Computational Statistics and Data Analysis 55 (2011) 2128–2136

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Penalized cluster analysis with applications to family data

Yixin Fang [a,*], Junhui Wang [b]

[a] Department of Mathematics and Statistics, Georgia State University, United States

[b] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, United States

ARTICLE INFO

Article history:
Received 5 July 2010
Received in revised form 2 January 2011
Accepted 4 January 2011
Available online 14 January 2011

Keywords:
Consistency
Cross-validation
Kinship
K-means
Stability

ABSTRACT

The goal of cluster analysis is to assign observations into clusters so that observations in the same cluster are similar in some sense. Many clustering methods have been developed in the statistical literature, but these methods are inappropriate for clustering family data, which possess intrinsic familial structure. To incorporate the familial structure, we propose a form of penalized cluster analysis with a tuning parameter controlling the tradeoff between the observation dissimilarity and the familial structure. The tuning parameter is selected based on the concept of clustering stability. The effectiveness of the method is illustrated via simulations and an application to a family study of asthma.

1. Introduction

The goal of cluster analysis is to assign observations into clusters so that observations in the same cluster are similar in some sense. Popular methods include *K*-means clustering (MacQueen, 1967), hierarchical clustering (Johnson, 1967), *K*-medoids clustering (Kaufman and Rousseeuw, 1990), and spectral clustering (Shi and Malik, 2000).

However, these methods are inappropriate for clustering family data, which are obtained from family study, an important type of sampling design in genetic epidemiology (e.g., Khoury et al., 1993). Family data often possess intrinsic familial structure. Ignoring the familial structure leads to loss of information, whereas forcing members in the same family into the same cluster can be too restrictive. In this manuscript we develop a form of penalized cluster analysis to find a reasonable compromise between these two extremes.

To illustrate our method, we use one dataset originally collected as part of the Collaborative Study on the Genetics of Asthma (The Collaborative Study on the Genetics of Asthma, CSGA). This dataset was also analyzed by Reilly et al. (2007). In the dataset we received (a little bit different from the one used in Reilly et al., 2007), there were 29 families, and totally there were 187 asthmatic members. Four phenotypes considered were the logarithm of the percent predicted of the following variables: volume exhaled during the first second of a forced expiratory maneuver (FEV1), forced expiratory vital capacity (FVC), maximum expiratory flow when half of the FCV has been exhaled (FEFM), and forced expiratory flow rate over the middle half FCV (FF25).

The method developed in Reilly et al. (2007) was based on seeking clusters of families that differ, between clusters, in the way affected individuals express the genotype. They proposed to use the distance of each individual to the cluster center for his family to define a quantitative trait for further genetic analysis. In order to do such cluster analysis, they obtained the set of averaged phenotypes within families, $\bar{\mathbf{y}}_i, i = 1, 2, \ldots, n$, where n is the number of families. Then they applied *K*-means to this set of averaged phenotypes. This method has the following three limitations. One, due to genetical heterogeneity within

* Corresponding address: Department of Mathematics and Statistics, Georgia State University, 750 COE, 30 Pryor Street, Atlanta, GA 30303, United States.
E-mail addresses: matyxf@langate.gsu.edu (Y. Fang), junhui@uic.edu (J. Wang).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.01.004

families, it can be too restrictive to force members in the same family into the same cluster. Two, the covariances of $\bar{\mathbf{y}}_i$ are heterogeneous as family sizes are different. Third, the number of families is small compared with the sample size, so the cluster analysis on the family level can be unreliable. Fortunately, these three limitations can be overcome by the proposed penalized cluster analysis.

The rest of the manuscript is organized as follows. In Section 2, we propose a form of penalized cluster analysis. In Section 3, we introduce the concept of clustering stability. In Section 4, we propose a cross-validation procedure based on clustering stability to select the tuning parameters in the penalized cluster analysis and discuss its consistency. In Section 5, the proposed method is illustrated via simulations and an application to the asthma data. Section 6 contains some discussion and the Appendix is devoted to the technical proofs.

2. Penalized cluster analysis

Let $\mathbf{y}_{ij} = (y_{ij}^1, \ldots, y_{ij}^p)^T$ be the p-vector of phenotypes measured for the jth member in the ith family, where $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, and $N = \sum_{i=1}^n n_i$. Let $d(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ be the distance between $\mathbf{y}_{ij}$ and $\mathbf{y}_{i'j'}$. Here $d(\cdot, \cdot)$ could be any kind of distance. For those quantitative phenotypes in the asthma data, as in Reilly et al. (2007), we consider the squared Euclidean distance, $d(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}) = \|\mathbf{y}_{ij} - \mathbf{y}_{i'j'}\|^2$.

