

Clustering Confidence Sets

Nicoleta Serban¹

Industrial Systems and Engineering School
Georgia Institute of Technology

In this article, we present a novel approach to clustering finite or infinite dimensional objects observed with different uncertainty levels. The novelty lies in using confidence sets rather than point estimates to obtain cluster membership and the number of clusters based on the distance between the confidence set estimates. The minimal and maximal distances between the confidence set estimates provide confidence intervals for the true distances between objects. The upper bounds of these confidence intervals can be used to minimize the within clustering variability and the lower bounds can be used to maximize the between clustering variability. We assign objects to the same cluster based on a min – max criterion and we separate clusters based on a max – min criterion. We illustrate our technique by clustering a large number of curves and evaluate our clustering procedure with a synthetic example and with a specific application.

Key words and phrases: single-linkage tree, distance confidence interval, gap sequence, clustering error rate, simultaneous confidence sets, Compustat Global database, Q-ratio.

1 Introduction

In this paper we introduce a technique for clustering a large set of objects observed with different uncertainty levels. Current approaches cluster their point estimates. We propose to use simultaneous confidence sets.

One advantage of clustering based on confidence sets is that we can cluster the confidence intervals of the distances between objects rather than the point estimates of these distances. Based on the upper and lower bounds of the distance confidence intervals, we define the estimated within and between variabilities. Two useful inequalities are provided in Section 3.1: the estimated within variability is larger than the true within variability and the

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 for their input. Importantly, the author is thankful to the two referees who provided useful feedback, and to the editor and associate editor whose input greatly helped improving the presentation of this paper.

estimated between variability is smaller than the corresponding true variability. Under these inequalities, we assign the clustering membership by minimizing the estimated within variability and separate the clusters or estimate the number of clusters by maximizing the estimated between variability. Under this algorithmic framework, we guarantee *small true within variability and large true between variability* with high probability. We describe the clustering algorithm in Section 4 and we derive the inequalities for the within and between variabilities in Section 3.1.

A second advantage is that we can allow for different degrees of uncertainty in the estimated clustering membership by controlling the resolution that one chooses to perform the clustering. The clustering resolution is provided by the number of optimal dimensions of the reduced transformed space after dimensionality reduction of the space where the objects lie. Therefore, one can warrant *coarse resolution clustering with low uncertainty or fine resolution clustering with high uncertainty*. We can control the resolution level using a multi-scale approach as exemplified for curve estimation in Section 5.1.

A third advantage of clustering based on confidence sets is that we can identify a set of *objects that have a very low contrast-to-noise ratio*. Generally, these objects make the clustering assignment unstable, and therefore, the number of clusters and the clustering membership will be unreliably estimated. Cluster instability in the presence of high uncertainty due to low contrast-to-noise ratio observations is discussed in Section 3 and Section 6.2.

In Figure 1, we present estimated confidence sets for 14 curves simulated from two similar patterns but with different levels of variability and on different scales. The confidence sets are estimated after dimensionality-reduction of the domain where the curves lie using an orthonormal basis transformation. Each circle represents the confidence set of one curve and it is a two-dimensional projection onto the reduced-dimensionality space. The confidence sets are colored according to their pattern. The blue points are the true transformed coefficients of the two similar patterns, and the red and black points are the estimated transformed coefficients after dimensionality reduction. The estimated coefficients are the centers of the two-dimensional confidence sets. Four of the two-dimensional confidence sets contain the ori-

gin - they correspond to low contrast-to-noise ratio observations since the noise dominates the pattern. Moreover, some of the observations have the center of their confidence sets away from the true center both in Euclidean and cosine measure. The point estimates of the transform coefficients cluster around the true coefficients into two distinct clusters. However, the confidence sets from both clusters overlap forming one cluster. Therefore, *when clustering confidence sets, similar patterns will be clustered together providing a more parsimonious clustering.* This will be revealed in our empirical example where we identify only four well separated clusters but small sub-clusters may be present. See Section 7.

We illustrate the general algorithm by clustering a large number of curves and evaluate the clustering procedure with a synthetic example. The confidence sets are estimated as introduced by Beran and Dúmbgen (1998) or using a multi-scale approach. We first transform using an orthonormal basis of functions and estimate the confidence sets in a low-dimensional transform space. See Section 5.2. When clustering curves, different distances can be considered as discussed in Section 3.1. Since we cluster in the transform domain, we need to find equivalent distances in the functional and transform domains. For example, clustering with the Euclidean distance in the functional domain is equivalent to clustering in frequency in the transform domain. Also clustering with the correlation measure in the functional domain is equivalent to clustering with the cosine measure in the transform domain.

We also apply our clustering method to a real example where we study the temporal Q-ratio patterns for companies listed in the Compustat Global database in manufacturing sectors tabulated in Table 1. Q-ratio is used to measure a firm's market value in corporate finance. We derived the yearly Q-ratios from financial statements included in Compustat database and adjusted for long-term macro-economic changes such as GDP, inflation, etc. Description of the data and the findings based on two clustering methods are provided in Section 7.

There is a large statistical literature on clustering multiple objects and more specifically, on clustering multiple curves. For a reference list on regularization and filtering clustering methods see James and Sugar (2003).

Generally, clustering techniques can provide hard or soft boundaries between clusters. The current soft clustering methods are model-based (see James and Sugar (2003) for a filtering method, Fraley and Raftery (2002); Wakefield, Zhou and Self (2002); Chudova, Gaffney, Mjolsness and Smyth (2003) for regularization methods). Hard clustering methods can also use regularization or filtering (see Ben-Dorr, Shamir and Yakhimi (1999), Hastie et al (2000) for regularization methods and Bar-Joseph, Gerber, Gifford and Jaakkola (2002); Serban and Wasserman (2005) for filtering methods). One of the difficulty in cluster analysis is the estimation of the number of clusters. Methods for estimating the number of clusters have been developed in Tibshirani, Walther and Hastie (2000), Sugar and James (2003) and the references therein. Model-based clustering methods commonly estimate the number of clusters using the Bayesian information criterion (BIC). Incorporating uncertainty of the estimated cluster separation in the clustering analysis has been also discussed in Medvedovic, Yeung and Bumgarner (2004).

2 General Algorithm

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 , $i = 1, \dots, N$ at a given significance level $1 - \alpha$:

$$P(o_i \in \mathbb{B}_i, \text{ for } i = 1, \dots, n) \geq 1 - \alpha.$$

2. Choose an appropriate measure of similarity for the objects to be clustered, denote this measure d , and measure the minimal and maximal distances 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.

In the following section, the algorithm above is implemented for a particular type of objects, specifically, for curves.

3 Clustering Curves

The general setting is:

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

where the errors ϵ_{ij} are assumed independent with $\mathbb{E}(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = 1$. Y_{ij} is the j^{th} observation on the i^{th} curve where N is typically much larger than m . Without loss of generality, we assume that the design points t_{ij} are in $[0, 1]$ and fixed, but they are not necessarily equally spaced. For ease of notation, we assume that all the curves are observed across equal number of design points m , but this assumption can be relaxed.

