Computational Statistics and Data Analysis 55 (2011) 2090–2103

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Identifying cluster number for subspace projected functional data clustering

Pai-Ling Li [a,*], Jeng-Min Chiou [b]

[a] Tamkang University, 151 Yingzhuan Rd., Danshui Dist., New Taipei City 25137, Taiwan

[b] Academia Sinica, 128 Sec.2 Academia Rd., Taipei 11529, Taiwan

ARTICLE INFO

Article history:
Received 4 April 2010
Received in revised form 21 December 2010
Accepted 4 January 2011
Available online 14 January 2011

Keywords:
Bootstrapping
Cluster analysis
Functional data analysis
Functional principal components
Gene expression profiles
Hypothesis test

ABSTRACT

We propose a new approach, the forward functional testing (FFT) procedure, to cluster number selection for functional data clustering. We present a framework of subspace projected functional data clustering based on the functional multiplicative random-effects model, and propose to perform functional hypothesis tests on equivalence of cluster structures to identify the number of clusters. The aim is to find the maximum number of distinctive clusters while retaining significant differences between cluster structures. The null hypotheses comprise equalities between the cluster mean functions and between the sets of cluster eigenfunctions of the covariance kernels. Bootstrap resampling methods are developed to construct reference distributions of the derived test statistics. We compare several other cluster number selection criteria, extended from methods of multivariate data, with the proposed FFT procedure. The performance of the proposed approaches is examined by simulation studies, with applications to clustering gene expression profiles.

1. Introduction

In cluster analysis, a partition-based clustering algorithm usually requires a predetermined number of clusters. A properly selected cluster number is critical to search for the optimal partitions within a data set. While numerous approaches have been proposed for choosing a reasonable cluster number in multivariate data, Gordon (1999) referred to the classical selection methods as stopping rules that optimize a particular objective function in terms of cluster numbers. These objective functions are mostly developed by observing the decreasing trend of within-cluster dispersion or the increasing trend of between-cluster dispersion as the cluster number increases. Examples in earlier literature include the index of Caliński and Harabasz (1974), rule of Hartigan (1975), silhouette statistic of Rousseeuw (1987), index of Krzanowski and Lai (1988), and approximate Bayes factor of a model-based approach (Fraley and Raftery, 1998). Studies for comparing various classical cluster number selection methods were provided by Milligan and Cooper (1985), Hardy (1996) and Gordon (1999). More recent studies in cluster number selection include methods based on the gap statistic (Tibshirani et al., 2001) and weighted gap statistic (Yan and Ye, 2007), the approach according to the largest jump in transformed distortion (Sugar and James, 2003), and the method depending on the prediction strength for cluster validation assessment and cluster number estimation (Tibshirani and Walther, 2005). While these aforementioned approaches are designed for multivariate data, the focus of this study is on functional data clustering.

Clustering the longitudinally collected functional data has been remarkably well discussed recently. Numerous functional clustering methods have arisen for finding homogenous subgroups of functional data according to the patterns of the curves. A simple approach is to cluster the finite-dimensional coefficients of basis function expansions using a classical

* Corresponding author.
E-mail address: plli@stat.tku.edu.tw (P.-L. Li).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.01.001

multivariate clustering algorithm (e.g. Abraham et al., 2003; García-Escudero and Gordaliza, 2005; Serban and Wasserman, 2005; Tarpey, 2007). Another popular approach is model-based functional clustering using mixed-effects models coupled with the smoothing techniques (e.g. Luan and Li, 2003; James and Sugar, 2003; Ray and Mallick, 2006; Ma and Zhong, 2008). Based on the nonparametric kernel approach, Ferraty and Vieu (2006) proposed a descending hierarchical method which combines functional data features with a kernel-type functional density estimation. Various depth-based methods for functional data classification have also been discussed recently (e.g. Cuevas et al., 2007; López-Pintado and Romo, 2006) with the major focus on supervised learning context. Furthermore, clustering curves that simultaneously considers curve registration for amplitude and phase variability is another interesting topic in functional data clustering (Liu and Yang, 2009; Sangalli et al., 2010; Tang and Müller, 2009). While most clustering approaches are developed for particularly defined centrality features of curve data, they are mainly based on the mean function. Functional clustering methods that simultaneously take into account the mean and the covariance structures of random functions are proven to be useful. The k-centers functional clustering (FC) of Chiou and Li (2007) regards cluster centers through projection onto the cluster functional principal component (FPC) subspaces such that individual cluster membership is determined by the minimum L^2 distance between the observed curve and the projected function. When the shape patterns of curves are of major interest, proper similarity measures and models are required to achieve the goal. Chiou and Li (2008) proposed a correlation-based FC method that considers shape similarity through maximization of functional correlations coupled with a more flexible shape function model.

