

Clustering Confidence Sets

Nicoleta Serban¹

Industrial Systems and Engineering School

Georgia Institute of Technology

We propose a method for clustering a large set of observed objects with different noise levels based on their confidence set estimates rather than their point estimates. The minimal and maximal distances between confidence sets provide confidence intervals for the true distances between objects. The upper bounds of these confidence intervals are used to minimize the within clustering variability and the lower bounds are used to maximize the between clustering variability. The underlying clustering algorithm is the single-linkage tree based on the matrix of the upper bounds of the distance confidence intervals. We illustrate our technique by clustering a large number of short curves within a synthetic example.

Key words and phrases: single-linkage tree, gap sequence, clustering error rate, simultaneous confidence sets.

1 Introduction

In this paper we introduce a technique for clustering a large set of objects observed with different noise levels. Current approaches cluster their point estimates. We propose to use simultaneous confidence sets rather than their point estimates. The primary advantages of using confidence sets over point

¹The author is grateful to Larry Wasserman for his research support. The author also thanks Alexander Gray and Ping Zhang for reading this manuscript and their input.

estimates are that we obtain a way to maximize the difference between the between and within variability and we may be able to make some probabilistic statements about our confidence of the estimated cluster membership.

In our technique, the objects can be d -dimensional points or more complex objects such as dependent measurements. The general framework of our technique is as follows:

1. For a set of N observed objects $\mathcal{C}_N = \{o_1, \dots, o_N\}$, with N large, estimate simultaneous (uniform) confidence sets \mathbb{B}_i , $i = 1, \dots, N$.
2. Choose an appropriate measure of similarity for the objects to be clustered and measure the minimal and maximal distance between confidence sets: $L = \{l_{ij}\}_{i=1, \dots, n, j=1, \dots, n}$ and $U = \{u_{ij}\}_{i=1, \dots, n, j=1, \dots, n}$ where

$$l_{ij} = \min_{\theta_i \in \mathbb{B}_i, \theta_j \in \mathbb{B}_j} d(\theta_i, \theta_j)$$

$$u_{ij} = \max_{\theta_i \in \mathbb{B}_i, \theta_j \in \mathbb{B}_j} d(\theta_i, \theta_j)$$

3. Assign two objects o_i and o_j to the same cluster if u_{ij} is small and assign the two objects to different clusters if l_{ij} large.

We exemplify our general technique for clustering curves, and apply it to a synthetic example. We estimate confidence sets for the observed curves as introduced in Beran and Dümbgen 1998 or using a multiscale approach.

Extensive literature is on clustering multiple curves. For a reference list on regularization and filtering clustering methods see James, Sugar 2003. Generally, clustering techniques can provide hard, or soft boundaries between clusters. The current soft clustering methods are model-based (see

James, Sugar 2003 for a filtering method, Fraley, Raftery 2002; Wakefield, Zhou, Self 2002; Chudova, Gaffney, Mjolsness, Smyth 2003 for regularization methods). Hard clustering methods are popular mostly because they are not computational intensive (Ben-Dorr, Shamir, Yakhimi 1999, Hastie et al 2000 for regularization methods and Bar-Joseph, Gerber, Gifford and Jaakkola 2002; Serban, Wasserman 2005 for filtering methods). One downside of hard clustering is that we need to estimate the number of clusters. Methods for estimating the number of clusters have been developed in Tibshirani, Walther, Hastie (2000), Sugar, James (2003) (see also the references therein).

2 Clustering Curves

The general setting is:

$$Y_{ij} = f_i(t_j) + \sigma_i \epsilon_{ij}, \quad i = 1, \dots, N, \quad j = 1, \dots, m. \quad (1)$$

assuming only $\mathbb{E}(\epsilon_{ij}) = 0$. Thus, Y_{ij} is the j^{th} observation on the i^{th} curve where N is typically much larger than m .

We assume that the curves f_i belong to a Sobolev space $\mathcal{F} \equiv \mathcal{F}_\beta(c)$ of unknown order β and unknown radius c :

$$\left\{ f(x) = \sum_{j=1}^{\infty} \beta_j \psi_j(x) : \sum_{j=1}^{\infty} \beta_j^2 j^{2\beta} \leq c^2 \right\}$$

where ψ_1, ψ_2, \dots is an orthonormal basis for L_2 .

We also assume that the curves are not simply flat over time. If the

set of curves contains a large number of constant curves, a screening step is necessary. For a screening step in the context of our model see Serban and Wasserman (2005).

2.1 Transforming the Data

Without loss of generality, assume that all time points lie in $[0, 1]$. We transform the data into the Fourier domain as follows. Let

$$\phi_0(t) \equiv 1, \quad \text{and} \quad \phi_j(t) = \sqrt{2} \cos(j\pi t), j \geq 1$$

denote the cosine basis. Define the $m \times (k+1)$ matrix $\Phi = \{\phi_j(t_i)\}_{i=1, \dots, m; j=0, \dots, k}$ and perform a Gram-Schmidt orthogonalization on the columns of Φ to make the columns orthogonal. Denote the new matrix by Ψ . Let $\hat{\theta}_i = (\hat{\theta}_{i1}, \dots, \hat{\theta}_{im})$

$$\hat{\theta}_{ir} = \frac{1}{m} \sum_{j=1}^m \psi_{rj} Y_{ij}.$$

The choice of basis does not matter much. We can use a basis as soon as we can construct uniform confidence sets. Uniform confidence sets for wavelet basis are derived in Genovese and Wasserman (2004). But the theory used to get nonparametric confidence sets has not been worked out for some other basis such as B-splines.

2.2 Confidence Set for f_i

The function $\hat{f}_i^J(t) = \sum_{j=0}^J \hat{\theta}_{ij} \psi_{jt}$ is the smoothed version of the i^{th} profile, where J is smoothing parameter. We estimate a global smoothing parameter

as in Serban and Wasserman 2005. The estimated smoothing parameter minimizes a total regret function.

