Penalized cluster analysis with applications to family data

Yixin Fang*
Department of Mathematics and Statistics
Georgia State University
Email: matyxf@langate.gsu.edu

Junhui Wang
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Email: junhui@uic.edu

Abstract

Cluster analysis is the assignment of observations into clusters so that observations in the same cluster are similar in some sense, and many clustering methods have been developed. However, these methods cannot be applied to family data, which possess intrinsic familial structure. To take the familial structure into account, we propose a form of penalized cluster analysis with a tuning parameter controlling its influence. The tuning parameter can be selected based on the concept of clustering stability. The method can also be applied to other cluster data such as panel data. The method is illustrated via simulations and an application to a family study of asthma.

Keywords: Consistency; Cross-validation; Kinship; K-means; Stability.

*Correspondence to: Yixin Fang, 750 COE, 30 Pryor Street, Atlanta, GA 30303, U.S.A.

1 Introduction

Cluster analysis is the assignment of 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 cannot be applied to family data, which possess intrinsic familial structure. Usually, family data are generated from family study, an important type of sampling design in genetic epidemiology (e.g., Khoury, Beaty, and Cohen, 1993). 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 possible compromise between these two extreme ways.

To illustrate our method, we use one dataset originally collected as part of the Collaborative Study on the Genetics of Asthma (CSGA, 1997). 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) is 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, $\overline{\mathbf{y}}_i$, $i = 1, 2, \cdots, n$, which are p-vectors, where n is the number of families and p is the number of phenotypes. Then they

applied K-means to this set of averaged phenotypes. This method has the following three limitations. One, due to genetical heterogeneity within families, it can be too restrictive to force members in the same family into the same cluster. Two, the covariances of $\overline{\mathbf{y}}_i$ 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 applying the 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 Appendix is devoted to the technical proofs.

2 Penalized cluster analysis

Let $\mathbf{y}_{ij} = (y_{ij}^1, \cdots, y_{ij}^p)^T$ be the p-vector of phenotypes measured for the jth member in the ith family, where $i = 1, \cdots, n$, $j = 1, \cdots, n_i$, 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 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 a 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 $\{1, \cdots K\}$, and $D_\lambda(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ is the dissimilarity between subjects $\mathbf{y}_{ij}$ and $\mathbf{y}_{i'j'}$,

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

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

Remark 1: If the data only contain the family index instead of the kinship information, we can define $F(\mathbf{y}_{ij}, \mathbf{y}_{i'j'})$ as 1 if $i = i'$ and 0 otherwise.

Remark 2: The method can also be applied to other clustered data such as panel data, where there is no kinship information. 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 cluster, and 0 otherwise.

Note that it is infeasible to optimize (1) over all the possible candidate clustering functions $\psi(\mathbf{y})$. In practice, one may restrict the candidate clustering functions to those obtained via some feasible strategies such as hierarchical clustering (Johnson, 1967), K-medoids (Kaufman and Rousseeuw, 1990), and spectral clustering (Ng *et al.,* 2002). In particular, if no kinship information is available and $F(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}) = 1$ if $i = i'$ and 0 otherwise, we can also consider K-means clustering. To perform K-means, we extend $\mathbf{y}_{ij}$ to $(p+n)$-vector $\mathbf{y}^*_{ij} = (\mathbf{y}^T_{ij}, 0, \cdots, 0, (\lambda/2)^{1/2}, 0, \cdots, 0)^T$, 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 no family structure is included; 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, Walther, and Hastie (2001), and the jump statistic proposed by Sugar and James (2003) can also be applied to select the number of clusters. However, these methods cannot be modified to be used to select an appropriate tuning parameter λ. 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

Recently, clustering stability has been suggested to assess the quality of clustering by measuring its robustness against the randomness in the sample. See, e.g., Ben-Hur, Elisseeff, and Guyon (2002), Lange *et al.* (2004), and Ben-David, von Luxburg, and Pal (2006). 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 should produce clusterings that do not vary much from one sample to another. The stability measure is assumption free and applicable to both distance based and non-distance based clustering algorithms.

In family data, families are sampling units. Assume that n families are randomly sampled from a population of families. Let $Y_i = (\mathbf{y}_{i1}, \cdots, \mathbf{y}_{in_i})$ stack the observations from the

ith families. Therefore, Y_i, $i = 1, \cdots, n$, can be assumed to be independent and identically distributed with a distribution, say $f(Y_0)$, where Y_0 is the random matrix including observations from a family randomly selected from the population. In particular, let $g(\mathbf{y}_0)$ be the marginal distribution of a subject randomly selected from the population. Denote the sample of n families as $Y^n = (Y_1, \cdots, Y_n)$.