In this study, cluster number selection is developed along the lines of subspace projection based functional data clustering. In particular, the method is proposed under the subspace projected functional clustering (SPFC) framework that simultaneously considers differences in the mean and the covariance structures, with the k-centers FC and the correlation-based FC mentioned above as special cases. We adopt the notion that each observed curve is viewed as a realization of a random function and is sampled from a mixture of stochastic processes. Each subprocess represents a cluster through the structure of a FPC subspace that corresponds to a Karhunen–Loève expansion coupled with a random scale. Under the SPFC framework with clusters defined via subspace projection, it is natural to develop a cluster number selection method based upon functional hypothesis tests on cluster subspaces such that the identified clusters are significantly distinct from each other in terms of cluster subspace structures. The null hypotheses comprise equalities between the cluster mean functions and between the sets of cluster eigenfunctions of the covariance kernels. Bootstrap resampling methods are developed to construct the reference distributions of the derived test statistics. The proposed forward functional testing (FFT) procedure aims at selecting the maximum number of distinctive clusters while retaining significant differences between cluster structures. The FFT starts with a small initial cluster number, and then it is increased in steps of one until it reaches the maximum number of distinguishable clusters through the functional hypothesis test procedure. It is shown that the proposed FFT procedure performs reasonably well for data under various cluster structures in our numerical studies.

The rest of this paper is organized as follows. Section 2 introduces the functional multiplicative random-effects model under the SPFC framework. Section 3 presents the procedure of testing differences between two cluster structures, and proposes the FFT algorithm and other relevant cluster number selection procedures for functional data clustering. Simulation studies to examine numerical performance of the proposed FFT and other selection procedures are presented in Section 4. Section 5 illustrates practical applications to two sets of gene expression profile data. Concluding remarks are summarized in Section 6.

2. Subspace projected functional clustering

2.1. Functional multiplicative random-effects model

Suppose that n independent random functions $X_1, X_2, \ldots, X_n$ are sampled from a stochastic process X in $L^2(\mathrm{d}\omega)$, where $L^2(\mathrm{d}\omega)$ represents a Hilbert space of square integrable functions with respect to the measure $\mathrm{d}\omega(t) = w(t)\mathrm{d}t$ on a real interval $\mathcal{T} = [a, b]$, for $a < b$, where $\mathrm{d}t$ is a Lebesgue measure and $w(t)$ is a nonnegative weight function such that $w(t) > 0$ for $t \in \mathcal{T}$ and $w(t) = 0$ otherwise. The inner product of two functions f and g in $L^2(\mathrm{d}\omega)$ is defined as $\langle f, g\rangle = \int f(t)g(t)\mathrm{d}\omega(t)$ and the L^2 norm is defined as $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$. Here, a constant weight function $w(t) = (b-a)^{-1}I_{[t\in\mathcal{T}]}$ is chosen in this study. Further, assume that the process X has a smooth mean function $\mu_x(t) = E(X(t))$ and a smooth covariance function $G_x(s, t) = \mathrm{Cov}(X(s), X(t))$ (twice continuously differentiable). We consider a functional multiplicative random-effects model for the random function $X_i(t)$ such that

$$X_i(t) = \theta_i\left(\mu_x(t) + \sum_{j=1}^{\infty}\xi_{ij}\varphi_{xj}(t)\right) = \theta_i\mu_x(t) + \sum_{j=1}^{\infty}\tau_{ij}\varphi_{xj}(t), \tag{1}$$

where the multiplicative random scales θ_i are positive and uncorrelated with $E(\theta_i) = 1$ and $\mathrm{Var}(\theta_i) = \sigma_\theta^2$. The mean function of X_i can be expressed as $\mu_x(t) = \mu_0 + \mu_z(t)$, where $\mu_0 = \langle\mu_x, 1\rangle$ is constant with respect to t, $\mu_z(t)$ is a fixed mean shape function and $\langle\mu_z, 1\rangle = 0$ consequently. The random effects ξ_{ij} are uncorrelated with zero mean and variance $\sigma_{\xi_j}^2$. Here, we assume that ξ_{ij} and θ_i are independent. The random effects $\tau_{ij} = \theta_i\xi_{ij} = \langle X_i - \theta_i\mu_x, \varphi_{xj}\rangle$ are uncorrelated with zero mean and variance $\sigma_{\tau_j}^2 = (\sigma_\theta^2 + 1)\sigma_{\xi_j}^2$. The set of functions $\{\varphi_{xj}\}$ forms an orthornormal basis in L^2 associated with the covariance