We use the method in Beran and Dümbgen (1998) for constructing a confidence set \mathbb{B}_i for f_i . Fix $\alpha > 0$ and let

$$\mathbb{B}_i = \left\{ (\theta_{i1}, \dots, \theta_{im}) : \sum_{j=1}^m (\theta_{ij} - \hat{\theta}_{ij})^2 \leq s_i^2 \right\} \quad (2)$$

where

$$s_i^2 = \frac{z_{\frac{\alpha}{N}} \hat{\tau}_i}{\sqrt{m}} + \hat{R}_i,$$

z_{α} is the α quantile of the standard normal and $\hat{\tau}_i$ is given in the Appendix. The corresponding confidence set for f_i is $\{\sum_{j=1}^m \theta_{ij} \psi_j(x) : \theta \in \mathbb{B}_i\}$. For notational convenience, the confidence set for f_i will also be denoted by \mathbb{B}_i .

The confidence sets are asymptotically uniform over the Sobolev space and uniform over curves. But the uniformity over both the Sobolev space and over curves will result in large confidence sets. We cannot relax the uniformity over curves since we observe the curves simultaneously. But we can relax the uniformity over the Sobolev space using a multiscale approach.

Rather than searching for an optimal amount of smoothing, we can instead consider all values of J simultaneously and choose the one that leads to the most efficacious clustering. More precisely, we consider all the estimates \hat{f}^J or $1 \leq J \leq J_m$ where $J_m = o(m)$. We recommend the value $J_m = \sqrt{m}$. This leads to confidence sets for the curves of size $O(m^{-1/4})$ which is the smallest possible in a nonparametric sense (Li, 1989). Note

that \widehat{f}_i^J is actually estimating

$$f_i^J(t) = \sum_{j=0}^J \theta_{ij} \phi_j(t).$$

We can think of $f_i^J(t)$ as the smoothed version of the true curve. When J is small, we give up high resolution information about f_i but we can estimate $f_i^J(t)$ accurately. We will probably not discover many clusters when J is small since there is not much shape information in f_i^J . As we increase J we can potentially discover more shape information leading to more refined clusters. However, the confidence sets for f_i^J get larger as J increases.

The confidence set for the multiscale method is slightly different. Here we need a confidence set for $(\theta_{i1}, \dots, \theta_{iJ})$ which is somewhat simpler. Since, $\widehat{\theta}_{ij} \approx N(\theta_{ij}, \sigma_i^2/m)$, we have that

$$\sum_{j=1}^J (\theta_{ij} - \widehat{\theta}_{ij})^2 \approx \frac{\sigma_i^2}{m} \chi_J^2.$$

Hence,

$$\mathbb{B}_i^J = \left\{ (\theta_1, \dots, \theta_J) : \sum_{j=1}^J (\theta_j - \widehat{\theta}_j)^2 \leq \frac{\widehat{\sigma}_i^2 \chi_{J, \alpha'}^2}{m} \right\} \quad (3)$$

is an approximate $1 - \alpha'$ confidence set for $(\theta_{i1}, \dots, \theta_{iJ})$. We take $\alpha' = \alpha/(NJ_m)$ to ensure that the coverage is uniform over curves i and scales J :

$$\inf_{1 \leq J \leq m} \mathbb{P} \left(f_i \in \mathbb{B}_i^J \text{ for all } i = 1, \dots, N \right) \gtrsim 1 - \alpha.$$

We can also use the confidence balls \mathbb{B}_i to screen out constant curves by

removing curve i if $(s, 0, \dots) \in \mathbb{B}_i$.

2.3 Distance Estimation

We want to cluster curves by shape similarity. This suggests using the correlation coefficient to measure the similarity between two curves. Let $f = \sum_{j=0}^{\infty} a_j \psi_j$ and $g = \sum_{j=0}^{\infty} b_j \psi_j$ be the Fourier decompositions for two functions. Then the correlation between the two curves can be expressed as a function of the curve coefficients:

$$\rho(f, g) = \frac{\sum_{j=1}^{\infty} a_j b_j}{\sqrt{\sum_{j=1}^{\infty} a_j^2} \sqrt{\sum_{j=1}^{\infty} b_j^2}} = \frac{\langle \tilde{a}, \tilde{b} \rangle}{\|\tilde{a}\| \|\tilde{b}\|} = 1 - \frac{\|\frac{\tilde{a}}{\|\tilde{a}\|} - \frac{\tilde{b}}{\|\tilde{b}\|}\|^2}{2}. \quad (4)$$

where $\tilde{a} = (a_1, \dots)$, $\tilde{b} = (b_1, \dots)$.

The result above applies to any decomposition based on an orthonormal basis, including wavelet basis for which we can derive nonparametric confidence sets.

If $\rho(f_1, f_2)$ is the distance between any two objects f_1 and f_2 (e.g. correlation when clustering by shape), define the maximal and minimal distances between the confidence sets of two curves:

$$MAX(\mathbb{B}_1, \mathbb{B}_2) = \sup_{f_1 \in \mathbb{B}_1, f_2 \in \mathbb{B}_2} \rho(f_1, f_2) \text{ and } MIN(\mathbb{B}_1, \mathbb{B}_2) = \inf_{f_1 \in \mathbb{B}_1, f_2 \in \mathbb{B}_2} \rho(f_1, f_2)$$

Estimating the confidence sets using the regret approach, we obtain the following result:

Theorem 1 *Let $\mathcal{F}_\beta(c)$ denote a Sobolev space of order β and radius c . Then, for any $\beta > 1/2$ and any $c > 0$,*

$$\liminf_{N \rightarrow \infty} \sup_{f_1, \dots, f_N \in \mathcal{F}_\beta(c)} \mathbb{P} \left(\text{MIN}(\mathbb{B}_i, \mathbb{B}_j) \leq d(f_i, f_j) \leq \text{MAX}(\mathbb{B}_i, \mathbb{B}_j), \forall i, j = 1, \dots, N \right) \geq 1 - \alpha.$$

Thus we estimate the distance between two curves with a simultaneous $(1-\alpha)$ confidence interval given by the distances between their confidence sets.

