Journal of Statistical Planning and Inference 166 (2015) 171–186

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Bayesian clustering of shapes of curves

Zhengwu Zhang*, Debdeep Pati, Anuj Srivastava

Department of Statistics, Florida State University, FL, 32306, United States

ARTICLE INFO

Article history:
Received 2 December 2013
Received in revised form 10 March 2015
Accepted 1 April 2015
Available online 28 April 2015

Keywords:
Clustering
Shapes of curves
Chinese restaurant process
Wishart distribution

ABSTRACT

Unsupervised clustering of curves according to their shapes is an important problem with broad scientific applications. The existing model-based clustering techniques either rely on simple probability models (e.g., Gaussian) that are not generally valid for shape analysis or assume the number of clusters. We develop an efficient Bayesian method to cluster curve data using an elastic shape metric that is based on joint registration and comparison of shapes of curves. The elastic-inner product matrix obtained from the data is modeled using a Wishart distribution whose parameters are assigned carefully chosen prior distributions to allow for automatic inference on the number of clusters. Posterior is sampled through an efficient Markov chain Monte Carlo procedure based on the Chinese restaurant process to infer (1) the posterior distribution on the number of clusters, and (2) clustering configuration of shapes. This method is demonstrated on a variety of synthetic data and real data examples on protein structure analysis, cell shape analysis in microscopy images, and clustering of shapes from MPEG7 database.

1. Introduction

The automated clustering of objects is an important area of research in unsupervised classification of large object databases. The general goal here is to choose groups (clusters) of objects so as to maximize homogeneity within clusters and minimize homogeneity across clusters. The clustering problem has been addressed by researchers in many disciplines. A few well-known methods are metric based e.g. K-means (MacQueen et al., 1967), hierarchical clustering (Ward, 1963), clustering based on principal components, spectral clustering (Ng et al., 2002) and so on (Jain and Dubes, 1988; Ozawa, 1985). Traditional clustering methods are complemented by methods based on a probability model where one assumes a data generating distribution (e.g., Gaussian) and infers clustering configurations that maximize certain objective function (Banfield and Raftery, 1993; Fraley and Raftery, 1998, 2002, 2006; MacCullagh and Yang, 2008). A model-based clustering can be useful in addressing challenges posed by traditional clustering methods. This is because a probability model allows the number of clusters to be treated as a parameter in the model, and can be embedded in a Bayesian framework providing quantification of uncertainty in the number of clusters and clustering configurations.

A popular probability model is obtained by considering that the population of interest consists of K different sub-populations and the density of the observation y from the kth sub-population is f_k. Given observations $y_1, \ldots, y_n$, we introduce indicator random variables $(c_1, \ldots, c_n)$ such that $c_i = k$ if y_i comes from the kth sub-population. The maximum likelihood inference is based on finding the value of $(c, f_1, \ldots, f_k)$ that maximizes the likelihood $\prod_{i=1}^{n} f_{c_i}(y_i)$. Typically K is assumed to be known or a suitable upper bound is assumed for convenience. When $y_i \in \mathbb{R}^p$, f_k is commonly parameterized by a multivariate Gaussian density with mean vector μ_k and covariance matrix Σ_k. An alternative is to use a nonparametric

* Corresponding author. Tel.: +1 850 644 7412.
E-mail addresses: zhengwu@stat.fsu.edu (Z. Zhang), debdeep@stat.fsu.edu (D. Pati), anuj@stat.fsu.edu (A. Srivastava).

http://dx.doi.org/10.1016/j.jspi.2015.04.007
0378-3758/© 2015 Elsevier B.V. All rights reserved.

Bayesian approach which has an appealing advantage of allowing K to be unknown and inferring it from the data. An advantage of such an approach is that it not only provides an estimate of the number of clusters, but also the entire posterior distribution.

