form. In any case, for $1 \leq k \leq n$, we associate with the curve $X_k(\cdot)$ the vector $\boldsymbol{X}_k = (X_k(t_1), \ldots, X_k(t_d))'$ in $\mathbb{R}^d$, corresponding to discretization using d equally spaced points

$$a \leq t_1 < t_2 < \cdots < t_d \leq b. \tag{1}$$

Thus we effectively use $\mathbb{X}_n = \{\boldsymbol{X}_1, \ldots \boldsymbol{X}_n\}$ as our basic data set for analytical work, while for plots we use the curves $X_k(\cdot)$, $1 \leq k \leq n$. Here no restriction such as $d \leq n$ or $d \geq n$ is imposed. Of course, in cases where n is very large, the displays of the data curves might be carried out just for a subsample, in order not to obscure visualization of the typical structure of the curves. Even so, the analytical work defining quantile curves and confidence bands to add to such plots is based on the full sample or at least a large subsample. Also, for technical reasons, the initial computation of the confidence band procedure is based on discretization to $d_0 = \min\{d, 20, n - 1\}$ as explained in Section 3.1.3.

Remark 1 (*On Discretization of Functional Data Curves*)**.** In discretizing into $\mathbb{R}^d$ to enable use of standard multivariate analysis techniques, it is helpful to keep in mind five important differences between the usual multivariate data setting and that of discretized functional data.

(i) Besides "location" outliers lying apart from the main body of data, a functional data set can also have "shape" outliers differing in structure or pattern from the typical cases, as well as outliers deviating in both location and shape. Shape outliers embedded within the main body of curves make outlier detection considerably more challenging.

(ii) In the functional data setting, full affine invariance of outlier detection procedures is neither required nor desired. In particular, invariance under heterogeneous scale changes of $X(t)$, $t \in [a, b]$, is irrelevant, it sufficing to have invariance merely under homogeneous scale changes and under orthogonal transformations of discretized versions. Similar comments apply regarding equivariance of the median and other quantile curves. Thus, fortuitously, invariance and equivariance requirements are less stringent in the functional data setting.

(iii) The component variables $X(t)$ defining a curve are ordered in the functional data setting, and this carries over to the components of their associated discretized versions. Therefore, in the functional data setting, we have the very attractive feature that each data point, discretized or not, may be represented and viewed as a curve in a 2-dimensional plot, $X(t)$ versus t, regardless of the inherent dimensionality of the data.

(iv) Further, since such plots naturally have an "upper" region and a "lower" region, the selection of representative quantile curves at various levels may accordingly be confined to "upper" and "lower" types relative to the location of the median curve. In comparison, in the standard multivariate data setting, the quantiles at a given level form a contour that is associated with all directions from the "center", and there is no simplifying "upper" versus "lower" orientation. (See Sections 3.1.1 and 3.1.2.)

(v) Finally, moreover, the directions that index the quantile representations in $\mathbb{R}^d$ of the discretized data curves also have ordered components and, therefore, like the data curves themselves, may be represented and viewed as curves in a 2-dimensional plot, regardless of the dimensionality of the discretization (see Fig. 1). □

Formally, the following objectives typify nonparametric description of functional data:

1. Identification of the sample median curve and selected sample quantile curves, for example a representative "90th percentile curve";
2. Identification and designation of "outlier" curves, both location and shape types;
3. Display of the data as a 2-dimensional functional boxplot exhibiting the "middle" 50% region of sample curves and "fences" demarking outlier regions;