3.1 Distance Definition

Our main objective is to cluster the curves f_i for $i = 1, \dots, N$ by shape similarity. Shape similarity can be defined in terms of the similarity features of primary interest. For example, if qualitative features such as number of modes are of main interest, shape similarity measures shall capture the modality of the curves. If global shape similarity is the feature of interest, a goodness-of-fit measure will suffice. Examples of goodness-of-fit distances are the L^p norms for $p = 1, 2, \dots, \infty$. To account for different variability within each curve and to cluster regardless of scale, one can apply the L^p norm to the standardized curves defined as:

$$\tilde{f}(x) = \frac{f(x) - \int_0^1 f(s) ds}{\sqrt{\int_0^1 \left(f(t) - \int_0^1 f(s) ds \right)^2 dt}}$$

The shape similarity measure for two curves f and g becomes:

$$\rho(f, g) = \|\tilde{f} - \tilde{g}\|_p = \left(\int_0^1 (\tilde{f}(x) - \tilde{g}(x))^p \right)^{1/p}, \quad (2)$$

where $1 - \rho(f, g)$ for $p = 2$ is the Spearman's correlation. Other shape similarity measures are the rank correlation as presented in Heckman and Zamar (2000) and the similarity measures introduced in Marron and Tsybakov (1995). The measures in Marron and Tsybakov (1995) quantify the similarity between two curves by measuring the distances between their graphs.

If $\rho(f_1, f_2)$ is the similarity measure between two curves/objects f_1 and f_2 (e.g. L^p norm for standardized curves), we define the maximal and minimal distances between the confidence sets of f_1 and f_2 , \mathbb{B}_1 and \mathbb{B}_2 , as:

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

If \mathbb{B}_1 and \mathbb{B}_2 are $1 - \alpha$ simultaneous confidence sets for f_1 and f_2 , since

$$P(f_1 \in \mathbb{B}_1 \text{ and } f_2 \in \mathbb{B}_2) \geq 1 - \alpha,$$

the maximal and minimal measures bound the true distance between f_1 and f_2 with probability $1 - \alpha$:

$$P(MIN(\mathbb{B}_1, \mathbb{B}_2) \leq \rho(f_1, f_2) \leq MAX(\mathbb{B}_1, \mathbb{B}_2)) \geq 1 - \alpha.$$

For a clustering of K clusters denoted as $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ where each cluster \mathcal{C}_k contains n_k objects for $k = 1, \dots, K$, we define the *estimated within and between clustering variabilities* as below:

$$\begin{aligned} \hat{W}(\mathcal{C}) &= \sum_{k=1}^K \frac{1}{n_k^2} \sum_{i,j \in \mathcal{C}_k} u_{ij} \\ \hat{B}(\mathcal{C}) &= \sum_{k=1, l=1, k \neq l}^K \frac{1}{n_k n_l} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_l} l_{ij} \end{aligned}$$

where $l_{ij} = MIN(\mathbb{B}_i, \mathbb{B}_j)$ and $u_{ij} = MAX(\mathbb{B}_i, \mathbb{B}_j)$, and the *true within and between clustering variabilities*:

$$\begin{aligned} W(\mathcal{C}) &= \sum_{k=1}^K \frac{1}{n_k^2} \sum_{i,j \in \mathcal{C}_k} d_{ij} \\ B(\mathcal{C}) &= \sum_{k=1, l=1}^K \frac{1}{n_k n_l} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_l} d_{ij} \end{aligned}$$

where d_{ij} is the true distance between f_i and f_j , $\rho(f_i, f_j)$. Based on the notations above, we obtain the following lemma. The proof of this lemma is briefly discussed in the Appendix.

Lemma 1 *For a set of curves or objects with $1 - \alpha$ simultaneous confidence sets*

$$P(f_i \in \mathbb{B}_i \text{ for } i = 1, \dots, n) \geq 1 - \alpha,$$

the relationship between the within and between variabilities becomes

$$\mathbb{P}\left(W(\mathcal{C}) \leq \hat{W}(\mathcal{C}) \text{ and } \hat{B}(\mathcal{C}) \leq B(\mathcal{C})\right) \geq 1 - \alpha.$$

Subsequently, we will need to assign the cluster membership by minimizing the estimated within variability and maximizing the estimated between variability as defined in this lemma.

To exemplify our general algorithm, we will consider the L^2 norm applied to standardized curves. We will show how to estimate the maximal and minimal measures in the next section.

3.2 Distance Estimation

Let $f = \sum_{r=0}^{\infty} \theta_{1r} \psi_r$ and $g = \sum_{r=0}^{\infty} \theta_{2r} \psi_r$ be the basis decompositions for two functions in L^2 , where $\{\psi_r\}$ is an orthonormal basis in L^2 with ψ_0 a constant function. Then the (function space) correlation of the two curves can be equivalently expressed in terms of the frequency coefficients:

$$\text{cor}(f, g) = \frac{\sum_{r=1}^{\infty} \theta_{1r} \theta_{2r}}{\sqrt{\sum_{r=1}^{\infty} \theta_{1r}^2} \sqrt{\sum_{r=1}^{\infty} \theta_{2r}^2}} = \frac{\langle \tilde{\theta}_1, \tilde{\theta}_2 \rangle}{\|\tilde{\theta}_1\| \|\tilde{\theta}_2\|} = 1 - \frac{\left\| \frac{\tilde{\theta}_1}{\|\tilde{\theta}_1\|} - \frac{\tilde{\theta}_2}{\|\tilde{\theta}_2\|} \right\|^2}{2}. \quad (3)$$

where $\tilde{\theta}_1 = (\theta_{11}, \dots)$, $\tilde{\theta}_2 = (\theta_{21}, \dots)$. The corresponding similarity distance becomes

$$\rho(f, g) = 1 - \text{cor}(f, g) = \frac{\left\| \frac{\tilde{\theta}_1}{\|\tilde{\theta}_1\|} - \frac{\tilde{\theta}_2}{\|\tilde{\theta}_2\|} \right\|^2}{2}. \quad (4)$$

The distance $\rho(f, g)$ is not defined for f or/and g constant since these functions correspond to the origin in the space where ρ is a measure between space vectors. Constant curves form a separate cluster and therefore, we exclude them from the further cluster analysis.

Remark: *When one of the functions f or g is close to being constant, the denominator in the definition of the distance measure provided in (3)*

is close to zero, which will cause estimates of the distance measures to be very unstable, and consequently, the clustering estimation becomes unstable also. Therefore, any clustering algorithm becomes more stable when we first remove those curves that are close to be constant. Therefore, a first step in our cluster analysis is to remove those observed curves that correspond to confidence sets including the origin point.

The result above applies to any decomposition based on an orthonormal basis (e.g. Fourier basis, wavelet basis, spline basis, functional principal components, etc.). Therefore, clustering based on correlation measure in the time domain is equivalent to clustering the standardized coefficients with Euclidean distance or the cosine of the angle between the vectors of coefficients in the transform domain.

