Means for Finding Meaningful Levels of a Hierarchical Sequence *Prior* to Performing a Cluster Analysis

David Allen Olsen
Department of Computer Science and Engineering, University of Minnesota-Twin Cities, Minneapolis, U.S.A.

Keywords: Intelligent Control Systems, Hierarchical Clustering, Hierarchical Sequence, Complete Linkage, Meaningful Level, Meaningful Cluster Set, Distance Graphs, Noise Attenuation.

Abstract: When the assumptions underlying the standard complete linkage method are unwound, the size of a hierarchical sequence reverts back from n levels to $\frac{n\cdot(n-1)}{2}+1$ levels, and the time complexity to construct a hierarchical sequence of cluster sets becomes $O(n^4)$. Moreover, the *post hoc* heuristics for cutting dendrograms are not suitable for finding meaningful cluster sets of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence. To overcome these problems for small-n, large-m data sets, the project described in this paper went back more than 60 years to solve a problem that could not be solved then. This paper presents a means for finding meaningful levels of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence *prior* to performing a cluster analysis. By finding meaningful levels of such a hierarchical sequence *prior* to performing a cluster analysis, it is possible to know which cluster sets to construct and construct only these cluster sets. This paper also shows how increasing the dimensionality of the data points helps reveal inherent structure in noisy data. The means is theoretically validated. Empirical results from four experiments show that finding meaningful levels of a hierarchical sequence is easy and that meaningful cluster sets can have real world meaning.

1 INTRODUCTION

Reasoning about hardware limitations while an application is being developed is a key aspect of computational thinking (Kirk and Hwu, 2013). This paper presents the second part of a three-part research project. The goal of this project was to develop a general, simplistic, complete linkage hierarchical clustering method that 1) substantially improves upon the accuracy of the standard complete linkage method and 2) can be fully automated or used with minimal operator supervision. The standard complete linkage method (Sorenson 1948) was the first of seven standard hierarchical clustering methods to be developed during the late 1940's to the mid-1960's (Everitt et al., 2011). At that time, clustering problems having about 150 data points were viewed as moderately-sized problems while problems having about 500 data points were viewed as large. *Cf.* (Anderberg, 1973).

To accommodate the hardware limitations of that time and solve these "large-scale" clustering problems, those who developed the standard hierarchical clustering methods made several assumptions. They assumed that cluster sets are nested partitions, i.e., that clusters are both indivisible and mutually exclusive (Jain and Dubes, 1988). Making this assumption reduces the size of a hierarchical sequence from $\frac{n\cdot(n-1)}{2}+1$ levels to n levels (Berkhin, 2006), where n is the number of data points in a data set. Further, the number of combinations that need to be examined at each level of the hierarchical sequence becomes much smaller than complete enumeration (Anderberg, 1973). Those who developed the standard hierarchical clustering methods also assumed that notions of distance between data points ("interpoint" distances) can be generalized to notions of distance between clusters of data points ("intercluster" distances). By making this assumption, proximity measures known as linkage metrics could be devised. Linkage metrics are used to combine clusters of data points or subdivide a cluster of data points at a time (Berkhin, 2006). Once the cluster sets of an n-level hierarchical sequence are constructed, a dendrogram is used to visually represent the hierarchical sequence, and *post hoc* heuristics for "cutting" dendrograms are used to find meaningful cluster sets. *See, e.g.,* (Jain and Dubes, 1988), (Johnson and Wichern, 2002), and (Everitt et al., 2011).

The above-described assumptions sacrifice accu-

Olsen D..
Means for Finding Meaningful Levels of a Hierarchical Sequence Prior to Performing a Cluster Analysis.
DOI: 10.5220/0005040600210033
In *Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2014), pages 21-33
ISBN: 978-989-758-039-0

racy for efficiency when the inherent (hierarchical) structure in a data set is not taxonomic. *See* (Lance and Williams, 1967), (Olsen, 2014). The standard complete linkage method has the following four weaknesses: First, when clusters are being combined or a cluster is being subdivided, the standard complete linkage method cannot resolve ties between intercluster distances. Consequently, either one of the distances is selected arbitrarily or alternative hierarchical sequences are constructed, and the results are no longer deterministic. Second, because the standard complete linkage method uses intercluster distances to construct clusters, does not allow clusters to overlap, and does not allow data points to migrate between clusters (Lance and Williams, 1967), cluster sets often are constructed inaccurately. Third, results obtained from the standard compete linkage method can depend on which end of a hierarchical sequence is treated as the beginning. Consequently, the dendrograms for agglomerative hierarchical clustering and divisive hierarchical clustering may be different, and finding the cause(s) for the difference is both inconvenient and time-consuming. Fourth, the standard complete linkage method does not find meaningful levels or meaningful cluster sets of hierarchical sequences[1]. It still is necessary to construct a dendrogram and determine where and how many times to cut the dendrogram, and *post hoc* heuristics are computationally expensive to run.

Because of these weaknesses, it can be difficult to interpret results obtained from the standard complete linkage method. Consequently, it is underutilized in automation and by intelligent control systems, including supervisory functions such as fault detection and diagnosis and adaptation. *Cf.* (Isermann, 2006). When the standard complete linkage method *is* used, stopping criteria often are used in place of *post hoc* heuristics. Stopping criteria are predetermined. If the model upon which they are based is inadequate or changes, the stopping criteria lose their usefulness. Moreover, the standard complete linkage method is an updating method, so it uses information from previously constructed cluster sets to construct subsequent cluster sets. It must construct the cluster set for every level of an n-level hierarchical sequence until the stopping criteria are met. *See, e.g.,* (Jain and Dubes, 1988), (Johnson and Wichern, 2002). These cluster sets must be either materially accurate or, if possible, amendable for material inaccuracies. *See, e.g.,* U.S. Patent No. 8,312,395 (defect identification in semiconductor production; operators must ensure that the results are 80 to 90 percent accurate). As much as 90 percent of the effort that goes into implementing the standard complete linkage method is used to develop stopping criteria or interpret results.

Notwithstanding these weaknesses, the standard complete linkage method is an important clustering method. The distributions of many real world measurements are bell-shaped, so the standard complete linkage method has broad applicability. Its simplicity makes it relatively easy to mathematically capture its properties. Of the standard hierarchical clustering methods, the standard complete linkage method is the only method that is invariant to monotonic transformations of the distances between the data points, that can cluster any kind of attribute, that is not prone to inversions, and that produces globular or compact clusters (Johnson and Wichern, 2002), (Everitt et al., 2011). Moreover, more sophisticated methods show no clear advantage for many purposes. Thus, the need exists to bring complete linkage hierarchical clustering over from the "computational side of things ... to the system ID/model ID kind of thinking" (Gill, 2011) as part of closing the loop on cyber-physical systems.

For the first part of the project, a new, complete linkage hierarchical clustering method was developed. *See* (Olsen, 2014). The new clustering method is consonant with the model for a measured value that scientists and engineers commonly use[2], so it substantially improves upon the accuracy of the standard complete linkage method. Further, it can construct cluster sets for select, possibly non-contiguous levels of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence. The new clustering method was designed with small-n, large-m data sets in mind, where n is the number of data points, m is the number of dimensions, and "large" means thousands and upwards (Murtagh, 2009).[3]

[1] A "meaningful cluster set" refers to a cluster set that can have real world meaning. Under ideal circumstances, a "meaningful level" refers to a level of a hierarchical sequence at which a new configuration of clusters has finished forming. These definitions appear to be synonymous for $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequences. The cluster set that is constructed for a meaningful level is a meaningful cluster set, so these terms are used interchangeably.

