Computational Statistics and Data Analysis 125 (2018) 10–26

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust template estimation for functional data with phase variability using band depth

Jason Cleveland, Weilong Zhao, Wei Wu *

Department of Statistics, Florida State University, 117 N Woodward Ave, Tallahassee, FL, 32306-4330, USA

ARTICLE INFO

Article history:
Received 3 July 2017
Received in revised form 5 March 2018
Accepted 10 March 2018
Available online 29 March 2018

Keywords:
Functional data analysis
Template estimation
Registration
Time warping
Band depth

ABSTRACT

Registration, or alignment, of functional observations has been a fundamental problem in functional data analysis. The creation of a template was the key step for alignment of a group of functions. Recent studies have defined the template with the notion of "mean" in the given observations. However, the mean can be sensitive to the, commonly observed, outlier functions in the data. To deal with this problem, a new approach is proposed to adopt the notion of "median" using the time warping functions in the alignment process, based on the recently developed band depth in functional data. A semi-parametric model is provided with an algorithm that yields a consistent estimator for the underlying median template. The robustness of this depth-based registration is illustrated using simulations and two real data sets. In addition, a new depth-based boxplot is proposed for outlier detection in functional data with phase variability.

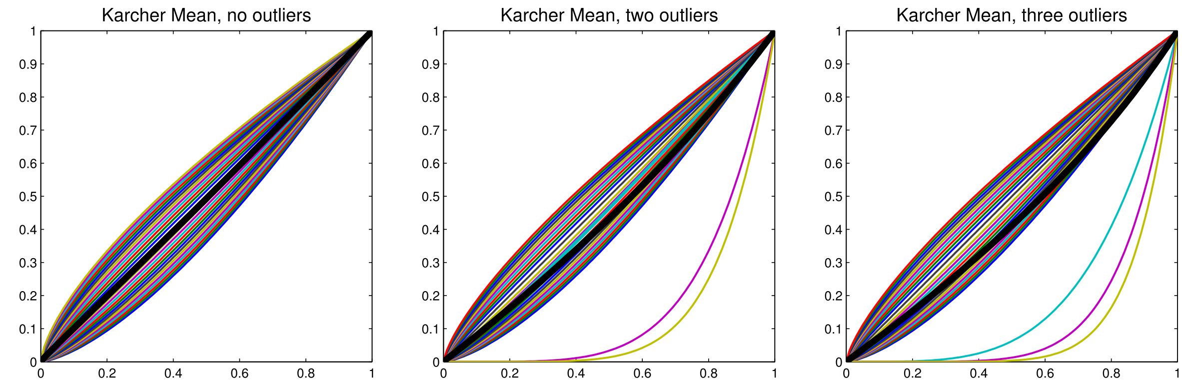
1. Introduction

The field of functional data analysis (FDA) has made great strides in mathematically modeling time-dependent observations. Progress towards handling temporal or phase variability along the time axis has addressed some of the issues once facing conventional analysis methods, such as spline smoothing, covariance or correlation analysis, functional principal component analysis, and functional linear or nonlinear regressions (Kneip and Gasser, 1992; Gasser and Kneip, 1995; Wang and Gasser, 1997; Ramsay and Li, 1998). The general solution to the phase variability, or time warpings, has been to apply a computational approach that removes such warpings prior to analysis of the data (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). Various approaches over the past two decades depend on finding a way to perform an optimal matching process (a procedure meant to match important functional features in the observations) called registration or curve alignment (Liu and Muller, 2004; Tang and Muller, 2008; Gervini and Gasser, 2004; Kneip and Ramsay, 2008; James, 2007; Capra and Muller, 1997; Ramsay and Silverman, 2005; Kneip and Gasser, 1992; Ramsay and Li, 1998; Hall et al., 2007; Srivastava et al., 2011). The basic procedure to register a set of functional observations is to find a template and then align each function to the template. The template is often defined as the *mean* or *average* of the data under a given metric. However, it is well known that the mean can be sensitive to outliers, which commonly appear in practical observations.

For example, the Karcher mean template with the elastic metric in the Fisher–Rao registration framework (Srivastava et al., 2011) is sensitive to outliers thus making it a poor choice for estimation. We can see in Fig. 1 the Karcher mean of a set of time warping functions can significantly change with respect to a small number of outliers in the data. In that regard, we will need a robust signal estimation with respect to potential outliers. The median is a naturally desired candidate template for our signal estimation. The challenging question is, "How do we find the median of functional observations"?