Let $F(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ be the kinship coefficient between subjects $\mathbf{y}_{ij}$ and $\mathbf{y}_{i'j'}$. Here the kinship coefficient between two subjects is defined as two times the probability that a randomly selected allele will be identical by descent (IBD) between them (e.g., Lange, 1997). It is 0 between unrelated individuals because there are no two alleles from them respectively coming from the same ancestor. If there is no inbreeding in the pedigree, it will be 1 for an individual with himself (we could choose the same allele twice), 1/2 between mother and child, 1/2 between siblings, 1/8 between first cousins, and so on. The kinship coefficients can be easily computed by R package "kinship" when the family kinship information is provided.

Now the penalized cluster analysis with a pre-specified number of clusters K can be defined as solving

$$\min_{\psi} W(\psi) = \frac{1}{2} \sum_{k=1}^{K} \sum_{\psi(\mathbf{y}_{ij}) = \psi(\mathbf{y}_{i'j'}) = k} D_\lambda(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}), \tag{1}$$

where $\psi(\mathbf{y})$ is a clustering function that maps $\mathbf{y}$ to its cluster membership in $\{1, \ldots, K\}$, and the inside summation is over all the pairs $\mathbf{y}_{ij}$ and $\mathbf{y}_{i'j'}$ that are in cluster k. Here $D_\lambda(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ is the dissimilarity between subjects $\mathbf{y}_{ij}$ and $\mathbf{y}_{i'j'}$,

$$D_\lambda(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}) = d(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}) + \lambda(1 - F(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})), \tag{2}$$

where λ is a tuning parameter that controls the tradeoff between the distance in phenotypes and the kinship coefficient.

In particular, if the data only contain the family indices instead of the kinship information, we can define $F(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ as 1 if $i = i'$ and 0 otherwise. The method can also be applied to other structured data such as panel data, where there is no kinship information. For this aim, to abuse the notation, we simply define $F(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ as 1 if subjects i and i' are in the same panel, and 0 otherwise.

Note that it is infeasible to optimize (1) over all possible candidate clustering functions $\psi(\mathbf{y})$. In practice, one may restrict the candidate clustering functions to those obtained via some feasible strategies. For example, if no kinship information is available and $F(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}) = 1$ if $i = i'$ and 0 otherwise, we can consider K-means clustering. We extend $\mathbf{y}_{ij}$ to $(p + n)$-vector $\mathbf{y}_{ij}^* = (\mathbf{y}_{ij}^T, 0, \ldots, 0, (\lambda/2)^{1/2}, 0, \ldots, 0)^T$, where $(\lambda/2)^{1/2}$ is at the $(p + i)$th component, leading to that $\|\mathbf{y}_{ij} - \mathbf{y}_{i'j'}\|^2 + \lambda I(i \neq i') = \|\mathbf{y}_{ij}^* - \mathbf{y}_{i'j'}^*\|^2$. Therefore, applying standard K-means clustering to $\mathbf{y}_{ij}^*$ is equivalent to solving the penalized cluster problem in (1).

Clearly, the effectiveness of the proposed penalized cluster analysis in (1) largely depends on the values of K and λ. For example, if $\lambda = 0$, $D_0(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ degenerates to regular distance and the familial structure is not accounted for; if λ is large enough, members in the same family are forced into the same cluster. Therefore, it is important to develop a tuning technique to appropriately select the number of clusters K and the tuning parameter λ.

In the literature, many model selection criteria have been proposed for selecting K. Most of them are based on between-cluster and/or within-cluster sum of squares; to name just a few, Calinski and Harabasz (1974), Hartigan (1975), and Krzanowski and Lai (1985). Additionally, the silhouette statistic proposed by Kaufman and Rousseeuw (1990), the gap statistic proposed by Tibshirani et al. (2001), and the jump statistic proposed by Sugar and James (2003) can also be applied to select the number of clusters. However, these methods are designed to select K only and it is unclear whether they can be modified to select λ as well. In the following two sections, we develop a tuning method based on clustering stability that can be used to select both K and λ.

3. Clustering stability

In the literature of cluster analysis, clustering stability was suggested to assess the quality of clustering by measuring its robustness against the randomness in the sample. See, e.g., Fowlkes and Mallows (1983), Gnanadesikan (1997), Ben-Hur et al. (2002), Lange et al. (2004), Ben-David et al. (2006), and the references therein. As discussed in Wang (2010), the intuition is that if we repeatedly draw samples from the population and apply the given clustering algorithm, a good one