When the similarity measure between two curves is the measure in (2) and the two curves are decomposed using an orthonormal basis, the maximal and minimal measures between two confidence sets of two curves become:

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

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

Both the minimal and the maximal measures have close form expressions when the measure of similarity between two curves is the correlation or the Euclidean distance. We derive the expressions of these measures in the Appendix. For other shape similarity measures we can obtain approximate maximal and minimal distances via Monte Carlo simulations.

4 Clustering Algorithm

Next we introduce a clustering algorithm that approximately maximizes the difference between the estimated within and between variabilities.

In the clustering procedure, we first “order” the objects according to their closest maximal distances (min – max criterion) and we call this the ‘min-max’ gap sequence. The order matters as we use it to assign the cluster mem-

bership. If two objects/curves f_i and f_j are close with respect to their maximal distance $MAX(\mathbb{B}_i, \mathbb{B}_j)$, then with high probability they are close with respect to the true distance $\rho(f_i, f_j)$. Since the algorithm assigns clusters using the maximal distances, we ensure that the within variability is small with high probability. To obtain the ordered distances, we apply the single-linkage tree algorithm to the maximal distance matrix $U = \{u_{ij} = MAX(\mathbb{B}_i, \mathbb{B}_j)\}$. But any other linkage algorithm can be used to obtain this order.

Second, we separate clusters based on the minimal distances: $L = \{l_{ij} = MIN(\mathbb{B}_i, \mathbb{B}_j)\}$ (max – min criterion). If two objects have large minimal distance, then with high probability they are well separated with respect to the true distance. The separation algorithm provides the estimated number of clusters, which ensures large between variability.

The clustering algorithm is as follows:

1. Let $I(j)$, $j = 1, \dots, N$ are the nodes and $G(j)$, $j = 1, \dots, N$ represent the links between the nodes of the ordered distances. Hartigan (1975) calls G the *gap sequence*.
 - (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 the corresponding distance.
2. Once we have the distances ordered, 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:

$$\hat{W}(j) = \frac{1}{j^2} \sum_{k=1}^j \sum_{l=1}^j u_{I(k), I(l)}, \quad \hat{B}(j) = \frac{2}{j} \sum_{k=1}^j l_{I(k), I(j+1)}$$

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

- (b) For K clusters in the sequence $I(1), \dots, I(j)$, where $\{I(1), \dots, I(j_1)\}$ is the first cluster, $\{I(j_1+1), \dots, I(j_2)\}$ is the second cluster, \dots , $\{I(j_{K-1}+1), \dots, I(j)\}$ is the K -th cluster. Assume $I(j+1)$ forms a $(K+1)$ th cluster and compute the estimated within and between variabilities between this cluster and the K^{th} cluster as in part (a): $\hat{W}(j, K)$ and $\hat{B}(j, K)$. Include $I(j+1)$ in the K^{th} cluster if $\hat{W}(j, K) \approx \hat{B}(j, K)$.

In the algorithm above, the approximation threshold is data driven. A large threshold will provide a very parsimonious clustering, whereas smaller meaningless clusters will appear as the threshold decreases.

The algorithm complexity is not larger than $O(N^2)$ since both ordering the distances using single-linkage and separating the clusters are of order $O(N^2)$. We have close form expressions for the minimal and maximal measures when the similarity measure is the Euclidean or the correlation, but for other similarity measures, we may need to solve complex optimization problems and the algorithm complexity may increase of an order higher than $O(N^2)$.

5 Confidence Set Estimation

In this section, we present an example of simultaneous confidence set estimation. For this example, 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 :

$$\left\{ f(x) = \sum_{r=0}^{\infty} \beta_r \psi_r(x) : \sum_{r=0}^{\infty} \beta_r^2 (r+1)^{2\beta} \leq c^2 \right\} \quad (5)$$

where $\psi_0, \psi_1, \psi_2, \dots$ is an orthonormal basis for L_2 . The smoothness assumption can be further extended to Besov spaces and other classes of functions depending on the basis one chooses for basis decomposition. The smoothness assumption is also important for estimating simultaneous confidence sets for

the curves f_i . Uniform confidence sets for Fourier basis are constructed in Beran and Dúmbgen (1998) and for wavelet basis are derived in Genovese and Wasserman (2004). Uniform confidence bands have been proposed for other basis of function, for example see Ruppert, Wand and Carroll (2003) for p-spline basis. However, using the confidence bands alone does not allow us to easily derive and compute confidence intervals on the true distances.

In time domain, the estimated confidence sets are infinite dimensional since they are subsets of functions in the Sobolev space (5). Since we use the estimated confidence sets to compute the maximal and minimal distances between two curves, we reduce the dimensionality of the space and estimate these distances in a low dimensional space. For dimensionality reduction, we transform the curves using a basis of functions in L^2 and threshold high resolution coefficients in the transform domain. The optimal number of dimensions/coefficients depends on the resolution level of the true curves and the noise structure. Further, by clustering coarse resolution curve estimates, we allow for low uncertainty in the estimated clustering membership. On the other hand, by clustering fine resolution curve estimates, we may discover more patterns. Therefore, in our procedure, the number of dimensions is a clustering tuning parameter. We can tune the clustering uncertainty using a multi-scale approach for the number of dimensions.

5.1 Multi-scale Method

In the example introduced in this paper, we will use a Fourier basis. If our observations are seasonal, we use the sine-cosine basis, and otherwise, we use the cosine basis. Define the $m \times m$ matrix $\Psi = \{\psi_r(t_{jr})\}_{j=1, \dots, m; r=0, \dots, m-1}$ and perform a Gram-Schmidt orthogonalization on the columns of Ψ to make the columns orthogonal. Denote the new matrix by Φ with components ϕ_{rj} for r and $j = 1, \dots, m$. Let $\hat{\theta}_i = (\hat{\theta}_{i1}, \dots, \hat{\theta}_{im})$ the frequency coefficients

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

For a given number of low frequency coefficients J , the function $\hat{f}_i^J(t) = \sum_{r=0}^J \hat{\theta}_{ir} \psi_r(t)$ is a smoothed version of the i^{th} curve. In this context, J

is a smoothing parameter. 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, we allow for higher uncertainty in the confidence set estimation as J increases.

The multi-scale approach is to consider all values of J simultaneously and choose the one that leads to the clustering with low uncertainty. 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 $O_P(m^{-1/4})$, which is the smallest possible in a nonparametric sense (Li, 1989).

The confidence set for $(\theta_{i1}, \dots, \theta_{iJ})$ is somewhat simple to estimate since $\hat{\theta}_{ir} \approx N(\theta_{ir}, \sigma_i^2/m)$. We have that

$$\sum_{r=1}^J (\theta_{ir} - \hat{\theta}_{ir})^2 \approx \frac{\sigma_i^2}{m} \chi_J^2.$$

Hence,

$$\mathbb{B}_i^J = \left\{ (\theta_1, \dots, \theta_J) : \sum_{r=1}^J (\theta_r - \hat{\theta}_r)^2 \leq \frac{\hat{\sigma}_i^2 \chi_{J, \alpha'}^2}{m} \right\} \quad (6)$$

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

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

If one prefers a clustering method without the input of a tuning parameter, in the next section, we show how to find a global estimate for the smoothing parameter and how to estimate uniform confidence sets.