A clustering function $\psi(\mathbf{y})$ is defined as a mapping $\psi(\mathbf{y}) : \mathbb{R}^p \to \{1, \cdots, K\}$, and a clustering algorithm $\Psi(\cdot; K, \lambda)$ given a number of clusters $K \geq 2$ and a tuning parameter λ yields a clustering function $\psi(\mathbf{y})$ when applied to the sample Y^n. Here the case of $K = 1$ is excluded, because any clustering algorithm with $K = 1$ leads to the same degenerate clustering that is trivially stable.

To select the number of clusters K, Ben-David *et al.* (2006) and Wang (2010) defined the clustering distance and clustering instability. However, for some reasons to be discussed soon, here we propose the following definition to measure the correlation between two clustering functions, similar to the one used to measure the closeness between two hierarchical clustering functions in Fowlkes and Mallows (1983).

Definition 1 *(Clustering Correlation and Stability) For any two clusterings $\psi_1(\mathbf{y})$ and $\psi_2(\mathbf{y})$, the correlation between them is defined as*

$$Corr(\psi_1, \psi_2) = \frac{P(I_1 = I_2 = 1) - P(I_1 = 1)P(I_2 = 1)}{\sqrt{P(I_1 = 1)(1 - P(I_1 = 1))P(I_2 = 1)(1 - P(I_2 = 1))}}, \tag{3}$$

where $I_1 = I\{\psi_1(\mathbf{x}) = \psi_1(\mathbf{y})\}$ and $I_2 = I\{\psi_2(\mathbf{x}) = \psi_2(\mathbf{y})\}$, with $\mathbf{x}$ and $\mathbf{y}$ being sampled from $g(\mathbf{y}_0)$. (The R.H.S. of (3) is defined as 0 if its denominator equals 0.) For any clustering algorithm $\Psi(\cdot; K, \lambda)$, the stability is defined as

$$Stab(\Psi, K, \lambda, n) = E\{Corr(\psi_1, \psi_2))\}, \tag{4}$$

where ψ_1 and ψ_2 are two clustering functions obtained by applying $\Psi(\cdot; K, \lambda)$ to Y_1^n and Y_2^n respectively, with Y_1^n and Y_2^n being two random samples of n families from $f(Y_0)$.

Actually, following our notation, the distance between two clustering functions was defined as $Dist(\psi_1, \psi_2) = P(I_1 + I_2 = 1)$ in Wang (2010), and its expectation was defined as

instability, $Instab(\Psi; K, \lambda, n) = E\{Dist(\psi_1, \psi_2)\}$. Clearly, $Corr(\psi_1, \psi_2)$ in Definition 1 is closely related to $Dist(\psi_1, \psi_2)$ in that $P\{I_1 = I_2 = 1\} = 1 - P(I_1 + I_2 = 1) - P(I_1 = I_2 = 0)$. On the other hand, it differs from $Dist(\psi_1, \psi_2)$ in two aspects. First, $Corr(\psi_1, \psi_2)$ excludes $P(I_1 = I_2 = 0)$ in measuring the agreement between ψ_1 and ψ_2. Second, $Corr(\psi_1, \psi_2)$ incorporates the standardization over the concordance frequency between ψ_1 and ψ_2. Therefore, $Instab(\Psi; K, \lambda, n)$ tends to underestimate the instability of a clustering algorithm especially when $P(I_1 = I_2 = 0)$ is large. For example, when the number of clusters is equal to the number of subjects, all clustering algorithms yield the same clustering function that maps each subject to its own cluster. For these reasons, throughout this manuscript, we focus on the clustering stability based on $Corr(\psi_1, \psi_2)$ and $Stab(\Psi, K, \lambda, n)$, although all the results can be extended to $Dist(\psi_1, \psi_2)$ and $Instab(\Psi; K, \lambda, n)$.

4 Selection of tuning parameters

4.1 Cross-validation procedure

To estimate the clustering stability for the penalized cluster method in (1), we develop a modified cross validation procedure. The key idea is to split the data into two training sets and one validation set, where the two training sets are used to construct two clustering functions via the same clustering algorithm, and the clustering stability is estimated as the correlation between the two clusterings measured on the validation set. Furthermore, to reduce the estimation variability due to splitting randomness, multiple data splittings can be performed. As discussed in Yang (2007), there are two ways to perform multiple data splittings; one is called cross-validation with voting (CV_v) and the other is called cross-validation with averaging (CV_a). Since CV_v and CV_a are of minor difference, only the CV_a procedure is described in the following.