* Corresponding author.
E-mail addresses: jcleveland@stat.fsu.edu (J. Cleveland), weilong.zhao@stat.fsu.edu (W. Zhao), wwu@stat.fsu.edu (W. Wu).

https://doi.org/10.1016/j.csda.2018.03.009
0167-9473/



(A) Zero outliers. (B) Two outliers. (C) Three outliers.

Fig. 1. Karcher mean with respect to outliers. **A.** Observed time warping functions with no outliers present. Each line denotes one function and the bold black line denotes the Karcher mean. **B.** Same as Panel A except that two outliers are added. **C.** Same as Panel A except that three outliers are added.

This question has been well addressed with the notion of depth on functional data over the past decade. The methods include h-depth (Cuevas et al., 2007), random Tukey depth (Cuesta-Albertos and Nieto-Reyes, 2008), band depth, modified band depth (Lopez-Pintado and Romo, 2009), half-region depth, modified half-region depth (López-Pintado and Romo, 2011), integrated depth (Fraiman and Muniz, 2001), multivariate functional halfspace depth (Claeskens et al., 2014), functional spatial depth (Chakraborty et al., 2014), etc. Of these methods, the band depth by Lopez-Pintado and Romo in Lopez-Pintado and Romo (2009) is probably the most commonly used. This depth is analogous to the well-known simplicial depth (Liu, 1990) on multivariate data and has plenty of computational and theoretical advantages. We will adopt this depth to estimate the template for functional observations with phase variability. Note that a robust method on functional observations has been studied with a Karcher median method (Kurtek et al., 2013), an elastic-metric-based measure of centrality for functional data. We will compare the proposed band depth approach with the Karcher median method in this manuscript.

In theory, functional observations with phase variability have been stated in the following general framework:

$$f_i(t) = (X_i \circ \gamma_i)(t) + e_i(t), \quad t \in [0, 1], \quad i = 1, \ldots, n, \tag{1}$$

where $\{f_i(t)\}$ are observations of $X_i(t)$ with compositional noise $\{\gamma_i(t)\}$ and additive noise $\{e_i(t)\}$. Various approaches have been proposed to address this problem (Tang and Muller, 2008; Gervini and Gasser, 2004; James, 2007; Kneip and Gasser, 1992; Ramsay and Li, 1998). In particular, if $X_i(t) = g(t)$ (independent of index i), then the model is in a template-based form.

In this paper, we focus on observations with nonlinear phase variability and investigate the robust estimation in the following semi-parametric, template-based framework (Kurtek et al., 2011):

$$f_i(t) = c_i(g \circ \gamma_i)(t) + e_i, \quad t \in [0, 1], \quad i = 1, \ldots, n, \tag{2}$$

where $g(t)$ is the underlying template, $\{f_i(t)\}$ are observations of $g(t)$ with random time warpings $\{\gamma_i(t)\}$, scaling coefficients $\{c_i\} \in \mathbb{R}^+$, and vertical translations $\{e_i\} \in \mathbb{R}$. This model has scale and translation variability on amplitude. We will provide an algorithm to estimate g from $\{f_i\}$ with the generative model in Eq. (2) and prove the estimator is asymptotically consistent.

We will also extend the notion of boxplot to time warping functions. Recent work has introduced a boxplot definition for functional observations (Sun and Genton, 2011). However, as the time warping functions are constrained (see the method section for details), the so-defined boxplot cannot capture the degree of variability over the time domain or properly identify outliers in the data. We propose a coordinate-rotation-based method to address this issue. We will demonstrate the important mathematical properties in the new method and compare it with the previous functional boxplot.

The rest of the paper is organized as follows: In Section 2, we provide details of our framework on phase band depth median, compare it with the Karcher median, and then propose a robust estimation algorithm with consistency theory supported. The effectiveness of this new framework is demonstrated using simulations and real data sets in Section 3. In Section 4, we extend the depth-based framework to a definition of boxplot for time warping functions. Section 5 provides discussion and conclusion. Finally, the Appendices give all mathematical details throughout the paper.

2. Methods

In this section we will propose a new framework for estimating the functional template by using the notion of depth in phase variability. We will, at first, briefly review the background information on phase variability in functional data.