[2] The model for a measured value is measured value = true value + bias (accuracy) + random error (statistical uncertainty or precision) (Navidi, 2006). This model has substantially broader applicability than the taxonomic model that is the basis for the standard complete linkage method.

[3] These data sets are used by many cyber-physical systems and include time series. For example, a typical automobile has about 500 sensors; a small, specialty brewery has about 600 sensors; and a small power plant has about 1100 sensors. The new clustering method may accommodate large-n, large-m data sets as well, and future work includes using multicore and/or heterogeneous processors to parallelize parts of the new clustering method, but large-n, large-m data sets are not the focus here.

Because the computational power presently exists to apply hierarchical clustering methods to much larger data sets than before, the new clustering method unwinds the above-described assumptions. However, by unwinding these assumptions and letting the size of a hierarchical sequence revert back from n levels to $\frac{n\cdot(n-1)}{2}+1$ levels, the time complexity to construct cluster sets becomes O(n^4). This is large even for small-n, large-m data sets. Moreover, the *post hoc* heuristics for cutting dendrograms are not suitable for finding meaningful cluster sets of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence. For example, in (Tibshirani et al., 2001), Tibshirani et al. present a gap statistic for determining an "optimal" number of clusters for a data set and use this technique to determine where to cut a dendrogram. Because the technique selects the number of clusters from a range of numbers, a range of cluster sets must be constructed as opposed to constructing only select cluster sets. Like other *post hoc* heuristics, *see, e.g.,* (Kim and Lee, 2000), (Daniels and Giraud-Carrier, 2006), the gap statistic is designed to find only one or maybe a few cluster sets. Further, it is not designed for hierarchical sequences where clusters are not well-separated but close together or overlap.

Thus, with today's technology, the project went back more than 60 years to solve a problem that could not be solved then. For the second part of the project, a means was developed for finding meaningful levels of an $\frac{n\cdot(n-1)}{2}+1$-level (complete linkage) hierarchical sequence *prior* to performing a cluster analysis. By finding meaningful levels of such a hierarchical sequence *prior* to performing a cluster analysis, it is possible to know which cluster sets to construct and construct only these cluster sets. This reduces the time complexity to construct cluster sets from $O(n^4)$ to $O(ln^2)$, where l is the number of meaningful levels. *These are the cluster sets that can have real world meaning.* It is notable that the means does not depend on dendrograms or *post hoc* heuristics to find meaningful cluster sets. The second part also looked at how increasing the dimensionality of the data points helps reveal inherent structure in noisy data, which is necessary for finding meaningful levels.

2 OTHER RELATED WORK

Researchers have avoided developing clique detection methods for hierarchical clustering, and at least one researcher has specifically taught away from using these methods (Jain and Dubes, 1988) (citing (Matula, 1977)). In (Peay, 1974) and (Peay, 1975), E.R. Peay presents a linkage-based clique detection method and applies the method to hierarchical clustering. For each level of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence for which a clique set is constructed, Peay's clique detection method recognizes every maximally complete subset of data points as a clique, including those from which the data points migrate. Because Peay's clique detection method is an updating method, it also constructs a clique set for every level of such a hierarchical sequence. It cannot construct only the clique sets that correspond to meaningful levels of a hierarchical sequence. A similar problem holds for flat clique detection methods. Without knowing which levels of a hierarchical sequence are meaningful, flat methods are ineffective.

Within a framework based on ultrametric topology and ultrametricity, F. Murtagh, in (Murtagh, 2009), observes that it is easier to find clusters in sparse or high dimensional spaces. This work does not describe how to find meaningful levels of a hierarchical sequence. Also, it assumes that the mean values and the standard deviations of all the dimensions of a data point are the same.

3 NOISE ATTENUATION

The means for finding meaningful levels is based on two assumptions. Let $\mathcal{X}=\{x_1,x_2,...,x_n\}$ be a data set that contains a finite number of data points n, where each data point has m dimensions. Further, suppose that each data point is a sequence of samples and that at any moment in time, with respect to each class or source, all the samples have the same true values and biases[4]. First, the means assumes that noise (random error) is the only random component in a measured value, that noise can be modeled as Gaussian random variables, and that the noise that is embedded in each dimension (sample) of each data point is statistically independent. Second, the means assumes that the dissimilarities between the data points are non-negative values. This latter assumption is needed because p-norm distance measures do not distinguish between positive and negative correlation.

Within the context of the nearest neighbor problem for database search, where high(er) dimensionality is considered to be a curse, Beyer et al., in (Beyer et al., 1998), show that under broadly applicable conditions, if

$$\lim_{m\to\infty} Var[\frac{\|Y_m\|^p}{E[\|Y_m\|^p]}]=0, \quad (1)$$

[4]In real world terms, this is the same as calibrating the sensors.

PAIRS	DIM	MEAN1	MEAN2	STD1	STD2	DMIN Normal	STDDIST Normal	Limit	DMIN Uniform	STDDIST Uniform
1000	10	2	2	2	200	200.6-256.1	131.7-143.5	141.4	490.0-553.1	158.5-163.8
1000	10	2	2	200	2	164.8-282.5	138.5-145.1	141.4	536.1-556.5	153.9-163.8
1000	10	2	2	20	20	20.9-36.4	19.2-20.2	20.0	47.5-65.2	28.8-30.0
1000	10	2	2B	20	20	$6.324\text{x}10^{9}$	27.9-29.0	28.2	$6.324\text{x}10^{9}$	47.7-49.2
1000	10	2	2B	200	200	$6.324\text{x}10^{9}$	278.1-291.8	282.8	$6.324\text{x}10^{9}$	464.3-497.0
1000	10	2	2B	2	200	$6.324\text{x}10^{9}$	195.8-207.0	200.0	$6.324\text{x}10^{9}$	337.6-352.3
1000	10	2	2B	200	2	$6.324\text{x}10^{9}$	198.2-204.9	200.0	$6.324\text{x}10^{9}$	335.4-351.1
1000	100	2	2	2	200	1507.8-1604.7	138.3-144.7	141.4	2891.3-3017.7	148.8-157.7
1000	100	2	2	200	2	1489.1-1603.1	137.5-144.8	141.4	2862.4-2919.1	153.4-155.7
1000	100	2	2	20	20	210.2-224.3	19.2-20.3	20.0	375.6-408.0	28.3-29.4
1000	100	2	2B	20	20	$2.000\text{x}10^{10}$	27.8-29.4	28.2	$2.000\text{x}10^{10}$	47.9-50.0
1000	100	2	2B	200	200	$2.000\text{x}10^{10}$	278.0-296.5	282.8	$2.000\text{x}10^{10}$	483.5-493.3
1000	100	2	2B	2	200	$2.000\text{x}10^{10}$	195.4-203.9	200.0	$2.000\text{x}10^{10}$	336.3-346.9
1000	100	2	2B	200	2	$2.000\text{x}10^{10}$	197.8-203.9	200.0	$2.000\text{x}10^{10}$	335.3-355.2
1000	1000	2	2	2	200	5756.0-5901.9	137.5-144.4	141.4	10,417.0-10,503.0	147.0-157.2
1000	1000	2	2	200	2	5772.8-5933.9	138.8-145.4	141.4	10,429.0-10,554.0	147.4-161.6
1000	1000	2	2	20	20	828.4-834.4	19.4-20.7	20.0	14,301.0-14,655.0	28.2-30.7
1000	1000	2	2B	20	20	$6.324\text{x}10^{10}$	27.4-28.7	28.2	$6.324\text{x}10^{10}$	47.5-50.8
1000	1000	2	2B	200	200	$6.324\text{x}10^{10}$	283.5-297.4	282.8	$6.324\text{x}10^{10}$	476.7-506.9
1000	1000	2	2B	2	200	$6.324\text{x}10^{10}$	194.8-202.7	200.0	$6.324\text{x}10^{10}$	335.3-353.2
1000	1000	2	2B	200	2	$6.324\text{x}10^{10}$	194.7-203.7	200.0	$6.324\text{x}10^{10}$	335.7-358.2