Consider a set of candidate algorithms $\{\Psi(\cdot; K, \lambda) : K = 2, \cdots, K_{\max}, \lambda \geq 0\}$, where

$K_{\max}$ is a fixed constant specifying the largest possible number of clusters in comparison.

CV$_a$ Algorithm

Step 1. Permute data $(Y_1, \cdots, Y_n)$ into $(Y_1^c, \cdots, Y_n^c)$, and then split them into three parts, $Y_1^{m,c} = (Y_1^c, \cdots, Y_m^c)$, $Y_2^{m,c} = (Y_{m+1}^c, \cdots, Y_{2m}^c)$, and $Y_3^{(n-2m),c} = (Y_{2m+1}^c, \cdots, Y_n^c)$. Denote the corresponding family sizes as n_i^c, $i = 1, \cdots, n$, and the number of subjects in the validation set is $l^c = \sum_{i=2m+1}^{n} n_i^c$.

Step 2. Let $\mathbf{u}(\Psi, K, \lambda, Y_1^{m,c})$ and $\mathbf{u}(\Psi, K, \lambda, Y_2^{m,c})$ be two $l^c(l^c-1)/2$ vectors with components being $I\{\psi_1(\mathbf{y}_{ij}^c) = \psi_1(\mathbf{y}_{i'j'}^c)\}$ and $I\{\psi_2(\mathbf{y}_{ij}^c) = \psi_2(\mathbf{y}_{i'j'}^c)\}$ respectively, where $\psi_k(\cdot) = \Psi(Y_k^{m,c}; K, \lambda), k = 1, 2$, and $(\mathbf{y}_{ij}^c, \mathbf{y}_{i'j'}^c)$ are pairs contained in the validation set. Denote the sample correlation between $\mathbf{u}(\Psi, K, \lambda, Y_1^{m,c})$ and $\mathbf{u}(\Psi, K, \lambda, Y_2^{m,c})$ as $\widehat{s}^c(\Psi, K, \lambda, m)$.

Step 3. Repeat the above two steps for C times. Let $\widehat{s}(\Psi, K, \lambda, m) = \sum_{c=1}^{C} \widehat{s}^c(\Psi, K, \lambda, m)/C$. Then find $(\widehat{K}, \widehat{\lambda}) = \arg\max \widehat{s}(\Psi, K, \lambda, m)$.

In the CV$_a$ Algorithm, two tuning parameters (K, λ) are considered simultaneously, which imposes significant computation burden. Based on our limited numerical experience, it seems that the penalized cluster analysis is much more sensitive to the selection of K than λ. Therefore, the following algorithm can be used to expedite the selection procedure.

Fast CV$_a$ Algorithm

Step 1. Given $\lambda = 0$, find $\widehat{K} = \arg\max_K \widehat{s}(\Psi, K, 0, m)$.

Step 2. Given $\widehat{K}$, find $\widehat{\lambda} = \arg\max_\lambda \widehat{s}(\Psi, \widehat{K}, \lambda, m)$.

4.2 Selection consistency

We establish an asymptotic theory regarding the selection consistency of the proposed cross-validation procedures. Let $K_0 \in \{2, \cdots, K_{\max}\}$ be the "true" number of clusters given that $\lambda = 0$; that is, the most appropriate number of clusters in the sense of clustering stability. In proving the consistency of $\widehat{K}$, we fix $\lambda = 0$ and exclude λ in all the expressions. The proof can be generated to the case where λ is fixed as any non-zero value.

To discriminate the candidate clusterings, a preference of K_0 over its competitors needs to be specified. Two assumptions are made as follows.

Assumption 1 Assume that $Corr(\Psi(\cdot; K, Y_1^{m,c}), \Psi(\cdot; K, Y_2^{m,c}))$ converges to one exactly at rate $r_{m,K}$ in probability as $m \to \infty$.

Assumption 2 For any $\epsilon > 0$, there exists $\delta > 0$ such that when m is sufficient large,

$$P\left(\frac{1 - Corr(\Psi(\cdot; K, Y_1^{m,c}), \Psi(\cdot; K, Y_2^{m,c}))}{1 - Corr(\Psi(\cdot; K_0, Y_1^{m,c}), \Psi(\cdot; K_0, Y_2^{m,c}))} > 1 + \delta\right) > 1 - \epsilon, \forall K \neq K_0.$$

Assumptions 1 and 2 are almost the same as the ones in Wang (2010). The exact convergence is defined in the sense of Definition 2 in Yang (2007). The selection consistency is stated in Theorem 1. For simplicity, Theorem 1 is presented for the case with only one splitting, which can be easily generalized to CV_a and CV_v.

