ACADEMIC
World Academic Union

ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 14, No. 3, 2019, *pp*.184-194

Functional clustering with application to air quality analysis

Ming He, Hairong Li, Xiaoxin Zhu, Chunzheng Cao[1]

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China

(Received March 21 2019, accepted June 20 2019)

Abstract. Based on the air quality status of 161 cities in China, this paper studies the temporal and spatial distribution characteristics of PM2.5 concentration of major pollutants affecting air quality index (AQI). We use improved functional clustering analysis methods and add priori information about location and human factors to make the clustering results more accurate. The improved functional clustering model is compared with the basic sparse data function clustering method, k-centres functional clustering method, functional principal component analysis and traditional K-means clustering method by repeated simulation. Finally, we use the PM2.5 concentration of selected 161 cities in China as an illustrative example.

Keywords: air quality index, PM2.5 concentration, functional clustering, priori information

1. Introduction

Air quality influences human health and economic development. Nowadays air quality is measured by air quality index (AQI), which is typically a temporal-spatial data. This research is motivated by an air quality influences human health and economic development.

Many existing studies have analyzed air quality but they are limited to use simple statistical methods and spatial correlations [1, 2]. And some researches just consider dozens of cities such as Chen used EPLS method to analysis air quality of 31 provincial capitals in China mainland [3]. Hamedian [4] mainly used fuzzy C-mean clustering to find the main pollutants which can influence the air quality. In this article, air quality is measured by Air Quality Index (AQI) and PM2.5, PM10, SO2, O3, CO, NO2 six pollutants of 161 cities in China. We leverage new statistical methods for estimating and describing air quality trends and distribution that can be used to inform about spatial and temporal distribution characteristics.

Cluster analysis is the art of identifying groups in data. Traditional clustering methods are focus on multivariate data and many clustering algorithms have been proposed when the data are curves or functions. In this context, Functional Data Analysis has received increasing attention recently [5].

Several clustering methods for functional data have been researched in recent years. The two-stage approach was proposed by Maharaj [6] and used by Iorio et al. [7] to handle time course data with observed measurements. P-splines smoothers was used to model the observed measurements and then to cluster functions by the optimal spline coefficients. They added penalty term based on the general basis expansion and fitted the curves well by choosing smoothing parameter. Traditional approaches based on clustering basis coefficients choose the same basis functions for all clusters to use the fitted coefficients to be clustered. There are some problems because basis functions should be chosen then the fitted coefficients can adequately adapt cluster differences. For the model-based clustering method, Same´ and Bouveyron used this approach in mixture model based on high dimensional data [8, 9]. Basically, the model parameters are always estimated by the maximum likelihood method solved by an Expectation-Maximization (EM) algorithm [10]. When the observations are sparse, irregularly spaced, or occur at different time points for each subject, James and Sugar proposed a particularly effective model-based approach for clustering functional data [11]. They produced low-dimension representation of the curves and then provided low-dimension graphical representations to show some direct clustering results in the pictures. In fact, various model-based approaches are under certain probability model assumptions. Just considering the information of curves themselves without some correlated variables may not cluster well. Chiou and Li [12] proposed a k-centres functional clustering method which can greatly improve cluster quality compared with the conventional clustering algorithms. The k-centres functional clustering method does not rely on any distribution assumptions and the mean and covariation structures of each cluster are explored using this approach.

[1] Corresponding author. *E-mail address*: caochunzheng@163.com.

This study is concemed with functional data clustering where the number of observations is 161 cities in China and the recording times are the same for individuals. Thus, the algorithm for sparse samples should be improved and we consider to add the position information as prior when fit the air quality curves. A logistic function and its similar form were considered to be probability by many model-based functional clustering approaches [13]. It is worth noting that the determination of a clustering technique is even more difficult under the possible presence of outlying curves. One possibility to improve the robustness of clustering algorithm is through the application of trimming tools. In this study, low-dimension representation and visual exhibition are considered. Combining logistic prior information and robust trimming tools, we compare with two typical functional clustering methods: k-centres functional clustering and sparsely functional clustering.

This paper is organized as follows. In section 2, we define the proposed model, also detail the method of parameter estimations, the curve clustering and model selection. In section 3, we compare the improved functional clustering model with other clustering methods through repeated simulations. In section 4, we conducted a cluster analysis of PM2.5 concentrations in selected 161 cities in China, and compared the differences in air quality between different types of cities. Finally, some conclusions are presented in section 5.