5.2 Global Method

Using the global smoothing parameter as in Serban and Wasserman (2005), we can apply the method in Beran and Dümbgen (1998) for constructing a

confidence set \mathbb{B}_i for f_i . Fix $\alpha > 0$ and let

$$\mathbb{B}_i = \left\{ (\theta_{i1}, \dots, \theta_{im}) : \sum_{r=1}^m (\theta_{ir} - \hat{\theta}_{ir})^2 \leq s_i^2 \right\} \quad (7)$$

where

$$s_i^2 = \frac{z_\alpha \hat{\tau}_i}{\sqrt{m}} + \hat{R}_i(J), \quad \hat{R}_i(J) = \frac{J \hat{\sigma}_i^2}{m} + \sum_{r=J+1}^m \left(\hat{\theta}_{ir}^2 - \frac{\hat{\sigma}_i^2}{m} \right)_+$$

z_α is the α quantile of the standard normal and $\hat{\tau}_i$ is given in the Appendix. In the formulation above, $R_i(J)$ is an estimate of the risk for \hat{f}_i^J as provided by Beran and Dümbgen (1998) and J is the the global smoothing parameter as derived by Serban and Wasserman (2005). The corresponding confidence set for f_i is $\{\sum_{r=1}^m \theta_{ir} \psi_r(x) : \theta \in \mathbb{B}_i\}$. 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 the multi-scale approach discussed in Section 5.1.

Based on the theorems in Beran and Dümbgen (1998), 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 minimal and maximal distances between their confidence sets.

The proof of this result derives from Theorem 1 in Serban and Wasserman (2005), which states that

$$\liminf_{N \rightarrow \infty} \sup_{f_1, \dots, f_N \in \mathcal{F}_\beta(c)} \mathbb{P} \left(f_i \in \mathbb{B}_i, \forall i = 1, \dots, N \right) \geq 1 - \alpha$$

under the conditions in theorem 1 of this paper.

As already mentioned in Section 1, we can derive an inequality of the within and between variabilities under the conditions stated in Theorem 1. Lemma 2 states that these inequalities hold with high probability asymptotically.

Lemma 2 *Let \mathcal{C} be a clustering of K clusters: $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$. Similar to the result in Lemma 1 and under the conditions of Theorem 1,*

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

The proof of this lemma follows directly from Theorem 1 and the proof is similar to lemma 1 presented in the Appendix.

Similar results can be derived for the multi-scale approach.

6 Method Evaluation: Synthetic Example

We evaluate our clustering technique using synthetic data. The synthetic example consists of 4 different patterns according to the regression model (1) with $m = 25$, $N = 500$ and $\sigma_i \in [.5, 1.5]$. The synthetic data are generated from the model below

$$Y_{ij} = \mu_i + F_k(t_j) + N_m(0, \sigma_i), \quad i = 1, \dots, 500, \quad j = 1, \dots, 25, \quad k = 1, 2, 3, 4$$

where $\mu_i \in [1, 5]$, 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 clusters are in Figure 2. We generate 50 curves from the first pattern (upper plot, left), 100 from the second pattern (upper plot, right), 150 from the third pattern (lower plot, left), and 200 from the fourth pattern.

6.1 Clustering Error Rate

We compare the performance of different clustering techniques using a measure for clustering error as described below.

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

$$T_{n,K}(f, g) = \begin{cases} 1 & \text{if } f \text{ and } g \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The two clustering maps depend on the number of clusters.

The *clustering error rate* for K clusters is

$$\eta(K) = \frac{1}{\binom{N}{2}} \sum_{r < s} I\left(T_{n,K}(f_r, f_s) \neq \hat{T}_{n,K}(f_r, f_s)\right) \quad (9)$$

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).

6.2 Other Methods

We will consider in this section the performance of three 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), and hierarchical clustering of the estimated curves. Last, we describe the performance of the method introduced in this paper.

Yeung et al (2001). The model-based clustering technique introduced in this paper estimates the number of clusters using the BIC approximation criterion. 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 3. The clustering error is about .25.

Bar-Joseph et al (2001). The clustering strategy introduced by Bar-Joseph and colleagues is a filtering method that provides hard clustering membership. We apply this technique to the synthetic data presented above. This method captures the main patterns in the synthetic data when fixing the number of clusters to the true number of patterns. Therefore, the method requires the input of the number of clusters. The clustering error is about

.09 before removing the unstable curves as defined in Section 3.2, and about .03 after removing these curves.

Hierarchical clustering. We apply the single-linkage clustering to the estimated curves. The clustering error is about .15. After discarding the confidence sets including the origin, the hierarchical clustering algorithm assigns the cluster membership correctly. Similar to the method introduced in Bar-Joseph et al (2001), for hierarchical clustering, we need to input the number of clusters.

Clustering Confidence Sets. In Figure 4, we present the gap sequence of the single-linkage tree based on the distance point estimates for our synthetic example. In Figure 5, we present the gap sequence for the single-linkage tree based on the maximal (upper bound) distances as provided by our method. In contrast to the point estimation based clustering, we clearly separate the four clusters when using the lower and upper distance bounds. Subsequently, using our algorithm, we estimate four clusters and the algorithm assigns the cluster membership errorless.

For all the methods compared in this paper, the clustering membership is more stable after removing the synthetic curves with low signal-to-noise ratio which result in large confidence sets containing the origin.

7 Example: Q-ratio Evaluation in Manufacturing Sector

The individual performance of a firm's business or operation is indicated by its financial status revealed in its financial statements such as balance sheets, income statements, statements of cash flows, and retained earning statements. Comparing companies within an industry sector requires adjusting the accounting information for size differences between firms. Subsequently, it is unreasonable to use raw items from financial statements when performing a comparative study across multiple companies.

There are two ways to remove size effects. The first is to express each item on the financial statement as a percentage of one representative item, which is called a common-size statement. The second method to evaluate

the current financial condition and performance of firms is to compute ratios from financial statements. In this paper, we use a market value ratio called Q-ratio to indicate the true worth of a firm's assets in the market. Ross, Westerfield, and Jaffe (2005) explain that Tobin's Q ratio divides the market value of all of the firm's debt plus equity by the replacement value of the firm's assets. This ratio was derived by James Tobin, Nobel laureate in Economics. A Q-ratio between 0 and 1 indicates that the value of a firm's stock is not enough to replace its assets. This suggests overvaluation of the stock. When the Q-ratio is larger than 1, the stock is undervalued. In Tobin's model, this method of stock valuation guides investment decisions (Investopedia, 2006).

We apply our clustering method to companies listed in the Compustat Global database in manufacturing sectors categorized using NAICS - North American Industry Classification System. The U.S., Canada, and Mexico, members of NAFTA, developed NAICS to provide statistical comparability of firm's operational activity across North America. (U.S. Census Bureau 2006). Table 1 shows the NAICS code and corresponding sector names of the companies studied in this paper. We investigate the annual financial statements starting with 1990 to 2005. Also we only consider the set of companies with none or a few missing/incomplete data values. The total number of companies is about 850.