PAIRS = Number of data point pairs
DIM = Number of dimensions in each data point
MEAN1 = True value plus bias of first data point dimensions
MEAN2 = True value plus bias of second data point dimensions
STD1 = Std. dev. of noise embedded in first data point dimensions
STD2 = Std. dev. of noise embedded in second data point dimensions

DMIN Normal = DMIN when noise is normally distributed (5 trials)
STDDIST Normal = Std. devs. of the distance distributions (5 trials) when noise is normally distributed
LIMIT = Limit calculations using Eq. 5
DMIN Uniform = DMIN when noise is uniformly distributed (5 trials)
STDDIST Uniform = Std. devs. of the distance distributions (5 trials) when noise is uniformly distributed

Figure 1: Exemplary results from a sensitivity analysis. The minimum distance and the maximum distance (not shown) between data points from two different classes are calculated. Limits calculated with Equation 5 are very consistent with the empirical results for STDDIST Normal. When noise is uniformly distributed, the results are analogous to those when noise is normally distributed, indicating that the Gaussian random variable assumption is reasonable.

then for every $\varepsilon > 0$,

$$\lim_{m\to\infty} Prob[DMAX_m^p \leq (1+\varepsilon)DMIN_m^p] = 1. \quad (2)$$

Y_m is the difference between any independent data point $P_{i,m}, i = 1,2,...,n$, and Q_m, a query point that is chosen independently of all the data points; m is the dimensionality of $P_{i,m}$ and Q_m; $DMAX$ is the distance between Q_m and the farthest away data point; $DMIN$ is the distance between Q_m and the nearest data point; and p is the index of the p-norm distance measure. In (Hinneburg et al., 2000), Hinneburg et al. extend this work by showing that

$$\lim_{m\to\infty} E[\frac{DMAX_m^p - DMIN_m^p}{m^{1/p-1/2}}] = C_p, \quad (3)$$

or

$$\lim_{m\to\infty} E[DMAX_m^p - DMIN_m^p] = C_p \cdot (m^{1/p-1/2}). \quad (4)$$

C_p is a constant that depends on p.

For the purposes of cluster analysis, these equations hint that classes of noisy data points may be spatially separable. However, they do not show how the distances between data points from different classes ("interclass" distances) relate to the distances between data points that belong to the same class ("intraclass" distances). Also, C_p is unknown. A set of theorems was proved to provide the missing pieces. Theorem 1, below, pertains to the 2-norm distance measure. Here, although it can have a much broader scope, it is written specifically for Euclidean distance. Since statistical independence is assumed only with respect to the Gaussian random variables (noise), the mean values (true values plus biases) may be highly correlated.

Theorem 1. *Let C_1 and C_2 be two clusters, each of which is comprised of a finite set of data points, i.e., $C_1 = \{x_{1,1}, x_{1,2}, ..., x_{1,n_1}\}$ and $C_2 = \{x_{2,1}, x_{2,2}, ..., x_{2,n_2}\}$. Let each data point have m dimensions, each of which is a statistically independent, Gaussian random variable, i.e., $X_{1,i,k} \sim N(\mu_{1,i,k}, \sigma^2_{1,i,k})$ and $X_{2,j,k} \sim N(\mu_{2,j,k}, \sigma^2_{2,j,k}), i = 1,2,...,n_1, j = 1,2,...,n_2$, and $k = 1,2,...,m$. When $Y_{k,(i,j)} = (X_{1,i,k} - X_{2,j,k})$, $Y_{k,(i,j)} \sim N(\mu_{k,(i,j)}, \sigma^2_{k,(i,j)})$. If $\sigma_{k,(i,j)}$ is bounded from below by $\varepsilon > 0$ and above by a constant S, and if $|\mu_{k,(i,j)}|$ is bounded from above by a constant M, then as $m \to \infty$, the variance $\sigma^2_{Z_{m,(i,j)}}$ of the random variable $Z_{m,(i,j)} = (\sum_{k=1}^m Y^2_{k,(i,j)})^{\frac{1}{2}}$ converges to $\frac{\sum_{k=1}^m \sigma^4_{k,(i,j)}}{2(\sum_{k=1}^m \sigma^2_{k,(i,j)} + \sum_{k=1}^m \mu^2_{k,(i,j)})} + \frac{\sum_{k=1}^m \sigma^2_{k,(i,j)} \mu^2_{k,(i,j)}}{\sum_{k=1}^m \sigma^2_{k,(i,j)} + \sum_{k=1}^m \mu^2_{k,(i,j)}}$.*

Proof for Theorem 1. *A sketch of the proof is in the Appendix to this paper.*

For $Y_{k,(i,j)} \sim N(0,1)$, $k = 1,2,...,m$,

$\lim_{m\to\infty}\sigma^2_{Z_{m,(i,j)}} = \frac{1}{2}$. For $Y_{k,(i,j)} \sim N(0,\sigma^2_{k,(i,j)})$ where $\sigma_{k,(i,j)} = \sigma_{(i,j)},\ k = 1,2,...,m,\ \lim_{m\to\infty}\sigma^2_{Z_{m,(i,j)}} = \frac{1}{2}\sigma^2_{(i,j)}$. When $\sigma_{k,(i,j)}$ and $\mu_{k,(i,j)}$ are chosen from uniform distributions, the Monte Carlo method shows that the limit in Theorem 1 converges from below to $\frac{m}{3}S^2$ as the bound M on $\mu_{k,(i,j)}$ increases. As m increases, the standard deviation of this number becomes smaller relative to its magnitude. When $\sigma_{k,(i,j)} = \sigma_{(i,j)},\ k = 1,2,...,m$, the Monte Carlo method shows that the limit in Theorem 1 converges from below to mS^2 as the bound M on $\mu_{k,(i,j)}$ increases. The standard deviation of this number decreases to zero absolutely. When $\sigma_{k,(i,j)} = \sigma_{(i,j)}$ and $\mu_{k,(i,j)} = \mu_{(i,j)}$, $k = 1,2,...,m$, the result in Theorem 1 becomes

$$\lim_{m\to\infty}\sigma^2_{Z_{m,(i,j)}} = \frac{\sigma^2_{(i,j)}}{2(1+\frac{\mu^2_{(i,j)}}{\sigma^2_{(i,j)}})} + \frac{\mu^2_{(i,j)}}{(1+\frac{\mu^2_{(i,j)}}{\sigma^2_{(i,j)}})}. \quad (5)$$

If $\sigma_{(i,j)}$ is held constant and $\mu_{(i,j)}$ is allowed to vary between 0 and $|\mu_{(i,j)}| \gg \sigma_{(i,j)}$, $\sigma_{Z_{m,(i,j)}}$ is a constant between $\frac{\sigma^2_{(i,j)}}{2}$ and $\sigma^2_{(i,j)}$. The graph for the first term in Equation 5 is monotonically decreasing while that for the second term is monotonically increasing. Moreover, as Fig. 1 shows, limits calculated with Equation 5 are very consistent with the empirical results from a sensitivity analysis.