2. Methodology

2.1. The model

Let $f_i(t)$ be the value for the i-th smooth underlying curve. The observed data can be expressed as

$$y_{ij} = f_i(t_{ij}) + \varepsilon_{ij}, \quad i = 1,\ldots,n,\ j = 1,\ldots,n_i, \tag{1}$$

where $\boldsymbol{y}_i = (y_{i1},\ldots,y_{in_i})^T$ are the vectors of observed values at time points $\boldsymbol{t}_i = (t_{i1},\ldots,t_{in_i})^T$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1},\ldots,\varepsilon_{in_i})^T$ are measurement errors following $N(0,\sigma^2\boldsymbol{I}_{n_i})$. The subject-specific random function $f_i(t)$ is a Gaussian process, which can be approximated as

$$f_i(t) = \mathbf{s}(t)^T\boldsymbol{\eta}_i, \tag{2}$$

where $\boldsymbol{s}(t)$ is p-dimensional vector of spline basis function with $\boldsymbol{\eta}_i$ is coefficient vector of the spline basis, which can be modeled using the following Gaussian distribution

$$\boldsymbol{\eta}_i\,|_{z_{ik}=1} = \boldsymbol{\mu}_k + \boldsymbol{\gamma}_i, \quad \boldsymbol{\gamma}_i \sim N(\mathbf{0},\boldsymbol{\Gamma}_k), \tag{3}$$

where the latent label z_{ik} denotes the cluster membership vector for the i-th individual, when $z_{ik} = 1$, $f_i(t)$ belongs to the k-th cluster and $z_{ik} = 0$ otherwise.

In model-based clustering it is assumed that the observations $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ follow a mixture distribution with K components. In addition, $\boldsymbol{z}_i = (z_{i1},\ldots,z_{iK})^T$ follows a multinomial distribution with parameter $(\pi_{i1},\ldots,\pi_{iK})^T$ and π_{ik} is the probability of the i-th observation belongs to the k-th cluster. Suppose there exists a p_w dimensional covariates $\boldsymbol{w}_i = (1, w_{1i},\ldots,w_{p_w-1,i})^T$ which can influence the categorical latent variable $\boldsymbol{z}_i$ through a logistic model

$$\pi_{ik} \triangleq P(z_{ik}=1) = \frac{\exp(\mathbf{w}_i^T\mathbf{v}_k)}{1+\sum_{j=1}^{K-1}\exp(\mathbf{w}_i^T\mathbf{v}_j)}, \quad k = 1,\ldots,K-1, \tag{4}$$

where $\boldsymbol{v}_K = \mathbf{0}$ for identifiability and $\sum_{k=1}^{K}\pi_{ik} = 1$. Thus, the functional clustering model can be written as

$$\begin{aligned} &\mathbf{y}_i\,|_{z_{ik}=1} = \mathbf{S}_i(\boldsymbol{\mu}_k + \boldsymbol{\gamma}_i) + \boldsymbol{\varepsilon}_i,\ i = 1,\ldots,n, \\ &\boldsymbol{\gamma}_i \sim N(\mathbf{0},\boldsymbol{\Gamma}_k), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0},\sigma^2\mathbf{I}_{n_i}), \end{aligned} \tag{5}$$

where $\boldsymbol{S}_i = (\boldsymbol{s}(t_{i1}),\ldots,\boldsymbol{s}(t_{in_i}))^T$ is the spline basis matrix for the i-th curve.

2.2. Parameter estimation

We recommend using the EM-algorithm to obtain the MLE of all the parameters. Since the $\boldsymbol{z}_i$'s and $\boldsymbol{\gamma}_i$'s are assumed independent each other, the combined density distribution of the complete data $\{\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{z}\}$ can be expressed as

$$p(\mathbf{y},\boldsymbol{\gamma},\mathbf{z}) = p(\mathbf{y}\,|\,\boldsymbol{\gamma},\mathbf{z})\,p(\boldsymbol{\gamma}\,|\,\mathbf{z})\,p(\mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{y}_i\,|\,\boldsymbol{\gamma}_i,\mathbf{z}_i)\,p(\boldsymbol{\gamma}_i\,|\,\mathbf{z}_i)\,p(\mathbf{z}_i), \tag{6}$$

where