Theorem 1 *For a single splitting, $Y_1^{m,c}$, $Y_2^{m,c}$, and $Y_3^{(n-2m),c}$, under Assumptions 1 and 2, we have $P(\widehat{K} = K_0) \to 1$, as long as $m \to \infty$ and $(n - 2m) \min_{K \neq K_0} r_{m,K}^2 \to \infty$.*

The proof of Theorem 1 is in Appendix. To understand the result, examine the regular case where $r_{m,K} = m^{-1/2}$. In this case, the requirement of consistency becomes $(n-2m)/n \to 1$, which agrees with the result for linear regression in Shao (1993).

Now we turn to the consistency of $\widehat{\lambda}$ for a given K. For simplicity, we fix K and remove K in all the expressions. Let $\lambda_m^* = \arg\max_\lambda Stab(\Psi, \lambda, m)$. We make the following two assumptions that are parallel to Assumptions 1 and 2.

Assumption 3 Assume that given λ, $Corr(\Psi(\cdot; \lambda, Y_1^{m,c}), \Psi(\cdot; \lambda, Y_2^{m,c}))$ converges to one exactly at rate $r_{m,\lambda}$ in probability as $m \to \infty$.

Assumption 4 Assume that for any series λ_m satisfying $\lambda_m / \lambda_m^* \to a \neq 1$ and any $\epsilon > 0$, there exists δ such as when m is sufficient large,

$$P\left(\frac{1 - Corr(\Psi(\cdot; \lambda_m, Y_1^{m,c}), \Psi(\cdot; \lambda_m, Y_2^{m,c}))}{1 - Corr(\Psi(\cdot; \lambda_m^*, Y_1^{m,c}), \Psi(\cdot; \lambda_m^*, Y_2^{m,c}))} > 1 + \delta\right) > 1 - \epsilon.$$

Theorem 2 *Suppose that Assumptions 3 and 4 hold. Given a single splitting,* $Y_1^{m,c}$, $Y_2^{m,c}$, *and* $Y_3^{(n-2m),c}$, *for any* $0 < \sigma < 1$, $\widehat{s}_c(\Psi, \lambda, m)$ *has a local maximum point* $\widehat{\lambda}$ *such that*

$$P((1-\sigma)\lambda_m^* < \widehat{\lambda} < (1+\sigma)\lambda_m^*) \to 1,$$

as long as $m \to \infty$ *and* $(n-2m)r^2_{m,(1\pm\sigma)\lambda_m^*} \to \infty$*; that is,* $\widehat{\lambda}/\lambda_m^* \to 1$ *in probability.*

The concept of local consistency in the above theorem is similar to the one in Speckman (1985, Theorem 4.3), where the selection of tuning parameter in smoothing splines was considered. For selecting tuning parameters in regularized linear regressions and smoothing splines, stronger asymptotic results can be derived, such as the weak consistency in Craven and Wahba (1979) and the strong consistency in Li (1986). However, more efforts must be taken to derive the property of weak consistency or strong consistency for our procedure.

5 Application to the asthma data

To examine the effectiveness of the procedure, we analyze the asthma data as described in Section 1. In Figure 2, the results from four modifications of the CV_a procedure given $\lambda = 0$ are reported: (a) based on Definition 1 with splitting over subjects, (b) based on Definition 1 with splitting over families, (c) based on Definition 2 with splitting over subjects, and (4) based on Definition 2 with splitting over families. Here K-means algorithm is applied, $K_{\max}$ is specified as 10, splitting ratio is taken as 1/3, and 50 splittings are generated.

Insert Figure 1 about here

First, from Figures 1(a) and 1(b), we see that the CV_a procedures based on instability in Definition 1 do not work well for the asthma data. Because the estimated clustering instabilities decrease significantly with the number of clusters after $K = 3$, the procedures based on instability select the number of clusters as $K_{\max}$, which is pre-specified as 10. Second, from Figures 2(c) and 2(d), we see that both the CV_a procedures with splitting over subjects and with splitting over families select the number of clusters as $\widehat{K} = 2$. This is

different from the result in Reilly *et al.* (2007), where they considered four clusters without any model selection process.

To demonstrate the influence of adding λ in the penalized cluster analysis, we continue to analyze the asthma data. Because the kinship information is not available to us, let $F_{ii'}$ be one if the two subjects are in the same family and zero otherwise. Instead of using K-medoids algorithm (Kaufman and Rousseeuw, 1990), we continue to use K-means algorithm. The results for selecting λ are reported in Figure 3: (a) $K = 2$ with splitting over subjects, (b) $K = 2$ with splitting over families, (c) $K = 3$ with splitting over subjects, and (4) $K = 3$ with splitting over families. Though the optimal K is selected as 2, $K = 3$ is included for illustration. Moreover, λ is taken from 0 to 1 by step 0.01, splitting ratio is taken as 1/3, and 50 splittings are generated.