The proof of the theorem follows from the theorems in Beran and Dümbgen (1998). A similar result can be obtained for the multiscale approach.

The maximal and minimal measures between two confidence sets of two curves when the similarity measure between two curves is the correlation coefficient become:

$$\begin{aligned} \text{MAX}(\mathbb{B}_1, \mathbb{B}_2) &= \sup_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \left[1 - \frac{\langle \tilde{\theta}_1, \tilde{\theta}_2 \rangle}{\|\tilde{\theta}_1\| \|\tilde{\theta}_2\|} \right] = 1 - \inf_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \frac{\langle \tilde{\theta}_1, \tilde{\theta}_2 \rangle}{\|\tilde{\theta}_1\| \|\tilde{\theta}_2\|} \\ \text{MIN}(\mathbb{B}_1, \mathbb{B}_2) &= \inf_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \left[1 - \frac{\langle \tilde{\theta}_1, \tilde{\theta}_2 \rangle}{\|\tilde{\theta}_1\| \|\tilde{\theta}_2\|} \right] = 1 - \sup_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \frac{\langle \tilde{\theta}_1, \tilde{\theta}_2 \rangle}{\|\tilde{\theta}_1\| \|\tilde{\theta}_2\|} \end{aligned}$$

Both the minimal and the maximal measures can be computed when the measure of similarity between two curves is the correlation or the Euclidean distance. We show how to compute these distances in the Appendix.

Remark 2 *When the confidence set \mathbb{B} of a curve f contains the origin $(0, 0, \dots)$ of the space, the maximal distance between these curve and any other curve f_i is 2 ($\text{MAX}(\mathbb{B}, \mathbb{B}_i) = 2$), the maximum value the maximal distance can take. Also the minimal distance is 0 ($\text{MIN}(\mathbb{B}, \mathbb{B}_i) = 0$), the minimum value*

the minimal distance can take. Any clustering algorithm becomes more stable when we first remove those curves whose uncertainty is so high due to a high noise level that they cannot be clustered. (A curve f cannot be clustered when $MAX(\mathbb{B}, \mathbb{B}_i) = 2$ and $MIN(\mathbb{B}, \mathbb{B}_i) = 0$ for any $i = 1, \dots, N$.)

2.4 Clustering Algorithm

Next we want to use a clustering algorithm which requires only the distance matrix for the objects, i.e. the matrix of all pairwise distances. One such algorithm is the single-linkage tree.

We first apply the single-linkage tree algorithm to the maximal distance matrix $U = \{u_{ij} = MAX(\mathbb{B}_i, \mathbb{B}_j)\}$. If two objects/curves are close with respect to the maximal distance, then with high probability they are close with respect to the true distance. Since the algorithm assigns clusters using the maximal distances in D , we are ensured that with high probability the curves are correctly assigned in the same cluster. Therefore, the within clustering variability is minimized.

Second, we separate clusters based on the minimal distances: $L = \{l_{ij} = MIN(\mathbb{B}_i, \mathbb{B}_j)\}$. If two objects have large minimal distance, with high probability their true distance is large, or with high probability the curves are correctly assigned in different clusters. Therefore, the between clustering variability is maximized.

The clustering algorithm is as follows:

1. The single-linkage tree for N curves/objects \mathcal{O}_N and their maximal distance matrix $U = u_{ij}$ is constructed as follows. ($I(j)$, $j = 1, \dots, N$)

are the nodes and $G(j)$, $j = 1, \dots, N$ represent the links between the nodes in the tree.)

- (a) Find two curves/objects f_i and f_j such that their distance is the smallest in the distance matrix U . Assign $I(1) = i$ and $I(2) = j$, $G(1) = 0$ and $G(2) = u_{I(1),I(2)}$.
- (b) For $j = 3, \dots, N$, find $f_{I(j)}$ the object in $\mathcal{O}_N \setminus \{f_{I(1)}, \dots, f_{I(j-1)}\}$ that has the minimum distance based on the maximal distance matrix U to any of the objects $f_{I(1)}, \dots, f_{I(j-1)}$. $G(j)$ is this minimum distance.

2. We separate clusters as follows:

- (a) For $j = 1, \dots, N - 1$, compute the estimated within and between variabilities assuming $\{I(1), \dots, I(j)\}$ and $I(j + 1)$ form two different clusters:

$$\widehat{W}(j) = \frac{1}{j^2} \sum_{k=1}^j \sum_{l=1}^j u_{I(k),I(l)}$$

$$\widehat{B}(j) = \frac{2}{j} \sum_{k=1}^j l_{I(k),I(j+1)}$$

If $\widehat{W}(j) \approx \widehat{B}(j)$ then put $I(j+1)$ in the same cluster as $I(1), \dots, I(j)$. Continue until $\widehat{W}(j) \ll \widehat{B}(j)$. That is, $I(j + 1)$ is in different cluster than $\{I(1), \dots, I(j)\}$.

- (b) For K clusters in the sequence $I(1), \dots, I(j)$ ($\{I(1), \dots, I(j_1)\}$ the first cluster, $\{I(j_1 + 1), \dots, I(j_2)\}$ the second cluster, ..., $\{I(j_{K-1} + 1), \dots, I(j)\}$ the K -th cluster), assume $I(j + 1)$ forms a

$(K + 1)$ th cluster and compute the estimated within and between variabilities:

$$\widehat{W}(j, K) = \sum_{k=2}^K \frac{1}{j_k^2} \sum_{s,t=j_{k-1}+1}^{j_k} u_{I(s),I(t)}$$

$$\widehat{B}(j, K) = \frac{2}{j} \sum_{k=1}^j l_{I(k),I(j+1)} + \sum_{k,l=1}^K \frac{1}{j_k j_l} \sum_{s=j_{(k-1)}+1}^{j_k} \sum_{t=j_{(l-1)}+1}^{j_l} l_{I(s),I(t)}$$

If $\widehat{W}(j) \approx \widehat{B}(j)$ then put $I(j + 1)$ in the same cluster as $I(j_{K-1} + 1), \dots, I(j_K = j)$. Continue until $\widehat{W}(j) \ll \widehat{B}(j)$. That is, $I(j + 1)$ is in different cluster than $I(j_{K-1} + 1), \dots, I(j)$.

