Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A Bayesian method for simultaneous registration and clustering of functional observations

Zizhen Wu, David B. Hitchcock*

Department of Statistics, University of South Carolina, Columbia, SC 29208, United States

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form 17 February 2016
Accepted 18 February 2016
Available online xxxx

Keywords:
Functional data
Time warping
Curve registration

ABSTRACT

We develop a Bayesian method that simultaneously registers and clusters functional data of interest. Unlike other existing methods, which often assume a simple translation in the time domain, our method uses a discrete approximation generated from the family of Dirichlet distributions to allow warping functions of great flexibility. Under this Bayesian framework, a MCMC algorithm is proposed for posterior sampling. We demonstrate this method via simulation studies and applications to growth curve data and cell cycle regulated yeast genes.

1. Introduction

An important example of exploratory data analysis, cluster analysis involves grouping observations that share similar characteristics. In many clustering methods, similarities or dissimilarities between pairs of observations are measured by some relevant distance metric. Methods based on dissimilarity measures include hierarchical clustering and the K-medoids method (Everitt et al., 2011). Another major clustering technique, model-based clustering, requires statistical assumptions about the observations. A popular model-based method (Fraley and Raftery, 2002) assumes a multivariate normal distribution for the measurements and assigns objects to clusters by comparing the posterior group probabilities given the observations.

If the observations to be clustered are functional data, i.e., repeated measures over time or some other domain, one may consider fitting a function to each observational unit using some basis function expansion (Ramsay and Silverman, 2005). Throughout this paper, we use the B-spline basis (De Boor, 2001). One advantage of functional data analysis over traditional multivariate analysis is the ability to examine higher-order derivatives of fitted functions. For example, the first-order derivative of a fitted monotone smoothing function (Ramsay and Silverman, 2005) measuring children's height over a given period represents the estimated growth velocity, and the second-order derivative is the estimated growth acceleration, etc.

Recently, several methods have been developed for clustering functional data. Luan and Li (2003) use mixed-effect models for time-course gene expression and cluster the curves by calculating their posterior cluster probabilities via the EM algorithm. This mixed-effect model is a special case of the model proposed by James and Sugar (2003).

A more challenging yet often-encountered problem is clustering the observations in the presence of time distortions, also known as phase variation. The time distortion is usually modeled by a warping function $h(\cdot)$ (Ramsay and Silverman, 2005), which is a non-decreasing continuous function defined on the time domain $\mathcal{T}$ satisfying the endpoint conditions $h(a) = a$, and $h(b) = b$, where a and b are two endpoints of the time domain. Fig. 1 shows eight warping functions, while

* Corresponding author. Tel.: +1 803 777 5346; fax: +1 803 777 4048.
E-mail address: hitchcock@stat.sc.edu (D.B. Hitchcock).

http://dx.doi.org/10.1016/j.csda.2016.02.010
0167-9473/© 2016 Elsevier B.V. All rights reserved.

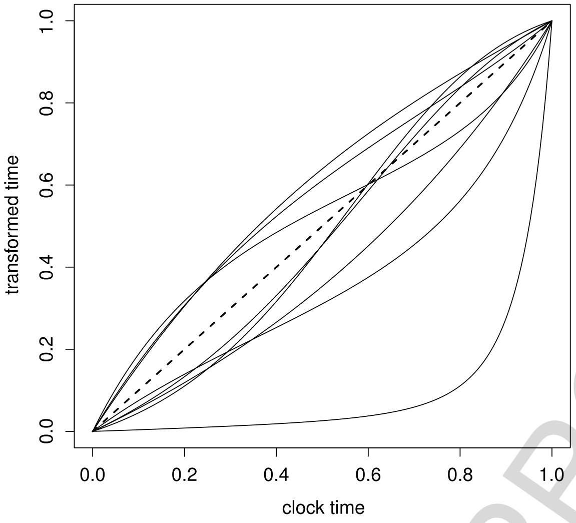

Fig. 1. Examples of warping functions.

the bold dashed line is the 45° reference line representing an identity warp. The cluster structure is blurred by the effect of the time distortions, which should be eliminated for the purpose of clustering. However, it is not feasible to estimate the warping functions without knowing the cluster memberships.

Liu and Yang (2009) propose the SACK model, which is capable of clustering functional data when a simple time translation is presented. They translate the shift in the time domain into variation in the measurement space by a first-order Taylor expansion on the B-spline basis functions. The conditional cluster probabilities are calculated via the EM algorithm. Also assuming a simple time translation, Sangalli et al. (2010) propose an iterative method based on a dissimilarity measure called k-mean alignment, which iterates among a template identification step, alignment and cluster step, and a normalization step until convergence.

To handle more realistic scenarios under arbitrary time warpings, Tang and Müller (2009) propose a method based on pairwise warping. However, this method assumes the mean curves in different clusters are well separated vertically to some degree, a condition potentially too strong for some applications. Zhang and Telesca (2014) propose a hierarchical model for joint curve clustering and registration. They use a reproducing kernel representation of phase variability for registration.

In this paper, we develop a Bayesian method for simultaneous clustering and registration of functional data when both arbitrary time distortions and vertical shifts are possible. We model the curves by B-spline basis functions and we approximate the time warping functions by the cumulative sum of realizations from a Dirichlet distribution (following Cheng et al., in press). The posterior cluster memberships are based on a multinomial distribution. Details of our Markov chain Monte Carlo algorithm are introduced in Section 3. Our method of choosing the number of clusters based on a log likelihood is discussed in Section 4. Various simulation studies indicate that our method is capable of clustering curves accurately. We apply our method to the well-known Berkeley growth data and compare our result with that of the SACK model by Liu and Yang (2009) and the kCFC model of Chiou and Li (2007). We also apply our method to the cell-cycle data collected by Alter et al. (2000).

2. Model assumption

In a functional dataset, we assume that there are N objects, on which we take K measurements over time. Given a certain number of repeated measurements, we may model the response trajectory as a function of time using some basis (such as splines) in the context of functional data analysis.

We assume that each observation is composed of a signal function and random error terms, that is,

$$\mathbf{Y} = af(\mathbf{t}) + \boldsymbol{\epsilon},$$

where $a \in \mathbb{R}^+$ is a stretching/shrinking factor (Zhang and Telesca, 2014), $f(\mathbf{t})$ is the set of underlying responses at the vector of time points $\mathbf{t}$, and $\boldsymbol{\epsilon}$ is an i.i.d. $N(0, \sigma^2)$ error vector.

When our observed data must be aligned, we model the effect of the warping function associated with $\mathbf{Y}$ as $\mathbf{Y} = f[h(\mathbf{t})] + \boldsymbol{\epsilon}$, where h is the underlying warping function, and therefore,

$$\mathbf{Y}|\boldsymbol{\beta}, \gamma, \sigma^2, a \sim \text{MVN}(af[h(\mathbf{t})], \sigma^2\mathbf{I}).$$

For the purpose of clustering, we introduce notation for different groups. For a fixed number of clusters C, we use the vector $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iC})$ to denote the cluster membership for the ith observation. Note that only one element of $\boldsymbol{z}_i$ equals