Insert Figure 2 about here

From Figure 2, we see that each curve of estimated clustering stability increases with λ, achieves a maximum, and then decreases to a small value quickly. Given $K = 2$, with splitting over subjects, the curve achieves the maximum at $\widehat{\lambda} = 0.51$, while with splitting over families, $\widehat{\lambda} = 0.41$. Given $K = 3$, $\widehat{\lambda} = 0.70$ and $\widehat{\lambda} = 0.27$, respectively, with splitting over subjects and with splitting over families.

Since K is selected as $\widehat{K} = 2$ and splitting over families seems more appropriate, in Tables 1 and 2, more details are presented for the setting where K and λ are specified as 2 and 0.41.

Insert Table 1 about here

Table 1 presents the cluster means from the regular cluster analysis with $\lambda = 0$ and from the penalized cluster analysis with $\lambda = 0.41$. Cluster 1 includes subjects with high FEFM and cluster 2 includes subjects with low FEV1 and FF25. In addition, to understand the influence of the penalty term with $\lambda = 0.41$, as an example, we present the cluster ID's for

those eight members in the family whose famID='33' in Table 2. If $\lambda = 0$, five members are assigned to cluster 2 and three members are assigned to cluster 1. If $\lambda = 0.41$, seven members are assigned to cluster 2 and only one member are assigned to cluster 1. It suggests that the penalty term in the penalize cluster analysis makes the members in the same family be more concordant.

Insert Table 2 about here

6 Discussion

We propose the penalized cluster analysis for analyzing family data, along with methods to select appropriate tuning parameters. The former is straightforward, but the latter is arguable, because clustering is an unsupervised learning problem. For a supervised problem, the prediction error serves as a universal criterion for model selection. However, for an unsupervised problem, it is hard to find a criterion that pleases everyone. Here the clustering stability is defined and then is used to select the tuning parameters in the penalized cluster analysis. It seems it works for those real examples, including the asthma example.

First of all, we should point out that we are not introducing any new clustering procedure. We just propose an idea of adding a penalty term in the distance defined in (1) and (2). This type of penalty only works for clustered data such as family data and panel data.

We should also point out that the true number of clusters is not well defined in the literature. When the data come from a mixture of well-separated distributions the number of components serves naturally as the true number of clusters, while it is not so clear when the components are overlapped or the data are even not from a mixture distribution. The clustering stability and Assumption 2 provide an alternative way of defining the true number of clusters, which, however, can be different from the number of components when the data comes from a mixture distribution with overlapped components; see Ben-David *et al.* (2006)

and Wang (2010) for more discussion. Yet it is unclear which one is more reasonable in such situations due to the lack of objective definition of a proper clustering. Therefore, like many other available criteria for cluster analysis, we can only claim that the clustering stability can be a useful criterion for assessing the goodness of any clustering algorithm.

Finally, we should be honesty that we only prove the local consistency of $\widehat{\lambda}$. It is challenging to prove the global consistency (it is even not clear what consistency means for an unsupervised problem), and probably more assumptions should be made. But it seems that, as the most popular model selection procedure in literature, the cross-validation procedure is trustworthy.

Acknowledgement

We thank Professor Cavan Reilly for providing the asthma dataset. The asthma dataset was originally collected as part of the Collaborative Study on the Genetics of Asthma sponsored by the National Heart, Lung, and Blood Institute.

Appendix

Proof of Theorem 1. For a given splitting, $Y_1^{m,c}$, $Y_2^{m,c}$, and $Y_3^{(n-2m),c}$, let $l^c = \sum_{i=2m+1}^{n} n_i^c$ and $L = l^c(l^c-1)/2$, and let $\mathbf{u}(\Psi, K, Y_j^{m,c}) = (u_1^{(j)}(K), \cdots, u_L^{(j)}(K))^T$, for $j = 1, 2$. Let $\mu^{(j)}(K)$ and $\sigma^{(j)}(K)$ be the conditional expectation and standard deviation of $u_1^{(j)}(K)$ given $Y_1^{m,c}$ and $Y_2^{m,c}$, for $j = 1, 2$. Then,

$$\widehat{s}^c(\Psi, K, m) = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{u_l^{(1)}(K) - \mu^{(1)}(K)}{\sigma^{(1)}(K)}\right)\left(\frac{u_l^{(2)}(K) - \mu^{(2)}(K)}{\sigma^{(2)}(K)}\right) + O_p(\frac{1}{n-2m}),$$

noting that $\sqrt{L}$ is of the same order of $n - 2m$ as $E\{n_1\}$ is finite.