4 FINDING MEANINGFUL LEVELS AND CLUSTER SETS

Often, as the dimensionality of the data points increases and the 2-norm interclass distances become larger, the standard deviations of the 2-norm interclass distances, i.e., $\sigma_{Z_{m,(i,j)}}$, nonetheless remain relatively small or constant. When $\sigma_{k,(i,j)} = \sigma_{(i,j)}$ and $\mu_{k,(i,j)} = \mu_{(i,j)}$, $k = 1,2,...,m$, this is certainly so, because Equation 5 shows that $\sigma_{Z_{m,(i,j)}}$ is a constant. In particular, when the distribution of the noise that is embedded in each dimension of each data point does not change, $\sigma_{Z_{m,(i,j)}}$ is a constant between $\frac{\sigma_{(i,j)}}{\sqrt{2}}$ and $\sigma_{(i,j)}$. As the Monte Carlo simulations show, this also is so when the 2-norm interclass distances grow at an expected rate that is much faster than $\frac{d(\sqrt{mS^2})}{dm} = \frac{S}{2\sqrt{m}}$.

When this scenario holds, data points that belong to the same class link at about the same time *even at higher dimensionalities*. As Fig. 2(a) depicts, classes of data points can be close together at lower dimensionalities. When they are, the magnitudes of many intraclass distances and interclass distances are about the same, so the two kinds of distances commingle. However, as Fig. 2(b) depicts, the classes of data points are farther apart at higher dimensionalities, so the intraclass distances and the interclass distances segregate into bands. Thus, higher dimensionalities can attenuate the effects of noise[5] that preclude finding meaningful levels of a hierarchical sequence at lower dimensionalities and distinguish between the classes. Moreover, as Figs. 2(b) and (c) show, this pattern repeats itself as clusters become larger from including more data points.

Consequently, as the dimensionality of the data points increases, the distance graphs for a data set can exhibit identifiable features that correlate with meaningful levels of the corresponding hierarchical sequences. *These levels are the levels at which multiple classes have finished linking to form new configurations of clusters.* In particular, assuming that the data set has inherent structure, a distance graph takes on a shape whereby sections of the graph run nearly parallel to one of the graph axes. Where there is very little or no linking activity, the sections run nearly vertically. Where there is significant activity, i.e., where new configurations of clusters are forming, the sections run nearly horizontally. Thus, portions of the graph that come after the lower-right corners and before the upper-left corners indicate where new configurations of clusters have finished forming. As the schematic in Fig. 2(c) shows, a distance graph can be visually examined *prior* to performing a cluster analysis. Since a distance graph is used to find meaningful levels of a hierarchical sequence *prior* to performing a cluster analysis, it is not a summary of the results obtained from the analysis. Instead, it enables a user to selectively construct only meaningful cluster sets, i.e., cluster sets where new configurations of clusters have finished forming.

Finding meaningful levels is remarkably easy:

Step 1. Calculate the dissimilarities between data points x_i and x_j in data set $\mathcal{X}$, $i,j = 1,2,...,n, x_i \neq x_j$. Then, calculate the lengths or magnitudes of the vectors that contain the dissimilarities between the data points. Here, the dissimilarity measures are simple value differences, and the 2-norm is used to obtain Euclidean distance.

Step 2. Construct ordered triples $(d_{i,j},i,j)$ from these distances and the indices of the respective data points, sort the ordered triples into rank or ascending order according to their distance elements, and assign indices to the sorted ordered triples (the "rank order indices"). The time complexity to calculate the distances is $O(\frac{n\cdot(n-1)\cdot m}{2})$. If ordinary merge sort

[5] Attenuating the effects of noise refers to reducing the effects of noise on cluster construction.

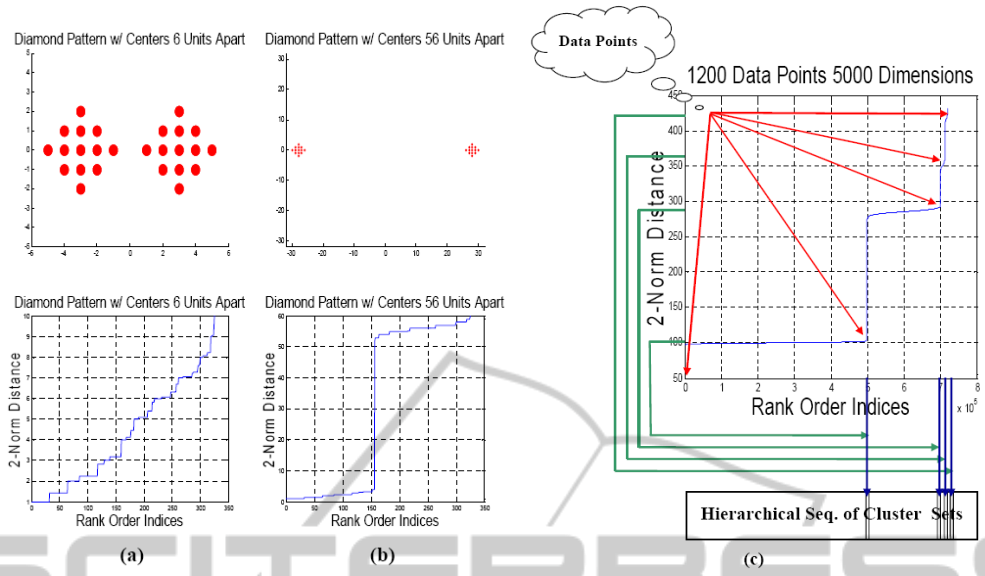


Figure 2: Simple illustration that shows how two classes of data points link as the distance between the classes increases (left) and schematic for finding meaningful levels of a hierarchical sequence (right). Inherent structure is revealed through identifiable features of the distance graph. These features correlate with those levels of the corresponding hierarchical sequence at which multiple classes have finished linking to form new configurations of clusters.

Rank Order Index =	Ordered Triple =	Hierarchical Level =	Threshold Distance d' =	Cluster Set
36	3136.20 4 8	n(n-1)/2 = 36	3136.20	{1,2,3,4,5,6,7,8,9} *
34	2958.51 4 6	n(n-1)/2 - 2 = 34	2958.51	{1,3,5,6,7,8,9}, {1,2,3,4,5,6,7,9}
27	2488.62 3 7	n(n-1)/2 - 9 = 27	2488.62	{1,3,5,6,7,8}, {2,3,4,5,7,9} *
12	1943.70 5 6	n(n-1)/2 - 24 = 12	1943.70	{5,7}, {2,4,9}, {1,6,8}, {1,3,5,6}
6	287.96 1 6	6	287.96	{1,6,8}, {2,4,9}, 3, 5, 7 *
5	277.63 4 9	5	277.63	{2,4,9}, {1,8}, {6,8}, 3, 5, 7
4	277.00 1 8	4	277.00	{1,8}, {2,4}, {4,9}, {6,8}, 3, 5, 7, 9
3	272.98 6 8	3	272.98	{1,8}, {2,4}, {6,8}, 3, 5, 7, 9
2	191.62 2 9	2	191.62	{2,4}, {6,8}, 1, 3, 5, 7, 9
1	157.97 2 4	1	157.97	{2,4}, 1, 3, 5, 6, 7, 8, 9
		0	0.00	1, 2, 3, 4, 5, 6, 7, 8, 9 *

Figure 3: Illustration that shows how rank order indices and distance elements align with levels and the respective threshold distances d' of the corresponding hierarchical sequence. The data come from the nine motes experiment described in Subsection 5.3. The 2-norm distance measure was used to calculate the distances. The arrow in the column for the threshold distance signifies that threshold distance d' is a continuous variable. The meaningful cluster sets in the last column have asterisks.