In the notation above, $I(j)$, $j = 1, \dots, n$ are the nodes and $G(j)$, $j = 1, \dots, n$ represent the links between the nodes in the tree. Hartigan (1975) calls G the *gap sequence*.

The statements above are supported by the next lemma, which follows from theorem 1.

Lemma 3 *Let \mathcal{C} be a clustering of K clusters: $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$. We define the estimated within and between clustering variabilities:*

$$\widehat{W}(\mathcal{C}) = \sum_{k=1}^K \frac{1}{n_k^2} \sum_{i,j \in \mathcal{C}_k} u_{ij}$$

$$\widehat{B}(\mathcal{C}) = \sum_{k=1, l=1}^K \frac{1}{n_k n_l} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_l} l_{ij}$$

and the true within and between clustering variabilities:

$$W(\mathcal{C}) = \sum_{k=1}^K \frac{1}{n_k^2} \sum_{i,j \in \mathcal{C}_k} d_{ij}$$

$$B(\mathcal{C}) = \sum_{k=1, l=1}^K \frac{1}{n_k n_l} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_l} d_{ij}$$

where d_{ij} is the true distance between f_i and f_j .

Then

$$\liminf_{N \rightarrow \infty} \sup_{f_1, \dots, f_N \in \mathcal{F}_\beta(c)} \mathbb{P} \left(W(\mathcal{C}) \leq \widehat{W}(\mathcal{C}) \leq \widehat{B}(\mathcal{C}) \leq B(\mathcal{C}) \right) \geq 1 - \alpha.$$

2.5 Clustering Error Rate

We evaluate our clustering technique with synthetic data. In order to compare the performance of different clustering techniques we use a measure for clustering error.

Let $T_{n,K}$ and $\widehat{T}_{n,K}$ denote the true clustering map and, respectively, the estimated clustering map:

$$T_{n,K}(f, g) = \begin{cases} 1 & \text{if } f \text{ and } g \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

The two clustering maps depend on the number of clusters.

The *clustering error rate* for K clusters is

$$\eta(K) = \frac{1}{\binom{N}{2}} \sum_{r < s} I \left(T_{n,K}(f_r, f_s) \neq \widehat{T}_{n,K}(f_r, f_s) \right) \quad (6)$$

This clustering error rate can also be expressed as

$$\eta = 1 - \mathcal{R}(T, \hat{T})$$

where \mathcal{R} is the Rand index (Rand, 1971).

3 Application - Synthetic Data

We generate data from 4 different curve shapes according to the regression model (1) with $m = 15$, $N = 400$ and $\sigma \in [1, 1.7]$ (at the latest time points the variance is slightly larger than at the early ones). The synthetic data are:

$$Y_{ij} = \mu_i + F_k(t_j) + N_m(0, (\sigma_i + \sigma(t_j))), \quad i = 1, \dots, 400, \quad j = 1, \dots, 15, \quad k = 1, 2, 3, 4$$

where $\mu_i \in [1, 20]$, $\sigma_i \in [1, 1.2]$ and $\sigma(t)$ is increasing with t (from 0 at $t = 0$ to .5 at $t = t_{15}$), and F_k for $k = 1, 2, 3, 4$ are four different patterns. The mean, 90th and 10th percentiles of the data in the four true clusters are in Figure 1.

3.1 Other Methods

We will consider in this section the performance of four different clustering algorithms. They are the model-based clustering ('mclust' R library) introduced in Yeung et al (2001), the filtering clustering technique introduced by Bar-Joseph et al (2001), hierarchical clustering of the estimated curves, and k -means clustering of the estimated curves.

Yeung et al (2001). It is a model-based clustering technique that estimates the number of clusters using the BIC approximation criteria. We apply this technique to the synthetic data presented above. The method assigns the four groups of curves to two clusters with the means, 10th and 90th percentiles presented in Figure 2.

Bar-Joseph et al (2001). The clustering strategy introduced by Bar-Joseph and colleagues is a filtering method that provide hard clustering membership. We apply this technique to the synthetic data presented above. We obtain the clustering assignment as shown in Figure 3. This method captures the main patterns in these simulated data when assigning four clusters. The misclustering error is about .1. However, the method requires the input of the number of clusters.

Hierarchical clustering. We apply the single-linkage clustering to the estimated curves. According to equation (4), clustering with Euclidean distance in the Fourier domain is equivalent with clustering with correlation measure in the functional domain. The resulting clusters are shown in Figure 4.

k -means clustering. Similarly, we apply k -means to the observed synthetic curves and their point estimates. The results are shown in Figure 5.

Remark 4 Except the model-based clustering method, all the other techniques require the input of the number of clusters. We assume 4 clusters, the number of true patterns. One has to bear in mind that both Yeung et al (2001) and Bar-Joseph et al (2001) are computational expensive since they are based on EM-type estimation. Also, EM may fail to provide accurate estimates when there are clusters with a few observations, when the number of compo-

nents/clusters is misspecified, or when there are outliers.

4 Discussion

In this article, we have presented a novel approach to clustering a large number of observed curves. The general strategy can be applied to other types of objects. The novelty of our technique consists of clustering based on the distances between confidence sets. Consequently, we obtain confidence interval estimates for the distances between curves. The upper limit of these confidence intervals can be used to assign curves in the same cluster and the lower limits can be used to assign curves in different clusters. We used the single-linkage tree to assign the cluster membership, but any other clustering algorithm can be used (e.g. the minimum spanning tree or k -ary clustering). One important advantage of the single-linkage tree is that we can infer the number of clusters from its links.