Denote the summands in the above formula as $V_l(K)$. Let $W_l(K) = V_l(K) - V_l(K_0)$ and $\Delta_K = -E(W_l|Y_1^{m,c}, Y_2^{m,c})$. Then, for $K \neq K_0$,

$$\begin{aligned}
& P(\widehat{s}^c(\Psi, K, m) \geq \widehat{s}^c(\Psi, K_0, m)|Y_1^{m,c}, Y_2^{m,c}) \\
= \; & P\left(\sum W_l/L + O_p(1/(n-2m)) \geq 0|Y_1^{m,c}, Y_2^{m,c}\right) \\
= \; & P\left(\sum (W_l - EW_l) + O_p(n-2m) \geq L\Delta_K|Y_1^{m,c}, Y_2^{m,c})\right) \\
\leq \; & P\left(\sum (W_l - EW_l) \geq L\Delta_K/2|Y_1^{m,c}, Y_2^{m,c}\right) + P\left(O_p(n-2m) \geq L\Delta_K/2|Y_1^{m,c}, Y_2^{m,c}\right).
\end{aligned}$$

For any $\epsilon > 0$, let A_K be the set in Assumption 2 such that $P(A_K) > 1-\epsilon$. By the Bernstein's inequality for U-statistics (e.g., Janson, 2004), on A_k,

$$P\left(\sum (W_l - EW_l) \geq L\Delta_K/2|Y_1^{m,c}, Y_2^{m,c}\right) \leq \exp(-\lfloor (n-2m)/2 \rfloor \Delta_K^2/8).$$

Therefore, if $m \to \infty$ and $(n-2m)r_{m,K}^2 \to \infty$, we have $P(\widehat{s}^c(\Psi, K, m) \geq \widehat{s}^c(\Psi, K_0, m)) \to 0$. Noting $P(\widehat{K} \neq K_0) \leq \sum_{K \neq K_0} P(\widehat{s}^c(\Psi, K, m) \geq \widehat{s}^c(\Psi, K_0, m))$, Theorem 1 is proved. ■

Proof of Theorem 2. For any $0 < \sigma < 1$, if we focus on three values of λ, $(1-\sigma)\lambda_m^*$, λ_m^*, and $(1+\sigma)\lambda_m^*$, following the same arguments in proving Theorem 1, we can show that

$$P\left(\widehat{s}^c(\Psi, (1 \pm \sigma)\lambda_m^*, m) \geq \widehat{s}^c(\Psi, \lambda_m^*, m)\right) \to 0,$$

as long as $m \to \infty$ and $(n-2m)r_{m,(1\pm\sigma)\lambda_m^*}^2 \to \infty$. Therefore, the probability that $\widehat{s}^c(\psi, \lambda, m)$ has a local maximum point $\widehat{\lambda}$ in the interval $((1+\sigma)\lambda_m^*, (1+\sigma)\lambda_m^*)$ goes to one, noting that the function $\widehat{s}^c(\psi, \lambda, m)$ with respect to λ is a step function. ■

References

[1] Ben-David, S., von Luxburg, U., and Pal, D. (2006). A sober look at stability of clustering, *19th Annual Conference on Learning Theory (COLT 2006)*.

[2] Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data, *Pacific Symposium on Biocomputing* **7**: 6-17.

[3] Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics - Simulation and Computation* **3**: 1-27.

[4] Craven, P. and Wahba, G. (1979), Smoothing Noisy Data With Spline Functions, *Numerische Mathematik* **31**: 377-403.

[5] The Collaborative Study on the Genetics of Asthma (CSGA) (1997). A genome-wide search for asthma susceptibility loci in ethnically diverse populations, *Nature Genetics* **15**: 389-392.

[6] Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierachical clusterings. *Journal of the American Statistical Association* **78**: 553-584.

[7] Hartigan, J. A. (1975). *Clustering Algorithms*, Wiley, New York.

[8] Janson, S. (2004). Large deviations for sums of partly dependent random variable. *Random Structures and Algorithms* **24**: 234-248.

[9] Johnson, S. C. (1967). Hierarchical Clustering Schemes, *Psychometrika* **2**: 241-254.

[10] Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An introduction to Cluter Analysis*, Wiley, New York.

[11] Khoury, M. J., Beaty, T. H., and Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology*, Oxford University Press, Inc.

[12] Krzanowski, W. J. and Lai, Y. T. (1985). A criterion for determining the number of clusters in a data set, *Biometrics* **44**: 23-34.

[13] Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*, Springer-Verlag, New York, 1997.

[14] Lange, T., Roth, V., Braun, M., and Buhmann, J. (2004). Stability-based validation of clustering solutions, *Neural Computation* **16**: 1299-1323.