(Cormen et al., 2004) is used, the time complexity to sort the ordered triples is $O(\frac{n\cdot(n-1)}{2} \cdot log(\frac{n\cdot(n-1)}{2})) = O(n\cdot(n-1)\cdot log(\frac{n\cdot(n-1)}{2})^{\frac{1}{2}})$.

Step 3. Use the rank order indices and the ordered triples to construct a distance graph. The distance graph will remain smooth, regardless of the dimensionality of the data points, when inherent structure is absent. Assuming that the data set has inherent structure, increase the dimensionality of the data points and repeat Steps 1 to 3 until the lower-right corners have good definition (or as good as is practically possible).

Step 4. Along the axes of the distance graph, locate the rank order indices and/or the distance elements that correspond to where the lower-right corners appear in the graph. Under ideal circumstances, these corners are nearly orthogonal. The rank order indices and the distance elements coincide with the meaningful levels and the respective threshold distances d' of the corresponding hierarchical sequence.

For an example that shows how these four variables align, see Fig. 3. As part of the cluster analysis described in (Olsen, 2014), the ordered triples are evaluated in ascending order for information about linkage. As the distance elements become larger, threshold distance d' increases implicitly from 0 to

the maximum of all the distance elements. Although threshold distance d' is a continuous variable that can vary from 0 (where each data point is a singleton) to at least this maximum distance (where all the data points belong to the same cluster), the only values that matter are those $\frac{n\cdot(n-1)}{2}$ values that are equal to the distance elements $d_{i,j}$. Since the number of data points in a data set is finite, the maximum number of levels of a hierarchical sequence is finite and equal to the number of ordered triples (distance elements) plus one. Thus, the rank order indices, by virtue of the distance elements $d_{i,j}$, coincide with the last $\frac{n\cdot(n-1)}{2}$ levels of the hierarchical sequence.

Step 5. Use a complete linkage hierarchical clustering method such as that in (Olsen, 2014) to construct only the cluster sets for the meaningful levels. By using a method that constructs cluster sets *de novo* instead of using an updating method, it is possible to construct only the cluster sets for meaningful levels of a hierarchical sequence. The number of clusters in a meaningful cluster set becomes an artifact of cluster set construction. Finding meaningful levels of a hierarchical sequence reduces the time complexity to construct cluster sets from $O(n^4)$ to $O(ln^2)$.

5 EMPIRICAL RESULTS

The remainder of this paper describes the empirical results from four experiments. The first experiment shows how the effects of noise are attenuated as the dimensionality of the data points increases. The second experiment looks at data sets having multiple attributes and how meaningful cluster sets can have real world meaning. The third experiment also shows how meaningful cluster sets can have real world meaning. The fourth experiment demonstrates other interesting properties of the means for finding meaningful levels of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence. The data sets are representative of other data sets that have inherent structure. The 2-norm distance measure (Euclidean distance) is used to calculate the distances. *level* is a variable that is used to refer to individual meaningful levels, and d' refers to the respective threshold distances d'.

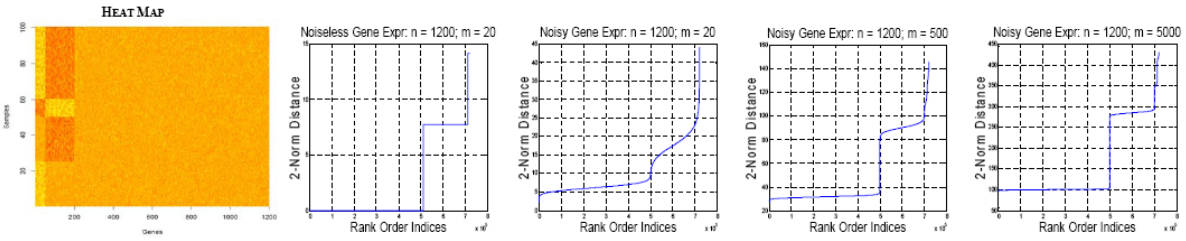
5.1 Synthetic Data Sets–Nearly Ideal Circumstances

This experiment shows how the effects of noise are attenuated and inherent structure emerges as the dimensionality of the data points increases. The heat map in Fig. 4 was provided by the Hollings Cancer Center at the Medical University of South Carolina. The data sets constructed from this heat map include three gene classes and four sample classes. The ratio for the gene classes is 50:150:1000 while the ratio for the sample classes is 25:25:10:40. The signal-to-noise ratio for the gene classes is 1.29/1.87, where noise is defined as the pooled estimate of the standard deviations for over ($N(2,4^2)$, mostly in red-orange (dark gray)), under ($N(-2,4^2)$, mostly in yellow (light gray)), and normally ($N(0,1^2)$, mostly in orange (medium gray)) expressed genes.

The mean values of the three gene classes are used to construct a noiseless data set. As the first graph in Fig. 4 shows, inherent structure emerges immediately for noiseless data. For the noisy data set, inherent structure emerges as early as m = 5000 dimensions, and the last graph suggests that the corresponding hierarchical sequence has five meaningful levels: *level* = 0 or d' = 0.00, *level* = 499,500 or d' = 105.28, *level* = 699,500 or d' = 297.65, *level* = 711,900 or d' = 365.58, and *level* = 719,400 or d' = 429.81. The cluster sets for these levels were constructed without constructing any of the other 719,396 cluster sets (which also is 1195 fewer cluster sets than an n-level hierarchical sequence). The gene classes are discernible by examining the meaningful cluster sets. The two tables in Fig. 4 show that noise attenuation is not the same as noise elimination.

5.2 Residential Heat Pump

This experiment looks at data sets that have multiple attributes. Three data sets were provided by the U.S. National Institute of Standards and Technology (NIST). The data sets originally were collected for a study described in (Kim et al., 2006). There, they were used to analyze the performance of a residential heat pump that was operating in the cooling mode when a single external fault was imposed. The data sets are comprised of numerous kinds of measurements that were collected at approximately 12 second intervals for at least 17 minutes. While two of the data sets were collected, the indoor air side flow rate ($ft.^3$/min., scfm) was changed from 1000 scfm to 500 scfm and from 1000 scfm to 1200 scfm, respectively. Using no-fault, third-order polynomial correlations as the basis for calculating residuals, the readings for the most informative seven kinds of measurements related to air flow are excerpted from each data set, and consecutive sequences of readings that include 15 consecutive time points are concatenated to construct data points. In all, 11 data points having 105 dimensions (7 measurements x 15 time points) are constructed from each data set, or 33 data points



Noiseless Data n = 1200; m = 20

Hierarchical Level	2-Norm Threshold Distance	Cluster Set
0	0.00	1, 2, …, 1200
511,900	0.00	{1-50}, {51-200}, {201-1200}
711,900	7.75	{1-50,201-1200}, {51-200,201-1200}
719,400	14.14	{1-1200}

Noisy Data n = 1200; m = 5000

Hierarchical Level	2-Norm Threshold Distance	Cluster Set
0	0.00	1, 2, …, 1200
499,500	105.28	1, 2, …, 200, {201-1200}
699,500	297.65	{1,201-1200}, {2,201-1200}, …, {200,201-1200}
711,900	365.58	{1-50,201-1200}, {51-200,201-1200}
719,400	429.81	{1-1200}

Figure 4: Heat map, distance graphs, and meaningful cluster sets for the synthetic data sets. The first graph pertains to the noiseless data set while the next three graphs pertain to the noisy data set.

in total.