Removing curves/objects that cannot be clustered (curves whose confidence sets contain the origin) we obtain a more reliable clustering. There are 30 curves for the synthetic data that cannot be clustered. After removing these curves the misclustering rate becomes 0 for the hierarchical clustering of the estimated curves. The misclustering rates for different clustering techniques and different settings are presented in Table 1.

One other advantage of using our approach is that we can establish a better separation between clusters. Figure 6 displays the gap sequence of the clustering assignment based on the estimated distances using single-linkage tree algorithm. Figure 7 displays the gap sequence of the clustering

assignment based on the maximal distances $U = \{u_{ij}\}$ using single-linkage algorithm. In the same figure, we include the lower bound of the gap sequence (the minimal distances of the objects in the sequence). It is a clearer separation in the latter gap sequence than in the former. Further we clearly identify four clusters using the gap sequence of the maximal distance.

There are a few different paths to take from here. One challenge is to design a clustering algorithm that uses the information of the upper and lower limits simultaneously. Another challenge is to extend the main strategy to clustering algorithms that do not use the distance matrix information only (e.g. k -means). Generally, clustering confidence sets may allow us to make probabilistic statements about the clustering membership.

Appendix

Computing the maximal and minimal distances

In this section we compute the maximal distance when the similarity measure between two objects is the Euclidean distance or the correlation coefficient. Under Euclidean distance, the maximal and minimal distances between two confidence sets are:

$$MAX_E(\mathbb{B}_1, \mathbb{B}_2) = \sup_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \|\theta_1 - \theta_2\|^2$$

$$MIN_E(\mathbb{B}_1, \mathbb{B}_2) = \inf_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \|\theta_1 - \theta_2\|^2.$$

Under the correlation measure, the maximal and minimal distances between two sets are:

$$MAX_C(\mathbb{B}_1, \mathbb{B}_2) = \sup_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \frac{\langle \theta_1, \theta_2 \rangle}{\|\theta_1\| \|\theta_2\|}$$

$$MIN_C(\mathbb{B}_1, \mathbb{B}_2) = \inf_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \frac{\langle \theta_1, \theta_2 \rangle}{\|\theta_1\| \|\theta_2\|}$$

Solution for the Euclidean distance.

When the Euclidean distance is the measure between two objects, the two distances are rather easy to compute:

$$MAX_E(\mathbb{B}_1, \mathbb{B}_2) = \|C_1 - C_2\| + (r_1 + r_2)$$

$$MIN_E(\mathbb{B}_1, \mathbb{B}_2) = (\|C_1 - C_2\| - (r_1 + r_2)) I(\|C_1 - C_2\| - (r_1 + r_2) \geq 0)$$

where r_1 and r_2 are the radius for ball \mathbb{B}_1 and, respectively, for ball \mathbb{B}_2 , and $\|C_1 - C_2\|$ is the Euclidean distance between the centers of the balls. $I()$ denotes the indicator function.

Solution for the correlation distance.

We assume that the smoothing parameter is J and thus only the first J coefficients of θ are non-zero. Thus in the J dimensional space

$$\frac{\langle \tilde{\theta}_1, \tilde{\theta}_2 \rangle}{\|\tilde{\theta}_1\| \|\tilde{\theta}_2\|} = \cos(\tilde{\theta}_1, \tilde{\theta}_2).$$

The problem is to find the maximum (for MAX_C) and the minimum (for

MIN_C) angle between any two points in the balls \mathbb{B}_1 and \mathbb{B}_2 . The centers of the two balls are C_1 and C_2 , and the origin of the J dimensional space is O . The tangents from the origin to the balls intersect the balls in T_1 and T_2 . We want to find the maximum and the minimum angle $\widehat{T_1OT_2}$.

Denote the coordinates of T_1, T_2, C_1, C_2 and O :

$$t_1 = (t_{11}, \dots, t_{1J}), t_2 = (t_{21}, \dots, t_{2J})$$

$$c_1 = (c_{11}, \dots, c_{1J}), c_2 = (c_{21}, \dots, c_{2J})$$

$$o = (o_1, \dots, o_J).$$

We want to optimize the function:

$$f(t_1, t_2) = \frac{\langle t_1, t_2 \rangle}{\|t_1\| \|t_2\|} = \frac{\sum_{j=1}^J t_{1j} t_{2j}}{\sqrt{\sum_{j=1}^J t_{1j}^2} \sqrt{\sum_{j=1}^J t_{2j}^2}}$$

over the set $(t_1, t_2) \in B_1 \times B_2$ where

$$B_i = \{t_i = (t_{i1}, \dots, t_{iJ}) : \sum_{j=1}^J t_{ij}^2 = \sum_{j=1}^J c_{ij}^2 - r_i^2 \text{ (1) and } \sum_{j=1}^J (t_{ij} - c_{ij})^2 = r_i^2 \text{ (2)}\} =$$

$$\{t_i = (t_{i1}, \dots, t_{iJ}) : \|t_i\|^2 = \|c_i\|^2 - r_i^2 \text{ and } \|t_i - c_i\| = r_i\}$$

where the first condition is due to the tangent from the origin to the balls and the second condition is that T_1 and T_2 are on the envelope of the two balls.

The equivalent optimization problem is to minimize/maximize:

$$\sum_{j=1}^J t_{1j} t_{2j} = \langle t_1, t_2 \rangle$$

with $t_1 \in B_1$ and $t_2 \in B_2$. \mathbb{B}_1 and \mathbb{B}_2 .

We solve this optimization problem using Lagrange multipliers. The objective function is:

$$f(t_1, t_2) = \langle t_1, t_2 \rangle$$

with the constraints

$$g_i(t_1, t_2) = \langle t_i, c_i \rangle - (\|c_i\|^2 - r_i^2)$$

$$h_i(t_1, t_2) = \|t_i\|^2 - (\|c_i\|^2 - r_i^2)$$

Denote $\|c_i\|^2 - r_i^2 = s_i^2$ for the ease of notation.

