Research Gallery: Clustering Spatially Correlated Functional Data with Multiple Scalar Covariates

Hui Wu and Yan-Fu Li
Department of Industry Engineering, Tsinghua University

1 Research Overview

Hui Wu and her team focus on developing statistical and machine learning models for complex data (e.g., mixed-type data, functional data, insufficient information data) in industrial systems. This research overview will present one of their works on a clustering algorithm for spatially correlated functional data with covariates, followed by future research plans. This research was awarded as the best student paper in DAIS of IISE conference 2021.

2 Clustering Algorithm

The objective of this research is to develop a probabilistic model for clustering spatially correlated functional data with multiple scalar covariates. The motivating application is to cluster the provinces of the Chinese mainland into three groups: low, medium, and high-risk areas, while the spatial dependence and effects of risk factors are considered. Since the epidemic situations in Tibet and Hubei are unique and thus readily identifiable by our method, we only consider the other 29 provinces. The dataset consists of daily confirmed-case records from January 20th to February 19th, 2020 (totally 31 records), longitude and latitude data, and nine risk factors: the number of permanent residents, the gross domestic product (GDP), the GDP index, the number of medical and health institutions, the number of beds in health institutions, the number of health technicians, the added value of the tertiary industry, passenger volume and the proportion of the tertiary industry, of the 29 provinces. An intuitive idea is to apply functional data clustering algorithm to the profiles of daily confirmed-case records. However, it would involve a waste of information on spatial location and risk impact factors, which significantly influences the accuracy of modeling and prediction. To account for the spatial influence, extensive clustering algorithms for spatially correlated functional data have been investigated, including discriminative and generative (i.e., model-based) methods. However, these existing methods are only applicable to the functional variables and do not account for dependencies between functional variables and scalar covariates. When there exists a clear regression relationship between them, significant insight can be gained by accounting for such dependencies.

To bridge these gaps, we propose a probability model for clustering spatially correlated functional data with multiple scalar covariates, which has the ability to interpret the clustering structure and the spatial influence. The proposed model can be regarded as an extension of mixture models, which allows different subsets of covariates to influence the component weights and the component densities by modeling the parameters of the mixture as functions of the covariates. In particular, the spatial dependence is introduced by allowing component weight to be location-specific, obtained as a multinomial logistic regression (MLR) model depending on spatial covariates. The heterogeneous relationship between functional responses and scalar covariates is represented by incorporating a regression model into the component densities. A graphical model representation is shown in Figure 1. Note that the probability assigned to the same cluster for two samples is influenced by similarity in three aspects: pattern of functional data, spatial location, and dependence on the related explanatory covariates. Such a model can be interpreted from two views. From a clustering perspective, once the estimates of the parameters are given, we can deduce data clustering by the so-called maximum a posteriori (MAP) principle. From a prediction perspective, the model helps us understand how the observations are affected by the covariates for different groups or subjects, which allows us to predict the output for a set of new covariates. Furthermore, we develop an L_1-penalized estimator to assist variable selection and robust estimation to cope with high-dimensional covariates. An efficient expectation-maximization algorithm is presented for parameter estimation. Theoretically, the identifiability of the proposed model is analyzed, and sufficient conditions to guarantee “generic” identifiability are provided.

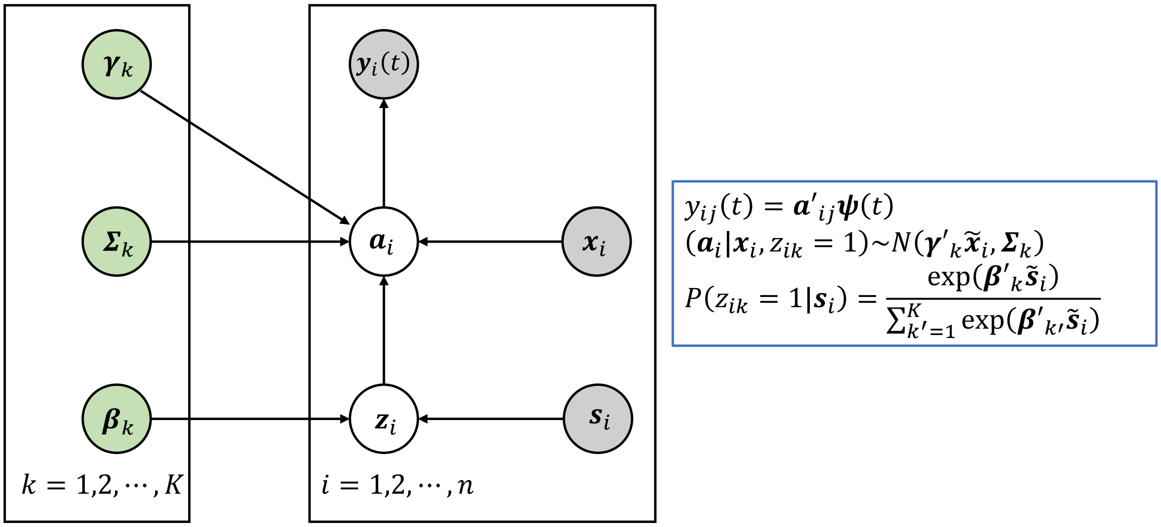
$\boldsymbol{\gamma}_k$
$\boldsymbol{y}_i(t)$
$\boldsymbol{\Sigma}_k$
$\boldsymbol{a}_i$
$\boldsymbol{x}_i$
$y_{ij}(t) = \boldsymbol{a}'_{ij}\boldsymbol{\psi}(t)$
$(\boldsymbol{a}_i|\boldsymbol{x}_i, z_{ik} = 1) \sim N(\boldsymbol{\gamma}'_k\tilde{\boldsymbol{x}}_i, \boldsymbol{\Sigma}_k)$
$P(z_{ik} = 1|\boldsymbol{s}_i) = \frac{\exp(\boldsymbol{\beta}'_k\tilde{\boldsymbol{s}}_i)}{\sum_{k'=1}^{K}\exp(\boldsymbol{\beta}'_{k'}\tilde{\boldsymbol{s}}_i)}$
$\boldsymbol{\beta}_k$
$\boldsymbol{z}_i$
$\boldsymbol{s}_i$
$k = 1,2,\cdots,K$
$i = 1,2,\cdots,n$

Figure 1: The graphical representation of the proposed method.

3 Future Research

This research can be extended in several directions. A possible direction is to extend it to semi-supervised model-based classification settings, where a part of observations are labeled. In addition, novel internal validity measures deserve to be developed to evaluate the clustering algorithms in real application scenarios, where the true data labels are not available. Furthermore, model selection is an interesting but difficult issue for a mixture model, especially for models with very complicated forms. In our work, we use the BIC to select the number of clusters. Further research might be worth extending to the mixture model with an unknown number of clusters by using a nonparametric prior.