[15] Li, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to splilne smoothing, *The Annals of Statistics* **14**: 1101-1112.

[16] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, **1**: 281-297

[17] Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural information Processing Systerms*. MIT Press, Cambridge.

[18] Reilly, C., Miller, M. B., Liu, Y., Oetting, W. S., King, R., and Blumenthal, M. (2007). Linkage analysis of a cluster-based quantitative phenotypes constructed from pulmonary function test data in 27 multigenerational families with multiple asthmatic members, *Human Heredity* **64**: 136-145.

[19] Shao, J. (1993). Linear model selection by cross-validation. *Journal of American Statistical Association* **88**, 486-494.

[20] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**: 888-905.

[21] Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models, *The Annals of Statistics* **13**: 970-983.

[22] Sugar, C. and James, G. (2003). Finding the number of clusters in a data set: an imformation theoretic approach, *Journal of American Statistical Association* **98**, 750-763.

[23] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, *Journal of Royal Statistical Society, Series B*, **63**: 511-528.

[24] Wang, J. (2010). Consistent selection of the number of clusters via cross-validation, *In Press, Biometrika.*

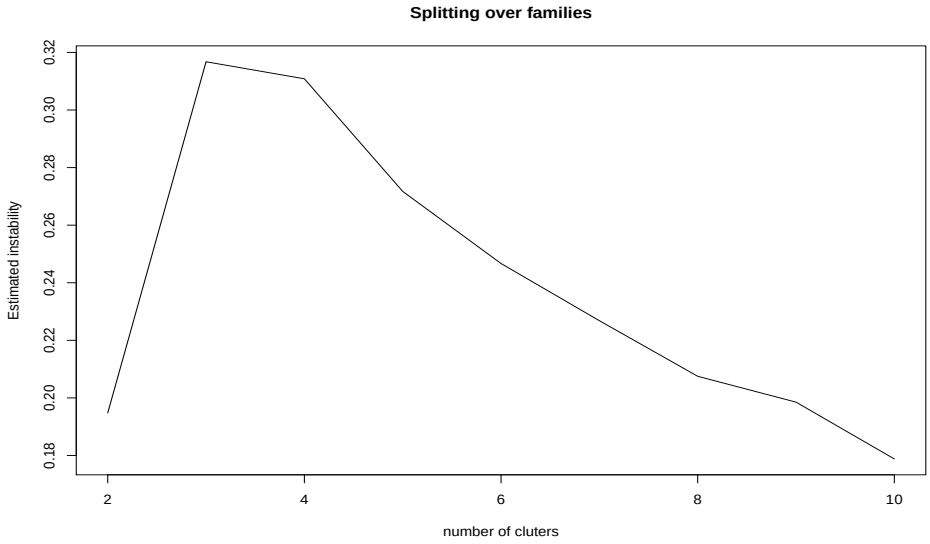
[25] Yang, Y. (2007). Consistency of cross validation for comparing regression procedures, *Annals of Statistics* **35**: 2450-2473.

Table 1: Cluster means given $K = 2$

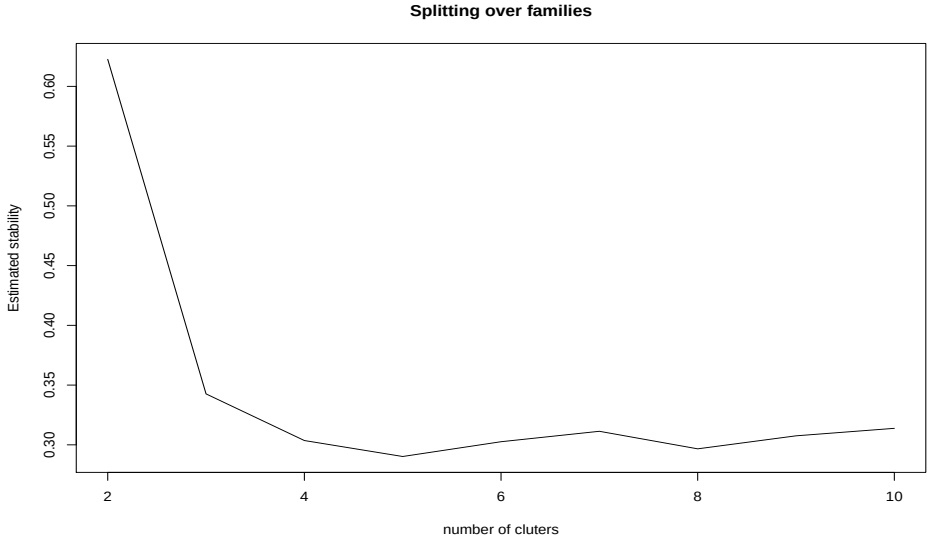
	cluster	FEV1	FVC	FEFM	FF25
$\lambda = 0$	1	-0.0060	0.0012	0.1231	-0.0357
	2	-0.1951	-0.0582	-0.0495	-0.5505
$\lambda = 0.41$	1	-0.0059	-0.0019	0.1257	-0.0305
	2	-0.1904	-0.0523	-0.0486	-0.5447