The optimization problem by Lagrange's theorem is equivalent to solving:

$$\Delta f(t_1, t_2) = \mu_1 \Delta g_1(t_1, t_2) + \mu_2 \Delta g_2(t_1, t_2) + \lambda_1 \Delta h_1(t_1, t_2) + \lambda_2 \Delta h_2(t_1, t_2) \quad (7)$$

with the first order derivatives

$$\Delta f(t_1, t_2) = (t_{21}, \dots, t_{2J}, t_{11}, \dots, t_{1J})$$

$$\Delta g_1(t_1, t_2) = (c_{11}, \dots, c_{1J}, 0, \dots, 0)$$

$$\Delta g_2(t_1, t_2) = (0, \dots, 0, c_{21}, \dots, c_{2J})$$

$$\Delta h_1(t_1, t_2) = (2t_{11}, \dots, 2t_{1J}, 0, \dots, 0)$$

$$\Delta h_2(t_1, t_2) = (0, \dots, 0, 2t_{21}, \dots, 2t_{2J}).$$

We translate the equation 7 into a $2J + 4$ equations with $2J + 4$ unknowns.

For $j = 1, \dots, J$,

$$\begin{cases} t_{1j} = 2\lambda_2 t_{2j} + \mu_2 c_{2j} \\ t_{2j} = 2\lambda_1 t_{1j} + 2\mu_1 c_{1j} \end{cases} \quad (8)$$

with the solution:

$$\begin{cases} t_{1j} = \frac{c_{2j}\mu_2 + 2c_{1j}\lambda_2\mu_1}{1-4\lambda_1\lambda_2} \\ t_{2j} = \frac{c_{1j}\mu_1 + 2c_{2j}\lambda_1\mu_2}{1-4\lambda_1\lambda_2}. \end{cases} \quad (9)$$

Thus we have to find only $\lambda_1, \lambda_2, \mu_1$, and μ_2 by using the constrains

$$g_i(t_1, t_2) = 0, \quad h_i(t_1, t_2) = 0.$$

which can be translated into 4 equations with 4 unknowns:

$$\begin{cases} \frac{\mu_2 \langle c_1, c_2 \rangle + 2\lambda_2 \mu_1 \|c_1\|^2}{1-4\lambda_1\lambda_2} = s_1^2 \\ \frac{\mu_1 \langle c_2, c_1 \rangle + 2\lambda_1 \mu_2 \|c_2\|^2}{1-4\lambda_1\lambda_2} = s_2^2 \\ \frac{\mu_2^2 \|c_2\|^2 + 4\lambda_2^2 \mu_1^2 \|c_1\|^2 + 4\lambda_2 \mu_1 \mu_2 \langle c_1, c_2 \rangle}{(1-4\lambda_1\lambda_2)^2} = s_1^2 \\ \frac{\mu_1^2 \|c_1\|^2 + 4\lambda_1^2 \mu_2^2 \|c_2\|^2 + 4\lambda_1 \mu_1 \mu_2 \langle c_1, c_2 \rangle}{(1-4\lambda_1\lambda_2)^2} = s_2^2. \end{cases} \quad (10)$$

This last system of equations can be solved using Mathematica. The system will have more than one solution. We take the solution which minimizes (for MAX_C) and maximizes (for MIN_C) $\langle t_1, t_2 \rangle$. This will be the maximal and the minimal distances between two confidence balls $M_C(\mathbb{B}_1, \mathbb{B}_2)$.

References

- [1] Ziv Bar-Joseph, Erik D. Demaine, David K. Gifford, Angle M. Hamel, Tommi S. Jaakkola and Nathan Srebro, “K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data”, *Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI 2002)* LNCS 2452, pp 506-520.
- [2] Bar-Joseph, Z., Gerber, G., Gifford, D.K., Jaakkola, T.S. (2002). “A new approach to analyzing gene expression time series data”, *Proceedings of the 6th Annual International Conference on RECOMB*, pp 39-48.
- [3] Ben-Dorr, A, Shamir, R. and Yakhimi, Z. (1999). “Clustering gene expression patterns”, *J. of Computational Biology*.
- [4] Beran, R., Dúmbgen, L. (1998), “Modulation of estimators and confidence sets”, *Annals of Statistics*, 26, 5, pp 1826-1856.
- [5] Chudova, D., Gaffney, S., Mjolsness, E., Smyth, P.(2003), ”Translation-invariant mixture models for curve clustering”, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 79 - 88.
- [6] Fraley, C., Raftery, E. (2002), ”Model-Based Clustering, Discriminant Analysis, and Density Estimation”, *Journal of the American Statistical Association*, 97, 611-631.
- [7] Hartigan, J.A. (1975), *Clustering Algorithms*, John Wiley & Sons, Inc.

- [8] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., Brown, P. (2000), "'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns", *Genome Biology*, I(2):research0003.1-0003.21.
- [9] Li, Ker-Chau (1989), "Honest confidence regions for nonparametric regression", *Annals of Statistics*, 3, pp 1001-1008.
- [10] Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* , 66, pp. 846-850.
- [11] Serban, N., Wasserman, L. (2005), "CATS: Cluster after transformation and smoothing", *Journal of the American Statistical Association*, 100, 990-999.
- [12] Tibshirani, R., Walther, G., Hastie, T. (Dec 2000), "Estimating the number of clusters in a dataset via the Gap statistic". Technical report, published in *Journal of the Royal Statistical Society, B*, 2000.
- [13] Yeung, K.Y., Murua, A., Raftery, A., Ruzzo, W.L. (2001). "Model-Based Clustering and Data Transformations for Gene Expression Data", *Bioinformatics*, 17, 977-987.
- [14] Wakefield, J., Zhou, C., Self, S. (2002), "Modelling Gene Expression Data over Time: Curve Clustering with Informative Prior Distributions", *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting*, 2003.