As the chart in Fig. 5 shows, the standard deviations for all the measurements are relatively small. Consequently, inherent structure emerges as early as $m = 105$ dimensions. The graph in Fig. 5 suggests that the corresponding hierarchical sequence has five meaningful levels. The fault pattern for the 500 scfm data appears at $level = 407$ or $d' = 40.10$ while that for the 1200 scfm data appears at $level = 286$ or $d' = 8.28$.

5.3 Motes Sensing Luminescence

This experiment shows that meaningful cluster sets can have real world meaning while other cluster sets generally do not. Nine Crossbow® MicaZ motes with MTS300CA sensor boards attached thereto are configured into a 1x1 meter grid. The motes are programmed to take light readings (lux) of an overhead light source every 1 second. After calibrating the motes, canopies are placed over motes 1, 6, and 8 during the entire experiment, so they are never exposed to direct light (the "full shade" motes); canopies are never placed over motes 2, 4, and 9, so they are always exposed to direct light (the "full sun" motes); and canopies are placed over motes 3, 5, and 7 for 1.5 minutes out of every 3 minute cycle (collectively, the "partial shade" motes). Further, the canopy for mote 3 is deployed at 30 seconds into each 3-minute cycle and removed at 120 seconds, the canopy for mote 5 is deployed at 60 seconds and removed at 150 seconds, and the canopy for mote 7 is deployed at 90 seconds and removed at 180 seconds. Data were collected for 15 minutes or 900 samples per mote (8100 samples in total), out of which 893 samples per mote (8037 samples in total) were usable[6].

Typical direct light readings were about 905 lux while typical indirect light readings were about 813 lux. The standard deviations of the readings collected by each mote are all less than 10 lux, so although some corners of the distance graph are not nearly orthogonal, inherent structure emerges as early as $m = 180$ dimensions. The graphs in Fig. 6 suggest that the corresponding hierarchical sequence has four meaningful levels. At $level = 6$ or $d' = 287.97$ ($m = 893$), the cluster set includes five non-overlapping clusters, one for the full sun motes, another for the full shade motes, and one for each of the partial shade motes. At $level = 27$ or $d' = 2488.63$ ($m = 893$), the cluster set includes two overlapping clusters, one for those motes that were exposed to direct light during all or part of the experiment (the full sun motes and the partial shade motes) and the other for those motes that were not exposed to direct light during all or part of the experiment (the full shade motes and the partial shade motes).

As the table in Fig. 6 illustrates, the cluster sets for the meaningful levels have real world meaning. The cluster sets for the other levels generally do not, and the more so for levels that are not proximate to the meaningful levels. When multiple classes of data points have not finished linking to form a new configuration of clusters, the cluster sets are comprised of overlapping clusters whose differences are not related to inherent structure. These cluster sets are much less transparent to domain experts.

[6]Seven packets from mote 9 were dropped during transmission.

Measurement	1000 scfm (Normal) Mean	Std. Dev.	500 scfm Mean	Std. Dev.	1200 scfm Mean	Std. Dev.
Thermopile Indoor Air Temperature Change (F)	18.19	0.079	25.28	0.082	16.80	0.066
Compressor Discharge Line Wall Temperature (F)	157.63	0.153	157.67	0.213	157.78	0.153
Indoor Vapor Temperature at Saturation (F)	50.15	0.125	43.57	0.113	51.18	0.129
Indoor Coil Liquid Subcooling (F)	5.62	0.159	5.25	0.120	5.62	0.173
Outdoor Liquid Line Subcooling (F)	8.39	0.155	7.33	0.138	8.47	0.164
Outdoor Inlet Refrigerator Vapor Temperature at Saturation (F)	102.92	0.103	100.75	0.086	103.28	0.101
Compressor Suction Superheat (F)	20.91	0.233	20.32	0.311	21.16	0.202

n = 33;	m = 105	
Hierarchical Level	2-Norm Threshold Distance	Cluster Set
0	0.00	1, 2, …, 33
165	4.06	{1-11}, {12-22}, {23-33}
286	8.28	{1-11,23-33}, {12-22}
407	40.10	{1-11,23-33}, {1-11,12-22}
528	46.42	{1-33}

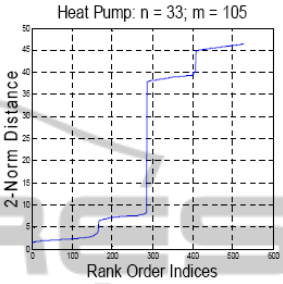
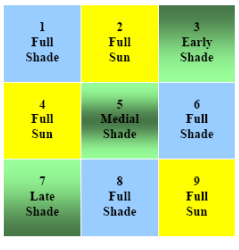
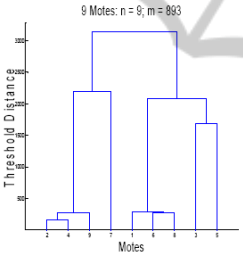
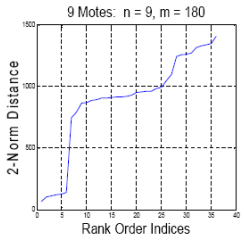
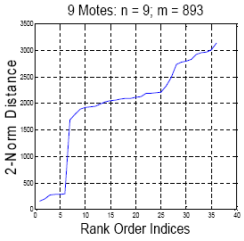


Figure 5: Mean values and standard deviations, distance graph, and meaningful cluster sets for the seven kinds of measurements that were excerpted from the NIST data sets.









n = 9;	m = 893	
Hierarchical Level	2-Norm Threshold Distance	Cluster Set
0	0.00	1, 2, 3, 4, 5, 6, 7, 8, 9*
1	157.98	{2,4}, 1, 3, 5, 6, 7, 8, 9
5	277.63	{2,4,9}, {1,8}, {6,8}, 3, 5, 7
6	287.97	{1,6,8}, {2,4,9}, 3, 5, 7*
7	1681.72	{1,6,8}, {2,4,9}, {3,5}, 7
12	1943.70	{5,7}, {2,4,9}, {1,6,8}, {1,3,5,6}
18	2088.70	{1,3,5,6,8}, {2,4,9}, {5,7}
24	2199.83	{1,3,5,6,8}, {2,3,4,5,9}, {2.4.5.7.9}, {1,5,7}
26	2322.14	{1,3,5,6,8}, {1,5,6,7,8}, {2,3,4,5,9}, {2.4.5.7.9},
27	2488.63	{1,3,5,6,7,8}, {2,3,4,5,7,9}*
28	2731.74	{1,3,5,6,7,8}, {2,3,4,5,7,9}
30	2791.14	{1,3,5,6,7,8}, {2,3,4,5,7,9}
35	3005.77	{1,2,3,5,6,7,8,9}, {1,2,3,4,5,6,7,9}
36	3136.20	{1,2,3,4,5,6,7,8,9}*

Figure 6: Configuration, dendrogram, distance graphs, and exemplary cluster sets for the nine motes data set. The motes are classified according to the data sequences that are collected. The different colors (gray scales) represent the different clusters at *level* = 6. The meaningful cluster sets have asterisks.

The number of meaningful levels does not appear to be a limiting factor. In experiments involving complex geometric patterns, as many as 19 meaningful levels have been found. In contrast, the *post hoc* heuristics are designed to find one or maybe a few cluster sets. The gap statistic found the cluster set at *level* = 6 but not that at *level* = 27, because the latter cluster set includes overlapping clusters.