We derived the Q-ratio from financial statements included in Compustat database and adjusted for long-term economic changes as provided by a research group located in Chicago (Ativo Research, Llc). The adjustments are documented in Madden (1999).

We apply the procedure presented in this paper to the Q-ratio patterns, and for $J = 3$, there are about 200 companies out of about 850 that have Q-ratios changing considerably over the past 16 years as their estimates confidence sets are away from origin. For the 200 time-varying patterns, we identify four well separated clusters with smaller sub-clusters as shown in Figure 6. The patterns within each cluster are presented in Figure 7. The largest cluster consists of curves with lower changes in Q-ratio values over time. However, most of the companies in cluster one to three are undervalued at different times over the 16 years: at the beginning of 1990 for companies in cluster 1, during the internet economic burst for companies in

cluster 2, and close to the macro-economic crisis in 2000 for companies in cluster 3. Examples of companies in each of the four clusters are listed in the Appendix.

Figure 8 shows the smoothed patterns as identified with the method in Bar-Joseph et al (2001) for four clusters. As provided in this figure, the four clusters contain mixed patterns.

8 Discussion

In this article, we have presented a novel approach to clustering a large number of observed curves but the general strategy can be applied to other types of objects. The novelty of our technique consists of clustering based on the maximal and minimal distances between confidence sets. Consequently, we obtain confidence interval estimates for the distances between curves. The upper limits of these confidence intervals are used to assign curves in the same cluster, and the lower bounds are used to assign curves in different clusters, which in turn will also give the estimated number of clusters. We used the single-linkage tree to assign the cluster membership, but other clustering algorithms can be used (e.g. the minimum spanning tree or k -ary clustering).

Removing curves/objects with confidence sets containing the origin, we obtain a more stable clustering. As we increase the smoothing parameter, the radius of the confidence sets increases also, and therefore, more confidence sets will contain the origin. In our synthetic example, for $J = 2$, about 36 out of 500 confidence sets contain the origin. For $J = 5$, the number of confidence sets containing the origin increases to 80. On the other hand, at higher uncertainty or less smoothing, we expect to discover more patterns. Consequently, we will choose a low uncertainty level or small smoothing parameter for which we detect all or most of the patterns discovered at finer resolution. In our synthetic example, we discover four patterns, for all smoothing levels from $J = 2$ to $J = 5$, and the separation between the four clusters is similar across all these smoothing levels. See Figure 5. Therefore, $J = 2$ will be low enough to identify the clustering structure in the data.

One other advantage of using our approach is that we can establish a better separation between clusters. Figure 4 displays the gap sequence of

the clustering assignment based on the point distance estimates using single-linkage tree algorithm. Figure 5 displays the gap sequence of the clustering assignment based on the maximal distances $U = \{u_{ij}\}$ using single-linkage algorithm for different smoothing parameters. In the same figure, we include the lower bound of the gap sequence. There is a clearer cluster separation in the gap sequence based on the upper and lower bounds of the distances (Figure 5) than in the gap sequence based on the point distance estimates (Figure 4).

Last, in our empirical example, the gap sequence divides in smaller sub-clusters at higher resolution level. On the other hand, at lower smoothing parameter J , the large clusters are well separated without being divided in many sub-clusters. Therefore, our method gives the flexibility to choose between a finer or coarser clustering structure by changing the resolution level.

There are a few different research directions 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).

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:

$$\begin{aligned} MAX_E(\mathbb{B}_1, \mathbb{B}_2) &= \sup_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \|\theta_1 - \theta_2\|^2 \\ MIN_E(\mathbb{B}_1, \mathbb{B}_2) &= \inf_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \|\theta_1 - \theta_2\|^2. \end{aligned}$$

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

$$\begin{aligned} MAX_C(\mathbb{B}_1, \mathbb{B}_2) &= \sup_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \left(1 - \frac{\langle \theta_1, \theta_2 \rangle}{\|\theta_1\| \|\theta_2\|} \right) \\ MIN_C(\mathbb{B}_1, \mathbb{B}_2) &= \inf_{\theta_1 \in \mathbb{B}_1, \theta_2 \in \mathbb{B}_2} \left(1 - \frac{\langle \theta_1, \theta_2 \rangle}{\|\theta_1\| \|\theta_2\|} \right) \end{aligned}$$

Solution for the Euclidean distance.

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

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

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 confidence sets. $I()$ 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

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

The problem reduces to finding the maximum (for MAX_C) and the minimum (for MIN_C) angle between any two points in the confidence sets \mathbb{B}_1 and \mathbb{B}_2 . The relationship between the maximum and minimum angles with the maximal and minimal distances are derived below

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

The centers of the two confidence sets are C_1 and C_2 , and the origin of the J dimensional space is O . The tangents from the origin to the confidence sets intersect the balls in T_1 and T_2 . We want to find the maximum and the minimum angle $T_1\hat{O}T_2$. See Figure 9.

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

$$\begin{aligned} t_1 &= (t_{11}, \dots, t_{1J}), \quad t_2 = (t_{21}, \dots, t_{2J}) \\ c_1 &= (c_{11}, \dots, c_{1J}), \quad c_2 = (c_{21}, \dots, c_{2J}) \\ o &= (o_1, \dots, o_J). \end{aligned}$$

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

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

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

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 $t_2 \in B_2$.

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:

$$\Delta f(t_1, t_2) = \mu_1 \Delta g_1(t_1, t_2) + \mu_2 \Delta g_2(t_1, t_2) + \lambda_1 \Delta h_1(t_1, t_2) + \lambda_2 \Delta h_2(t_1, t_2) \quad (10)$$

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 10 into a $2J + 4$ equations with $2J + 4$ unknowns.

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

$$\begin{cases} t_{1j} = 2\lambda_2 t_{2j} + \mu_2 c_{2j} \\ t_{2j} = 2\lambda_1 t_{1j} + 2\mu_1 c_{1j} \end{cases} \quad (11)$$

with the solution:

$$\begin{cases} t_{1j} = \frac{c_{2j}\mu_2 + 2c_{1j}\lambda_2\mu_1}{1 - 4\lambda_1\lambda_2} \\ t_{2j} = \frac{c_{1j}\mu_1 + 2c_{2j}\lambda_1\mu_2}{1 - 4\lambda_1\lambda_2}. \end{cases} \quad (12)$$

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:

$$\begin{cases} \frac{\mu_2 \langle c_1, c_2 \rangle + 2\lambda_2 \mu_1 \|c_1\|^2}{1 - 4\lambda_1 \lambda_2} = s_1^2 \\ \frac{\mu_1 \langle c_2, c_1 \rangle + 2\lambda_1 \mu_2 \|c_2\|^2}{1 - 4\lambda_1 \lambda_2} = s_2^2 \\ \frac{\mu_2^2 \|c_2\|^2 + 4\lambda_2^2 \mu_1^2 \|c_1\|^2 + 4\lambda_2 \mu_1 \mu_2 \langle c_1, c_2 \rangle}{(1 - 4\lambda_1 \lambda_2)^2} = s_1^2 \\ \frac{\mu_1^2 \|c_1\|^2 + 4\lambda_1^2 \mu_2^2 \|c_2\|^2 + 4\lambda_1 \mu_1 \mu_2 \langle c_1, c_2 \rangle}{(1 - 4\lambda_1 \lambda_2)^2} = s_2^2. \end{cases} \quad (13)$$

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) $\langle t_1, t_2 \rangle$.