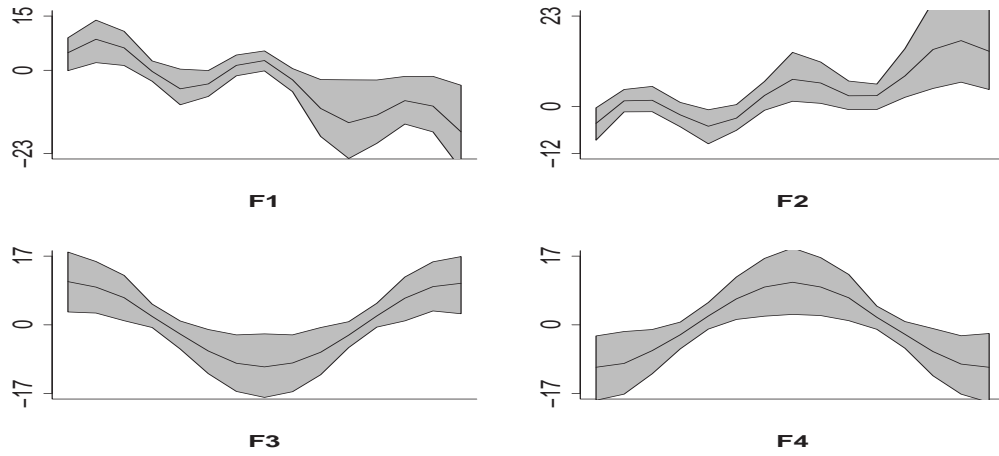


Figure 1: Mean, 10th and 90th percentile for the four true clusters in the synthetic data with $\sigma \sim \text{Unif}(1.0, 1.7)$ and $m = 15$.

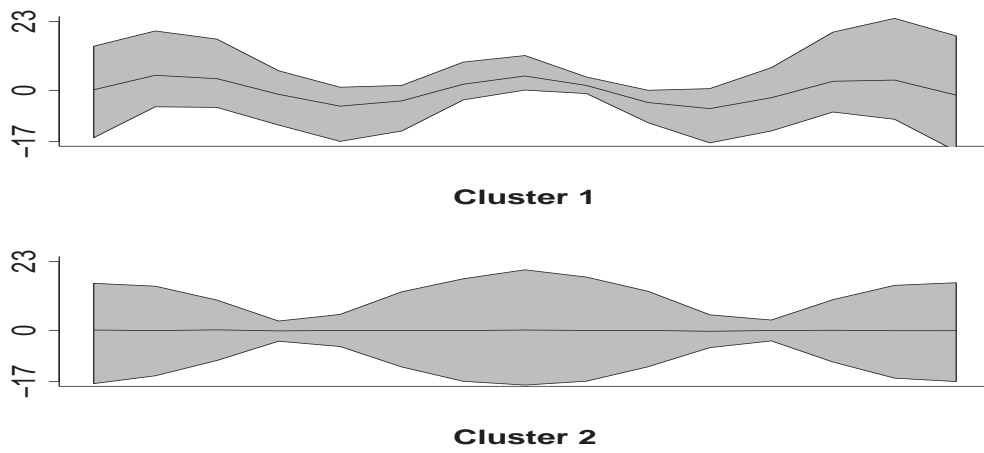


Figure 2: Model-based clustering assignment: Mean, 10th and 90th percentile.

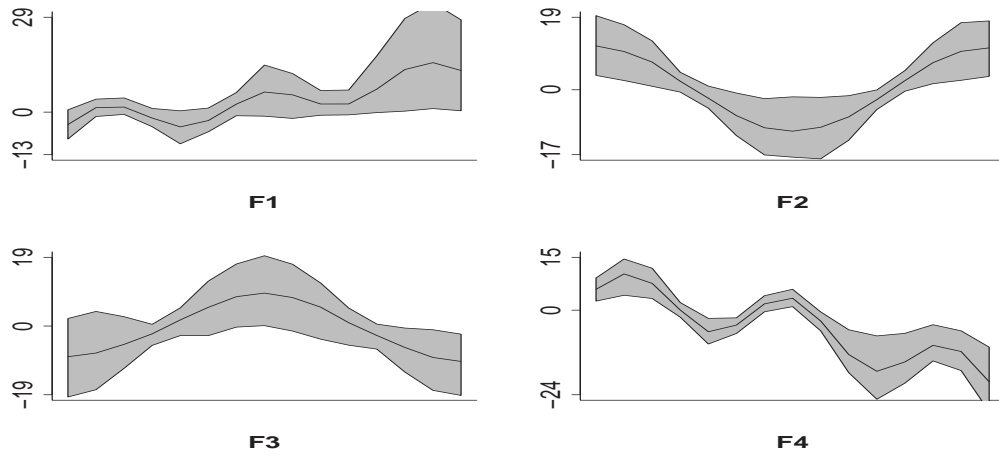


Figure 3: Bar-Joseph method assignment: Mean, 10th and 90th percentile for four clusters (input variable)

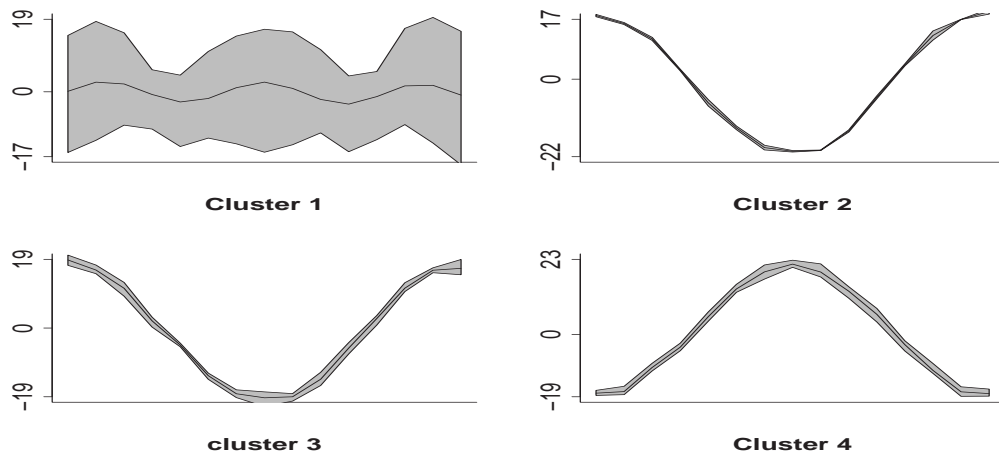


Figure 4: Hierarchical clustering of the estimated curves assignment: Mean, 10th and 90th percentile for four clusters (input variable).