The vast majority of the literature on model-based clustering is almost exclusively focused on Euclidean data. This is primarily due to the easy availability of parametric distributions on the Euclidean space as well as computational tractability of estimating the cluster centers. For clustering functional data, e.g. shapes of curves, one encounters several challenges. Unlike Euclidean data, where the notions of cluster center and cluster variance are standard, these quantities and the resulting quantification of homogeneity within clusters are not obvious for shape spaces. Moreover, it is important to use representations and metrics for clustering objects that are invariant to shape-preserving transformations (rigid motions, scaling, and re-parameterization). For example, Kurtek et al. (2012) take a model-based approach for clustering of curves using an elastic metric that has proper invariances. However, under the chosen representations and metrics, even simple summary statistics of the observed data are difficult to compute. Other existing shape clustering methods (Belongie et al., 2002; Liu et al., 2012) either extract finite-dimensional features to represent the shapes or project the high-dimensional shape space to a low-dimensional space (Yankov and Keogh, 2006; Auder and Fischer, 2012), and then apply clustering methods for Euclidean data; these approaches are not necessarily invariant to shape preserving transformations. Also, several methods (Srivastava et al., 2005; Gaffney and Smyth, 2005) have been proposed to cluster non-Euclidean data based on a distance-based notion of dispersion, thus, avoiding the computation of shape means (e.g. Karcher means), but they all assume a given number of clusters.

In this paper we develop a model-based clustering method for curve data that does not require the knowledge of cluster number K a priori. This approach is based on modeling a summary statistic that encodes the clustering information, namely the inner product matrix. The salient points of this approach are: (1) The comparison of curves is based on the inner product matrix under elastic shape analysis, so that the analysis is invariant to all desired shape-preserving transformations. (2) The inner product matrix is modeled using a Wishart distribution with prior on the clustering configurations induced by the Chinese restaurant process (Vogt et al., 2010). A model directly on the inner product matrix has an appealing advantage of reducing computational cost substantially by avoiding computation of the Karcher means. (3) We formulate and sample from a posterior on the number of clusters, and use the mode of this distribution for final clustering. We illustrate our ideas through several synthetic and real data examples. The results show that our model on the inner product matrix leads to a more accurate estimate of the number of clusters as well as the clustering configurations compared to a Bayesian nonparametric model directly on the data, even in the Euclidean case.

This paper is organized as follows. We start by introducing two case studies in Section 2. The mathematical details of the metric used for computing the inner product and the model specifications are presented in Section 3. In Section 4, we illustrate our methodology on several synthetic data examples and the case studies on clustering cell shapes and protein structures. Section 5 closes the paper with some conclusions.

2. Case studies

We propose to undertake two specific case studies involving clustering of curve data.

2.1. Clustering of protein sequences

Protein structure analysis is an outstanding scientific problem in structural biology. A large number of new proteins are regularly discovered and scientists are interested in learning about their functions in larger biological systems. Since protein functions are closely related to their folding patterns and structures in native states, the task of structural analysis of proteins becomes important. In terms of evolutionary origins, proteins with similar structures are considered to have common evolutionary origin. The Structural Classification of Proteins (SCOP) database (Murzin et al., 1995) provides a manual classification of protein structural domains based on similarities of their structures and amino acid sequences. Refer to Fig. 1 for a snapshot of the proteins in $\mathbb{R}^3$ and the 3-coordinates of the protein sequences. Clustering protein sequences is extremely important to trace the evolutionary relationship between proteins and detect conserved structural motifs. In this article, we focus on an automated clustering of protein sequences based on their global structures.

2.2. Clustering of cell shapes

The problem of studying shapes of cellular structures using microscopic image data is very important medical diagnosis (Rohde et al., 2008) and genetic engineering (Thomas et al., 2002). This research involves extracting cell contours from images using segmentation techniques (Hagwood et al., 2012) and then studying shapes of these extracted contours for medical diagnosis. We will focus on the problem of clustering of cells according to their shapes; these clusters can be further used for statistical modeling and hypothesis testing although these steps are not pursued in the current paper. The specific database used here was obtained by segmenting the 2D microscopy images, as described in Hagwood et al. (2012, 2013). Fig. 2(b) shows some examples of the cell contours used in this paper. In this article, we consider two types of cell shapes: DLEX-p46 cell shapes and NIH-3T3 cell shapes. Visually, DLEX-P46 cells are round, denoting normal cell shapes whereas, NIH-3T3 cells have an elongated, spindly appearance, denoting progression of some pathological conditions.