Proof of Lemma 1

The result is derived from the following inequality:

$$P\left(f_i \in \mathbb{B}_i, f_j \in \mathbb{B}_j, i, j = 1, \dots, N\right) \leq \mathbb{P}\left(\text{MIN}(\mathbb{B}_i, \mathbb{B}_j) \leq d(f_i, f_j) \leq \text{MAX}(\mathbb{B}_i, \mathbb{B}_j), i, j = 1, \dots, N\right).$$

Therefore, if \mathbb{B}_i for $i = 1, \dots, N$ are $1 - \alpha$ simultaneous confidence sets for the set of N curves, then

$$\mathbb{P}\left(\text{MIN}(\mathbb{B}_i, \mathbb{B}_j) \leq d(f_i, f_j) \leq \text{MAX}(\mathbb{B}_i, \mathbb{B}_j), i, j = 1, \dots, N\right) \geq 1 - \alpha.$$

From the definition of u_{ij} and l_{ij} distances, the inequality above can be re-expressed as:

$$\mathbb{P}\left(\rho(f_i, f_j) \leq u_{ij} \text{ and } \rho(f_i, f_j) \geq l_{ij} \text{ } i, j = 1, \dots, N\right) \geq 1 - \alpha$$

Since $\rho(f_i, f_j) \leq u_{ij}$ for all $i, j = 1, \dots, N$ implies that $W(\mathcal{C}) \leq \hat{W}(\mathcal{C})$ and $\rho(f_i, f_j) \geq l_{ij}$ for all $i, j = 1, \dots, N$ implies that $B(\mathcal{C}) \geq \hat{B}(\mathcal{C})$, together with the inequality above we have the result in this lemma.

τ^2 estimation

For the high component variance estimator σ_i^2 for curve f_i

$$\hat{\sigma}_i^2 = \frac{1}{m - L} \sum_{i=L+1}^m \hat{\theta}_{ij}^2.$$

we estimate

$$\hat{\tau}_i^2 = \sum_{j=1}^m \left(1 + \frac{1 - 2c_j}{m - J}\right) (4(\hat{\theta}_{ij}^2 - \hat{\sigma}^2)_+ \hat{\sigma}^2 + 2\hat{\sigma}^4) + 4\hat{\sigma}^2 \sum_{j=1}^k \frac{2c_j}{m - J} (\hat{\theta}_{ij}^2 - \hat{\sigma}^2)_+$$

where c_j is 1 for $j \geq J + 1$ and 0 otherwise.

The estimated variance of $\hat{\tau}^2$ is derived in Serban and Wasserman (2005).

Companies names by clusters

Below I included samples of companies and their cluster memberships.

Cluster 1:

Healthcare Technologies Ltd, Cooper Tire & Rubber Co, Ap Pharma Inc, Lancer Orthodontics Inc, Phazar Corp.

Cluster 2:

3Com Corp, Topps Co Inc, Mgp Ingredients Inc Scientific, Technologies Inc, Spectrum Signal Processing, Aleris International Inc.

Cluster 3:

Pfizer Inc, Medtronic Inc, Biogen Idec Inc, Bristol-Myers Squibb Co, Nanometrics Inc, Lilly (Eli) & Co, Colgate-Palmolive Co.

Cluster 4:

Allergan Inc, Avon Products, Varian Medical Systems Inc, Alcoa Inc, Novo-Nordisk A/S -Adr, Protein Design Labs Inc, Tektronix Inc, Procter & Gamble Co, Amylin Pharmaceuticals Inc, Wireless Telecom Group Inc, Lipid Sciences Inc, Texas Instruments Inc.

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] Genovese, C., Wasserman, L. (2005), "Confidence sets for nonparametric wavelet regression", *The Annals of Statistics*, Volume 33, Number 2, 698-729.
- [8] Hartigan, J.A. (1975), *Clustering Algorithms*, John Wiley & Sons, Inc.
- [9] 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.
- [10] Heckman, N., Zamar, R. (2000), "Comparing the shapes of regression functions", *Biometrika*, 87, 135-144.
- [11] James, G.M., Sugar, C.A. (2003), "Clustering for sparsely sampled functional data", *Journal of the American Statistical Association*, 98, p397.
- [12] Li, Ker-Chau (1989), "Honest confidence regions for nonparametric regression", *Annals of Statistics*, 3, pp 1001-1008.
- [13] Marron, J.S., Tsybakov, A.B. (1995), "Visual Error Criteria for Qualitative Smoothing", *Journal of the American Statistical Association*, 90, 499-507.

- [14] Madden B. J. (1999), “CFRoI Valuation - A Total System Approach to Valuing the Firm”, 1st ed., Butterworth-Heinemann, Oxford.
- [15] Medvedovic, M., Yeung, K.Y., Bumgarner, R. (2004), “Bayesian mixture model based clustering of replicated microarray data”, *Bioinformatics*, 20,1222-1232.
- [16] Rand, W. M. (1971), “Objective criteria for the evaluation of clustering methods”. *Journal of the American Statistical Association* , 66, pp. 846-850.
- [17] Ruppert, D., Wand, M.P., Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge University Press.
- [18] Serban, N., Wasserman, L. (2005), “CATS: Cluster after transformation and smoothing”, *Journal of the American Statistical Association*, 100, 990-999.
- [19] Tibshirani, R., Walther, G., Hastie, T. (2001), “Estimating the number of clusters in a dataset via the Gap statistic”, *Journal of the Royal Statistical Society, B*, 63, 411-423.
- [20] 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.
- [21] 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.
- [22] Ross, S.A., Westerfield, R.W. and Jaffe, J.F. (2005), “Corporate Finance”, 7th ed. McGraw-Hill, New York
- [23] Sugar, C., and James, G. (2003), ”Finding the Number of Clusters in a Data Set : An Information Theoretic Approach”, *Journal of the American Statistical Association*, 98, 750-763.

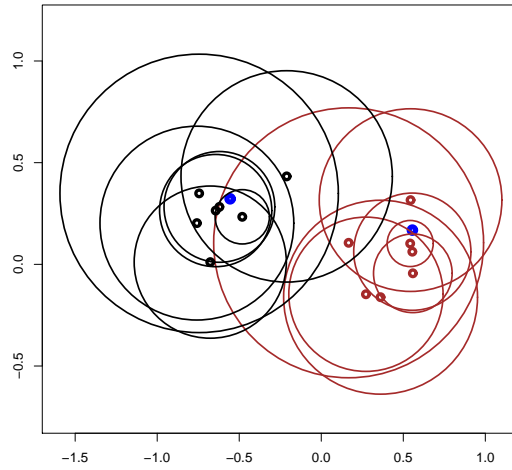


Figure 1: 2-dimensional estimated confidence sets for seven observations simulated from two similar patterns but with different variability levels and at different scales.