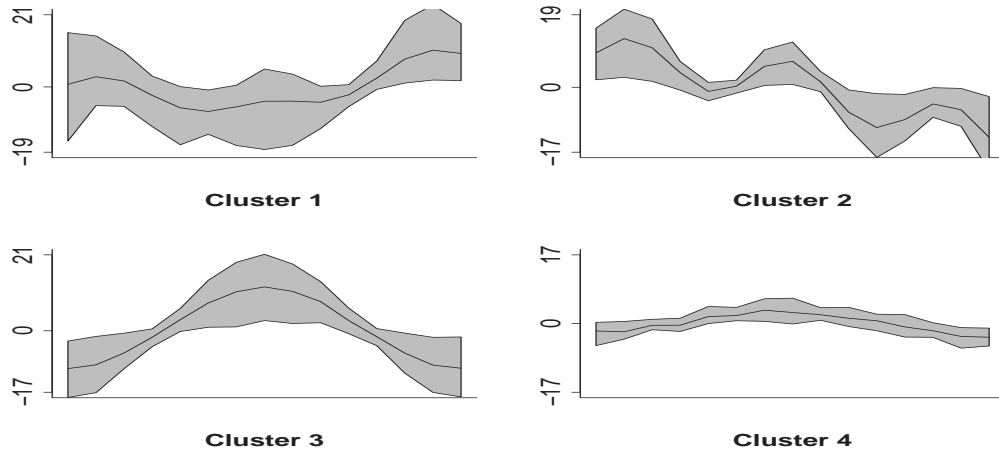


Figure 5: K-means assignment based on the estimated curves: Mean, 10th and 90th percentile for four clusters (input variable).

Clustering Technique	Misclustering Error before screening	Misclustering Error after screening
Yeung et al (2001)	.26	.25
Bar-Joseph et al (2001)	.09	.03
Hierarchical clustering of the estimated distances	.17	0
k -means clustering of the estimated distances	.25	0
Hierarchical clustering of the maximal distances	–	0

Table 1: Clustering error rates based on Rand’s index criteria. The measure of the similarity in shape of any two curves is the correlation. The smoothing parameter is $J = 4$ and the number of clusters is $K = 4$. We compare four different methods before screening the curves/objects that cannot be clustered and after screening.

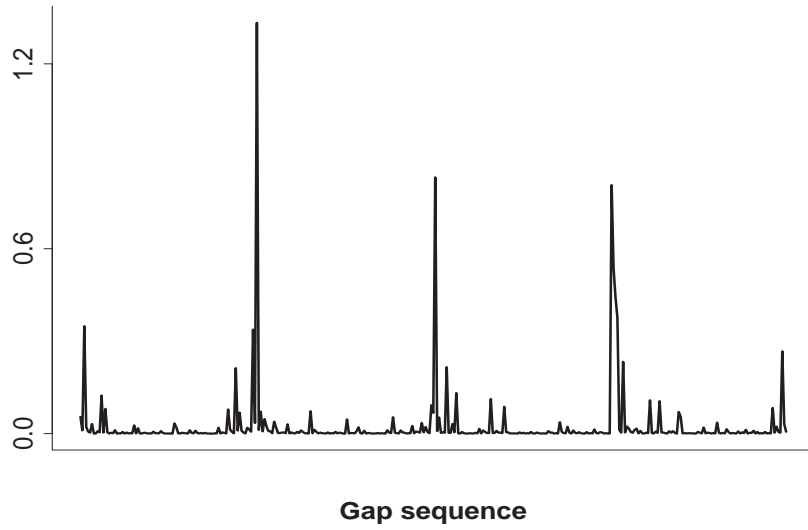


Figure 6: Gap sequence of the single-linkage tree of the estimated distances.

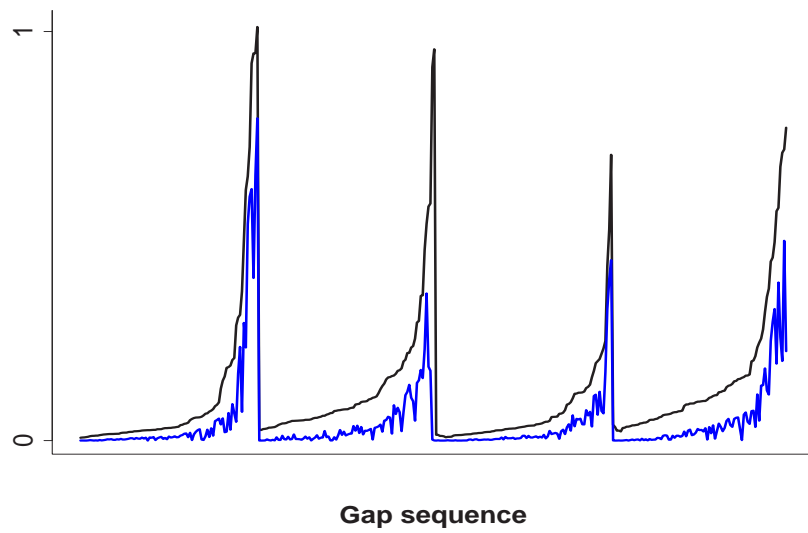


Figure 7: Gap sequence for the single-linkage tree of the maximal distances (shown in black). The minimal distances are shown in blue.