When the standard complete linkage method is used to cluster a data set, some cluster sets that should be meaningful are obscure. As the dendrogram in Fig. 6 shows, while mote 7 combines with the full sun motes at $d' = 2488.63$, motes 3 and 5 combine with the full shade motes. This disparity among the partial shade motes is difficult to understand without taking into consideration how the standard complete linkage method imposes taxonomic structure onto data sets.

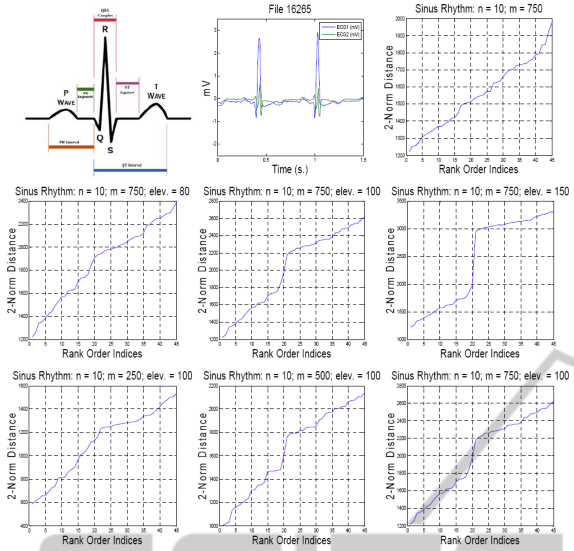


Figure 7: ECG and distance graphs for the data that are excerpted from file 16265 of the MIT-BIH Normal Sinus Rhythm database.

5.4 Health Monitoring

The data used in this experiment come from file 16265 of the MIT-BIH PhysioNet Normal Sinus Rhythm database (Goldberger et al., 2000). This file contains ECG readings collected at 128 hertz. The P,Q,R,S,T interval of each heart beat, illustrated by the first two graphs in the top row of Fig. 7, describes how a heart pumps blood to other parts of a body. Here, 25 samples per beat that include the Q,R,S complex and at least the left side of the ST element are excerpted from the first 300 consecutive beats of the file, and the data set is divided into ten segments (approx. 25 seconds each). The last graph in the first row of Fig. 7 shows that this data set has almost no inherent structure.

An elevating ST element is simulated by adding a constant c_{elevST} to samples 11-22 of the excerpts in the last five segments. Increasing c_{elevST} from 10 mV to 150 mV adds structure to the data set. The graphs in the second row show that the elevating ST is detectable as early as 80 mV, when the first five segments and the last five segments are grouped into different clusters. The graphs in the bottom row show how inherent structure emerges as the dimensionality of the segments increases. Increasing the dimensionality of the segments does not add structure to the data set, however, and the law of diminishing returns eventually sets in. At these elevations, the damage from ischemia and the risk of sudden death still are low.

6 CONCLUSION

When the assumptions underlying the standard complete linkage method are unwound, the size of a hierarchical sequence reverts back from n levels to $\frac{n\cdot(n-1)}{2}+1$ levels, and the time complexity to construct a hierarchical sequence of cluster sets becomes $O(n^4)$. Moreover, the *post hoc* heuristics for cutting dendrograms are not suitable for finding meaningful cluster sets of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence. To overcome these problems, this paper presents three contributions. First, using the 2-norm distance measure as an example, it presents a means for finding meaningful levels of an $\frac{n\cdot(n-1)}{2}+1$-level hierarchical sequence *prior* to performing a cluster analysis. By finding meaningful levels of such a hierarchical sequence *prior* to performing a cluster analysis, it is possible to know which cluster sets to construct and construct only these cluster sets. This reduces the time complexity to construct cluster sets from $O(n^4)$ to $O(ln^2)$. Second, it shows how increasing the dimensionality of the data points helps reveal inherent structure in noisy data. Third, it provides working definitions for the notions "meaningful level" and "meaningful cluster set". The empirical results from four experiments show that finding meaningful levels of a hierarchical sequence is easy and yields results that can have real world meaning. Future work includes mathematically capturing and integrating the means into the new clustering method, so that the new clustering method is self-contained, and working with more complex beta applications.

ACKNOWLEDGEMENTS

The author thanks Dr. Larry Gray, Department of Mathematics, University of Minnesota, for his help with the proof, and Dr. John Carlis, Department of Computer Science and Engineering, University of Minnesota, for his general guidance and advice on technical writing. The author also thanks the paper's reviewers for reviewing the paper and for their helpful feedback.

REFERENCES

Anderberg, M. (1973). *Cluster Analysis for Applications*. Academic Press.

Berkhin, P. (2006). A survey of clustering data mining techniques. In Kogan, J., Nicholas, C., and Teboulle, M., editors, *Grouping Multidimensional Data: Re-*

cent Advances in Clustering, chapter 2, pages 25–71. Springer-Verlag.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1998). When is "nearest neighbor"meaningful? Technical report, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI.

Cormen, T., Leiserson, C., Rivest, R., and Stein, C. (2004). *Introduction to Algorithms*. MIT Press, 2nd edition.

Daniels, K. and Giraud-Carrier, C. (2006). Learning the threshold in hierarchical agglomerative clustering. In *Proceedings of the Fifth International Conference on Machine Learning and Applications (ICMLA '06)*, pages 270–278, Orlando, FL.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. John Wiley and Sons, 5th edition.

Gill, H. (2011). CPS overview. In *Symposium on Control and Modeling Cyber-Physical Systems (www.csl.illinois.edu/video/csl-emerging-topics-2011-cyber-physical-systems-helen-gill-presentation)*, Champaign, IL.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., and Stanley, H. (June 13, 2000). *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://cir.ahajournals.org/cgi/content/full/101/23/e215]*.

Hinneburg, A., Aggarwal, C., and Keim, D. (2000). What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB 2000)*, pages 506–516, Cairo, Egypt.

Isermann, R. (2006). *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer-Verlag.

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice Hall.

Johnson, R. and Wichern, D. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall, 5th edition.

Kim, H. and Lee, S. (2000). A semi-supervised document clustering technique for information organization. In *Proceedings of the 9th ACM International Conference on Information and Knowledge Management (CIKM '00)*, pages 30–37, McLean, VA.

Kim, M., Payne, W. V., and Domanski, P. (2006). Performance of a residential heat pump operating in the cooling mode with single faults imposed. Technical report, U.S. National Institute of Standards and Technology, Gaithersburg, Maryland.

Kirk, D. and Hwu, W. (2013). *Programming Massively Parallel Processors*. Elsevier Inc., 2nd edition.

Lance, G. and Williams, W. (1967). A general theory of classificatory sorting strategies ii clustering systems. *Computer J.*, 10(3):271–277.

Matula, I. (1977). Graph theoretic techniques for cluster analysis algorithms. In Ryzin, J. V., editor, *Classification and Clustering*, pages 95–129. Academic Press.

Murtagh, F. (2009). The remarkable simplicity of very high dimensional data: Application of model-based clustering. *J. of Classification*, 26:249–277.

Navidi, W. (2006). *Statistics for Engineers and Scientists*. McGraw-Hill.

Olsen, D. (2014). Include hierarchical clustering: A hierarchical clustering method based solely on interpoint distances. Technical report, Minneapolis, MN.

Peay, E. (1974). Hierarchical clique structures. *Sociometry*, 37(1):54–65.

Peay, E. (1975). Nonmetric grouping: Clusters and cliques. *Psychometrika*, 40(3):297–313.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63(2):411–423.