- [24] 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.

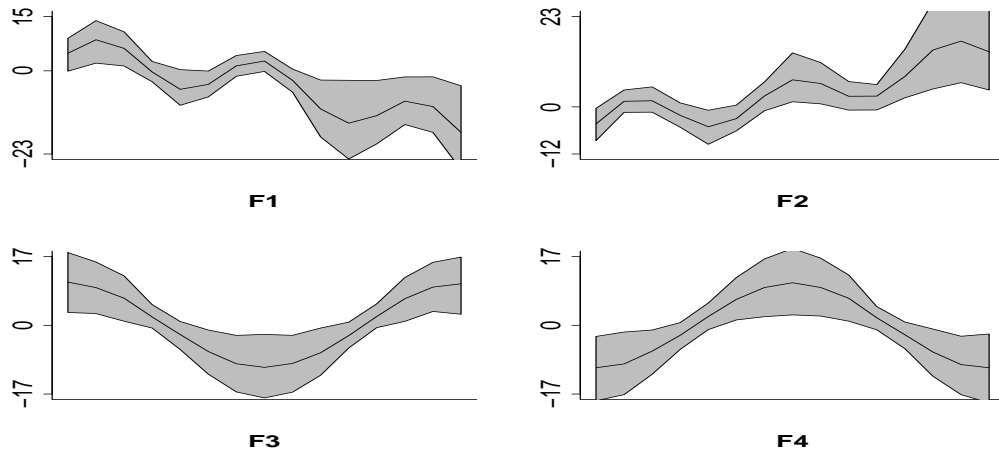


Figure 2: Mean, 10th and 90th percentile for the four true clusters in the synthetic data after rescaling.

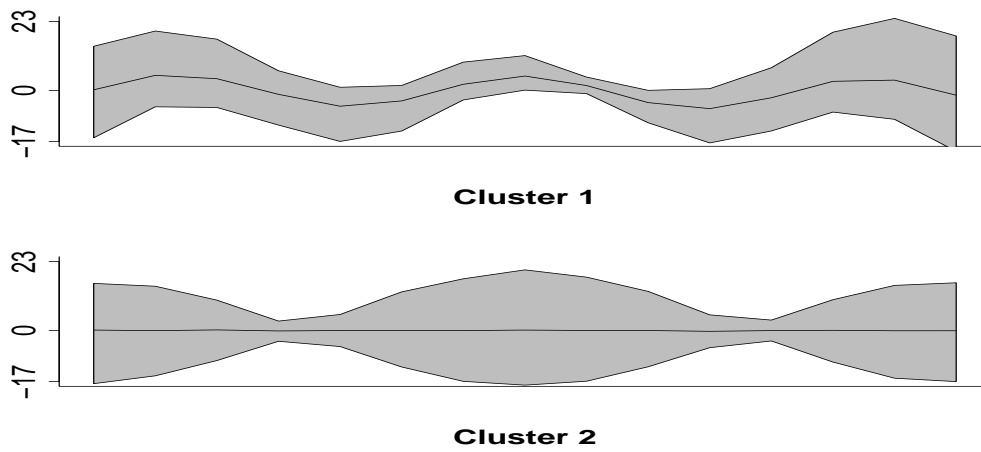


Figure 3: Model-based clustering assignment: Mean, 10th and 90th percentile.

31	Manufacturing
321	Wood Product Manufacturing
322	Paper Manufacturing
323	Printing and Related Support Activities
324	Petroleum and Coal Products Manufacturing
325	Chemical Manufacturing
326	Plastics and Rubber Products Manufacturing
327	Nonmetallic Mineral Product Manufacturing
331	Primary Metal Manufacturing
332	Fabricated Metal Product Manufacturing
333	Machinery Manufacturing
334	Computer and Electronic Product Manufacturing
335	Electrical Equipment, Appliance, and Component Manufacturing
336	Transportation Equipment Manufacturing
337	Furniture and Related Product Manufacturing
339	Miscellaneous Manufacturing

Table 1: Manufacturing Industries in 32 and 33 NAICS listings.

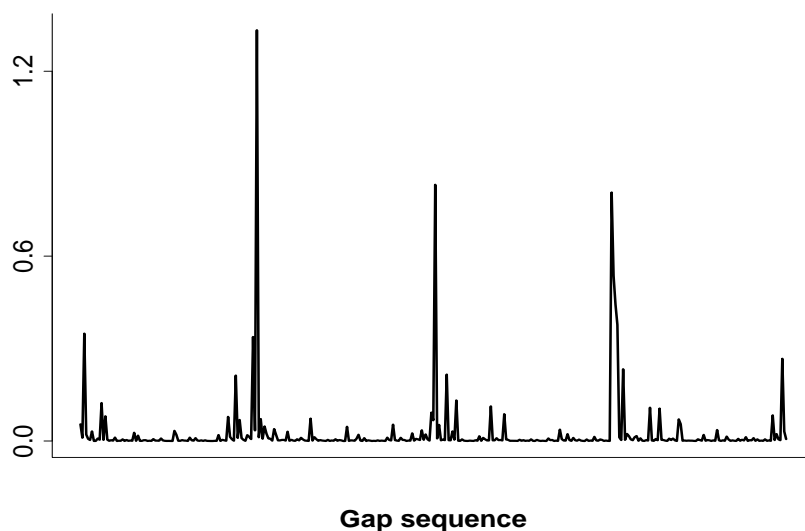
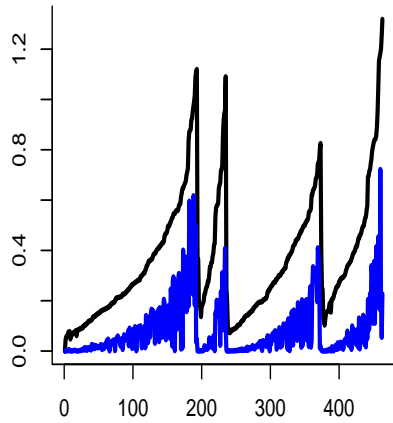
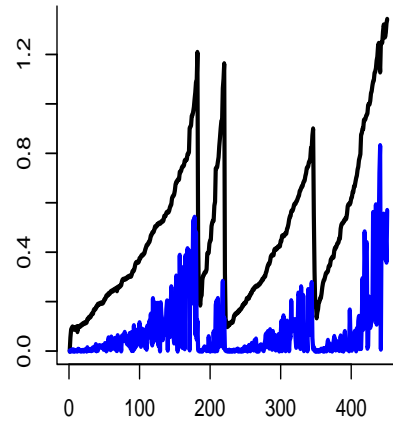


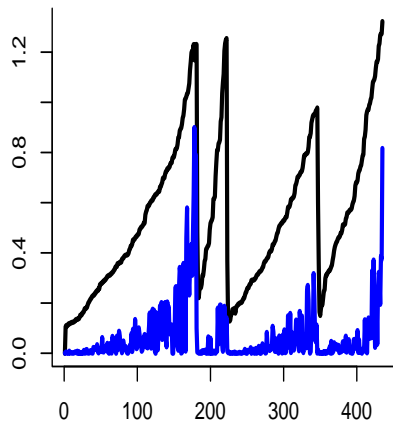
Figure 4: Gap sequence of the single-linkage tree of the point distance estimates.