Table 2: One example − family with famID='33'

	member ID	1	2	3	4	5	6	7	8
$\lambda = 0$	cluster ID	2	2	1	1	2	2	1	2
$\lambda = 0.41$	cluster ID	2	2	1	2	2	2	2	2

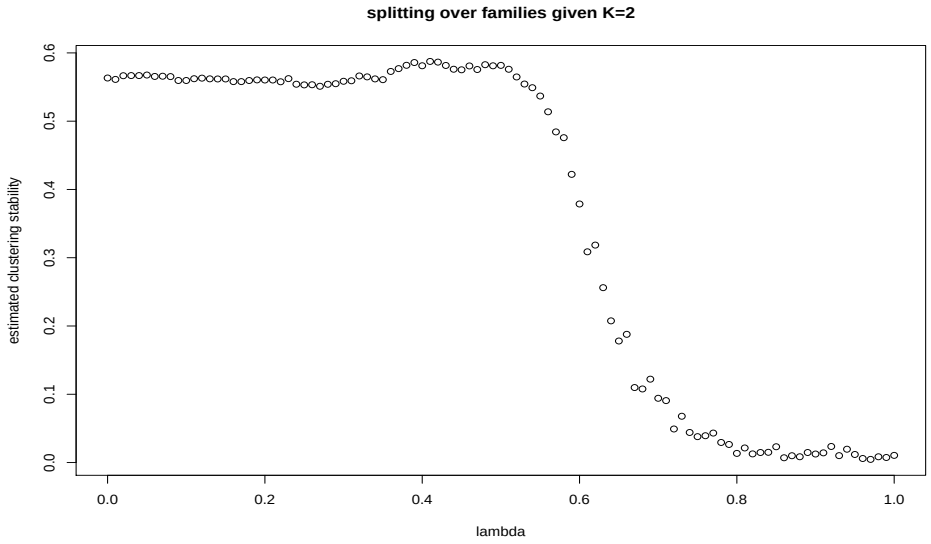


(a) Based on the instability defined in Wang (2010) ($\widehat{K} = 10$)

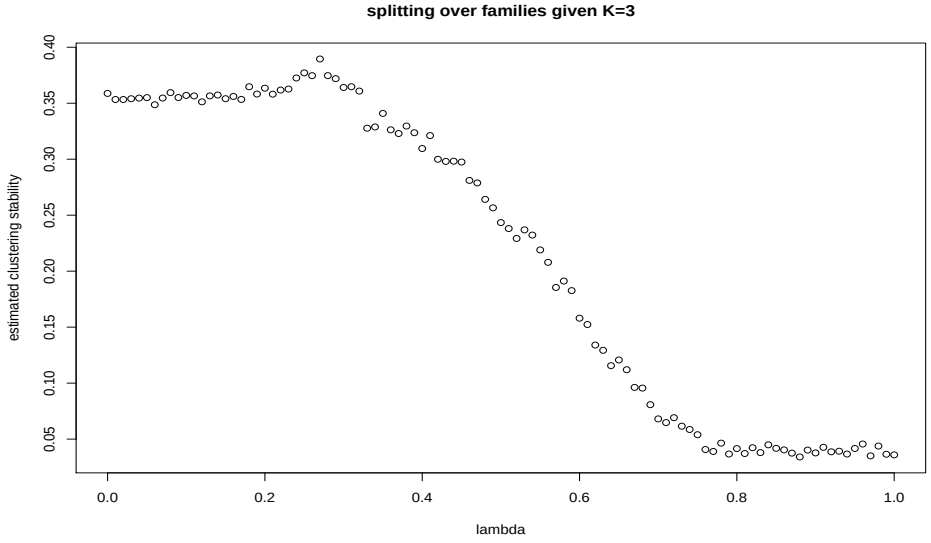


(b) Based on our definition of stability ($\widehat{K} = 2$)

Figure 1: Selection of K for the asthma data



(a) Estimated stability achieves maximum at $\widehat{\lambda} = 0.41$



(b) Estimated stability achieves maximum at $\widehat{\lambda} = 0.27$

Figure 2: Selection of λ for the asthma data