APPENDIX

Proof for Theorem 1. *(sketch only)*
Assume that the stated conditions are true. The second moment of

$$Z_{m,(i,j)} \text{ is } E[Z^2_{m,(i,j)}] = E[(\sum_{k=1}^{m} Y^2_{k,(i,j)})^{\frac{1}{2}\cdot 2}]$$
$$= \sum_{k=1}^{m} E[Y^2_{k,(i,j)}]$$
$$= \sum_{k=1}^{m} E[(\sigma_{k,(i,j)} W_{k,(i,j)} + \mu_{k,(i,j)})^2]$$
$$= \sum_{k=1}^{m} E[(\sigma_{k,(i,j)} W_{k,(i,j)})^2 + 2\sigma_{k,(i,j)} W_{k,(i,j)} \mu_{k,(i,j)} + \mu^2_{k,(i,j)}]$$
$$= \sum_{k=1}^{m} E[(\sigma_{k,(i,j)} W_{k,(i,j)})^2] + \sum_{k=1}^{m} E[2\sigma_{k,(i,j)} W_{k,(i,j)} \mu_{k,(i,j)}] + \sum_{k=1}^{m} E[\mu^2_{k,(i,j)}],$$

where W is a normally distributed random variable. The expected value of the middle term in the last expression equals zero and drops out, so

$$E[Z^2_{m,(i,j)}] = \sum_{k=1}^{m} \sigma^2_{k,(i,j)} + \sum_{k=1}^{m} \mu^2_{k,(i,j)}.$$

To find the expected value $E[Z_{m,(i,j)}]$ in terms of the standard deviations $\sigma_{k,(i,j)}$ and mean values $\mu_{k,(i,j)}$, Taylor's series is used to expand $E[Z_{m,(i,j)}]$.

Let

$$x_0 = \sum_{k=1}^{m} ((\sigma_{k,(i,j)} W_{k,(i,j)})^2 + \mu^2_{k,(i,j)})$$

and

$$h = \sum_{k=1}^{m} 2\sigma_{k,(i,j)} W_{k,(i,j)} \mu_{k,(i,j)}.$$

Then,

$$
\begin{aligned}
E[Z_{m,(i,j)}] &= E[(\sum_{k=1}^{m} Y_{k,(i,j)}^2)^{\frac{1}{2}}] \\
&= E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2 + 2\sigma_{k,(i,j)}W_{k,(i,j)}\mu_{k,(i,j)}))^{\frac{1}{2}}] \\
&\approx E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}} + \frac{h}{2(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}}} \\
&\quad - \frac{h^2}{8(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{3}{2}}} + \frac{3h^3}{48(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{5}{2}}} + ...] \qquad (6) \\
&\approx E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}}] - E[\frac{h^2}{8(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{3}{2}}}]. \qquad (7)
\end{aligned}
$$

In Equation 6, $2\sigma_{k,(i,j)}W_{k,(i,j)}\mu_{k,(i,j)}$ is symmetric, so $E[\frac{h}{2(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2+\mu_{k,(i,j)}^2))^{\frac{1}{2}}}] = 0$ and drops out. As $m \to \infty$, the third-order term and all higher order terms converge to 0. Thus,

$$
\begin{aligned}
(E[Z_{m,(i,j)}])^2 \approx (E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}}])^2 \\
- 2(E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}}]E[\frac{h^2}{8(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{3}{2}}}]) \\
+ (E[\frac{h^2}{8(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{3}{2}}}])^2. \qquad (8)
\end{aligned}
$$

Using the dominated convergence theorem and Taylor's series, where $g = \sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 - \sigma_{k,(i,j)}^2)$, the first term in Equation 8 evaluates to

$$
(E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}}])^2 \approx \sum_{k=1}^{m} \sigma_{k,(i,j)}^2 + \sum_{k=1}^{m} \mu_{k,(i,j)}^2 - \frac{\sum_{k=1}^{m} \sigma_{k,(i,j)}^4}{2(\sum_{k=1}^{m} \sigma_{k,(i,j)}^2 + \sum_{k=1}^{m} \mu_{k,(i,j)}^2)}. \qquad (9)
$$

When h^2 is expanded, the terms with $(\sum_{k_1=1}^{m} 2\sigma_{k_1,(i,j)}W_{k_1,(i,j)}\mu_{k_1,(i,j)})(\sum_{k_2=1}^{m} 2\sigma_{k_2,(i,j)}W_{k_2,(i,j)}\mu_{k_2,(i,j)})$ in the numerator drop out, leaving only those terms with $4\sum_{k=1}^{m}(\sigma_{k,(i,j)}W_{k,(i,j)}\mu_{k,(i,j)})^2$ in the numerator. Using the dominated convergence theorem and Taylor's series once more, where $g = \sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 - \sigma_{k,(i,j)}^2)$, the second term in Equation 8 evaluates to

$$
\begin{aligned}
&- 2(E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}}]E[\frac{h^2}{8(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{3}{2}}}]) \\
&\approx -E[(\sum_{k=1}^{m} ((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{1}{2}}]E[\frac{\sum_{k=1}^{m}(\sigma_{k,(i,j)}W_{k,(i,j)}\mu_{k,(i,j)})^2}{(\sum_{k=1}^{m}((\sigma_{k,(i,j)}W_{k,(i,j)})^2 + \mu_{k,(i,j)}^2))^{\frac{3}{2}}}] \\
&\approx -\frac{\sum_{k=1}^{m} \sigma_{k,(i,j)}^2\mu_{k,(i,j)}^2}{\sum_{k=1}^{m} \sigma_{k,(i,j)}^2 + \sum_{k=1}^{m} \mu_{k,(i,j)}^2}. \qquad (10)
\end{aligned}
$$

As $m \to \infty$, the third term in Equation 8 converges to 0.

Thus, the variance $\sigma^2_{Z_{m,(i,j)}}$ *of* $Z_{m,(i,j)}$ *is*

$$\begin{aligned}\sigma^2_{Z_{m,(i,j)}} &= E[Z^2_{m,(i,j)}] - (E[Z_{m,(i,j)}])^2 \\ &\approx \sum_{k=1}^{m} \sigma^2_{k,(i,j)} + \sum_{k=1}^{m} \mu^2_{k,(i,j)} - (\sum_{k=1}^{m} \sigma^2_{k,(i,j)} + \sum_{k=1}^{m} \mu^2_{k,(i,j)} \\ &\quad - \frac{\sum_{k=1}^{m} \sigma^4_{k,(i,j)}}{2(\sum_{k=1}^{m} \sigma^2_{k,(i,j)} + \sum_{k=1}^{m} \mu^2_{k,(i,j)})} - \frac{\sum_{k=1}^{m} \sigma^2_{k,(i,j)} \mu^2_{k,(i,j)}}{\sum_{k=1}^{m} \sigma^2_{k,(i,j)} + \sum_{k=1}^{m} \mu^2_{k,(i,j)}}) \\ &= \frac{\sum_{k=1}^{m} \sigma^4_{k,(i,j)}}{2(\sum_{k=1}^{m} \sigma^2_{k,(i,j)} + \sum_{k=1}^{m} \mu^2_{k,(i,j)})} + \frac{\sum_{k=1}^{m} \sigma^2_{k,(i,j)} \mu^2_{k,(i,j)}}{\sum_{k=1}^{m} \sigma^2_{k,(i,j)} + \sum_{k=1}^{m} \mu^2_{k,(i,j)}}. \end{aligned} \tag{11}$$

QED