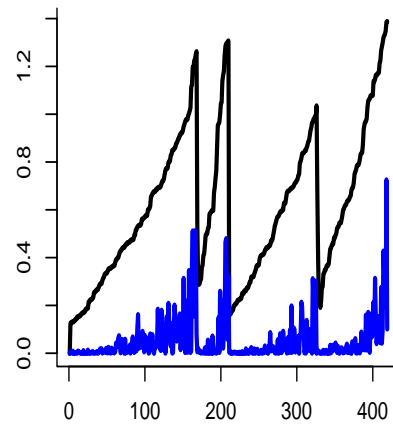
Gap sequence for $J=2$



Gap sequence for $J=3$



Gap sequence for $J=4$



Gap sequence for $J=5$

Figure 5: Gap sequence of the maximal distances (shown in black) and the corresponding minimal distances (shown in blue) for different smoothing levels.

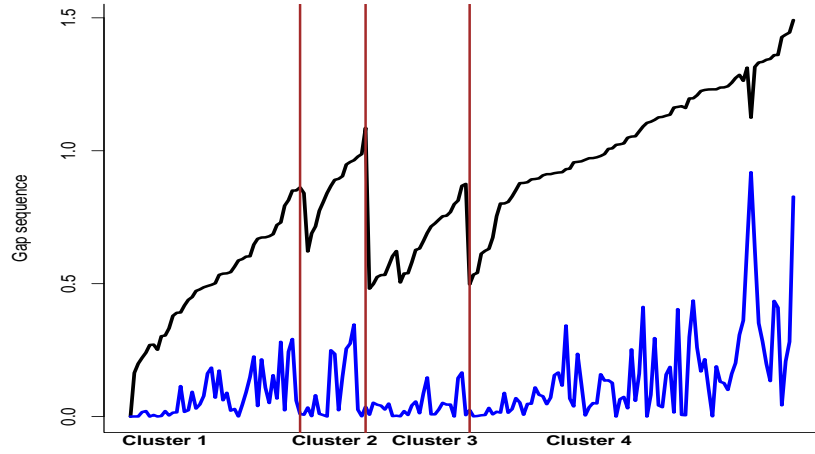


Figure 6: Gap sequence of the maximal distances (shown in black) and the corresponding minimal distances (shown in blue) with cluster separation thresholds shown in brown.

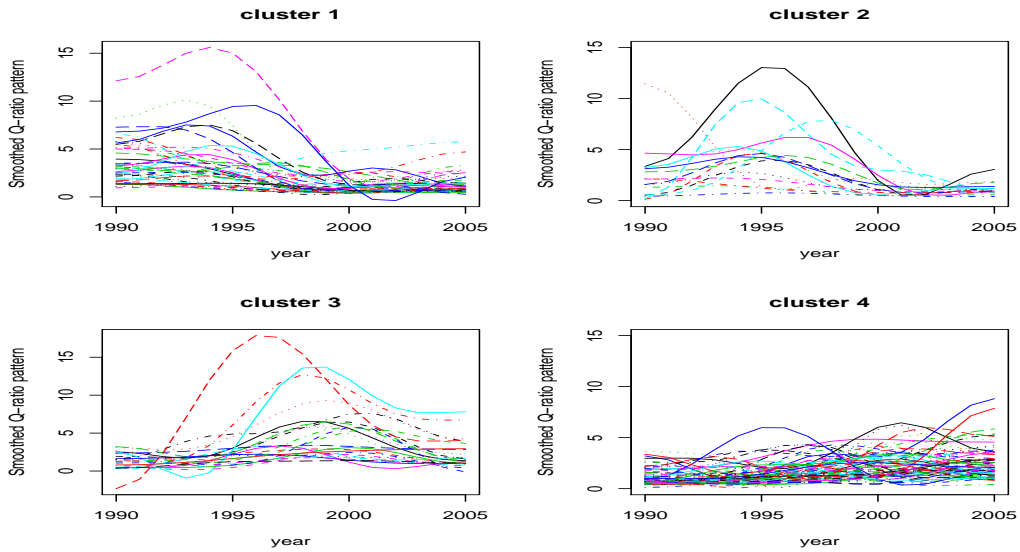


Figure 7: Clustering patterns as identified by clustering maximal and minimal distance.

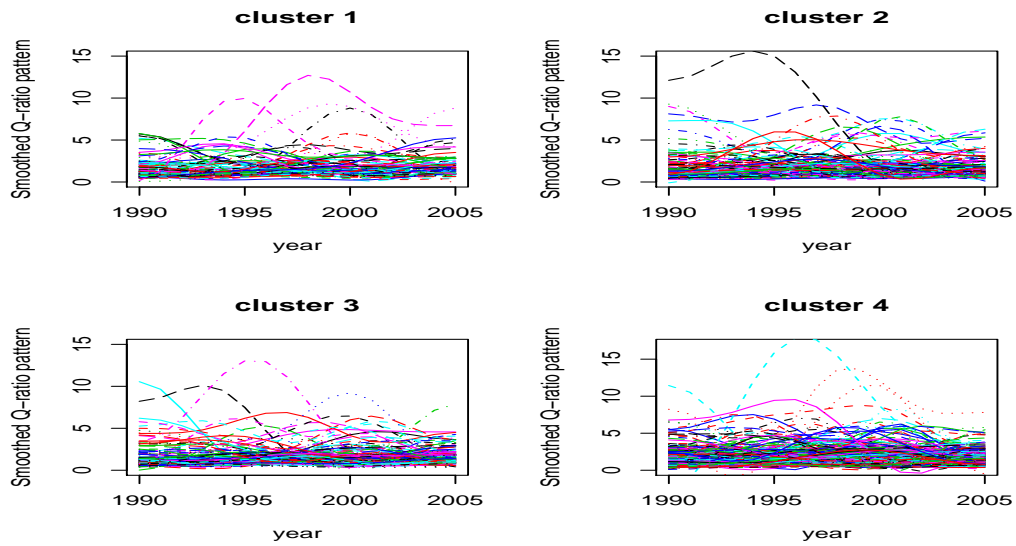


Figure 8: Clustering patterns as identified by the Bar Joseph et al (2001) method.

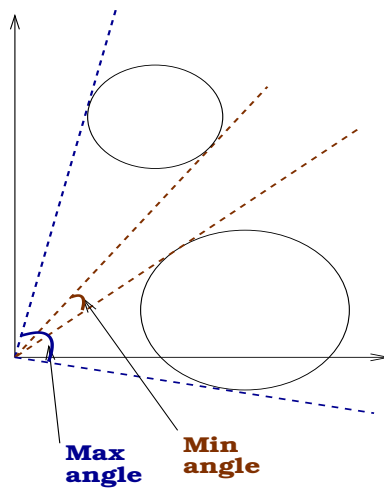


Figure 9: 2-dimensional example for maximum and minimum angle: Computing the maximal and minimal distances under correlation measure.