ASYMPTOTIC AND COMPUTATIONAL PROBLEMS IN SINGLE-LINK CLUSTERING

by
Evangelos Tabakis
Ptychion, University of Athens, 1987

Submitted to the Department of Mathematics
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology
July 1992

Signature of Author ______________________________
Department of Mathematics
July 31, 1992

Certified by ______________________________
Richard M. Dudley
Professor of Mathematics
Thesis Supervisor

Accepted by ______________________________
Alar Toomre
Chairman
Committee on Applied Mathematics

Accepted by ______________________________
Sigurdur Helgason
Chairman
Departmental Graduate Committee

ASYMPTOTIC AND COMPUTATIONAL PROBLEMS IN SINGLE-LINK CLUSTERING

by
Evangelos Tabakis

Submitted to the Department of Mathematics on July 31, 1992,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Mathematics

The main theme of this thesis is the study of the asymptotic and computational aspects of clustering analysis for samples of iid observations in an effort to improve upon the older methods. We are concerned with hierarchical clustering methods and we focus on the single link method. First, a detailed general framework is developed to deal with hierarchical structure in either the sample or the population case. In this general setting, we establish the equivalence of hierarchies and ultrametric distances, define single-link distances and derive the connection to minimal spanning trees.

The next step is to study the behavior of single-link distances between iid observations drawn from probability distributions whose support is compact and has a finite number of connected components. For such distributions, we prove the consistency of single-link distances and in the case of one dimensional distributions we obtain an asymptotically normal distribution for the average single link distance using facts about spacings. In the case of multivariate distributions and under some conditions, we obtain the rate of convergence for the maximum single-link distance (which is equal to the length of the longest edge of the minimal spanning tree) and give upper and lower bounds.

To deal with the chaining problem in real data, we combine kernel density estimation with the computation of minimal spanning trees to study the effect of density truncation on single-link partitions. New statistics are proposed to help decide on the best truncation level, leading to an improved version of the single-link method. Simulation studies show how these statistics perform with unimodal and bimodal densities. Finally, these tools are applied to two clustering examples: One involves grouping several foods according to the nutrients they contain. The other is a market segmentation study, concerning an Atlanta manufacturer of prefabricated homes.

Thesis supervisor: Richard M. Dudley.
Title : Professor of Mathematics.

ASYMPTOTIC AND COMPUTATIONAL PROBLEMS IN SINGLE-LINK CLUSTERING

by

Evangelos Tabakis

Τοῦτ’ αὐτὸ τοίνυν ἡμᾶς ὁ πρόσθεν λόγος ἀπαιτεῖ,
πῶς ἔστιν ἕν καί πολλά αὐτῶν ἑκάτερον,
καί πῶς μὴ ἄπειρα εὐθύς,
ἀλλὰ τινὰ ποτε ἀριθμόν ἑκάτερον ἔμπροσθεν κέκτηται
τοῦ ἄπειρα αὐτῶν ἕκαστα γεγονέναι;

This is exactly what the previous discussion requires from us:
How is it possible for each of them
to be one and many at the same time
and how is it they do not immediately become Infinity
but instead they first acquire a finite number
before each of them becomes Infinity?

Plato, Philebus 19A.

To my family, for their love and support.

Acknowledgements

New ideas rarely come out of nowhere and this work is no exception. The inspiration can often be traced back to the enlightening lectures I was fortunate to attend both at MIT and at Harvard University. Other contributions came in the form of informal discussions with a number of distinguished professors, friends and colleagues. For one or both of these reasons I feel I must mention, in particular, the names of: Dimitris Bertsimas, Kjell Doksum, Michael Economakis, Wayne Goddard, John Hartigan, Greta Ljung, Panagiotis Lorentziadis, Walter Olbricht, Adolfo Quiroz, Helmut Rieder, David Schmoys, Hal Stern, Achilles Venetoulias and Jeff Wooldridge.

Special thanks are due to Peter Huber, Gordon Kaufman and Mark Matthews for reading this thesis and making several suggestions which resulted in substantial improvements. I must also thank Drazen Prelec for helping me find the data set used in chapter 7 and teaching me about marketing research. And, of course, I owe a lot to the continuous and patient guidance of my thesis supervisor, Richard Dudley. It would be impossible to mention the many ways in which he has contributed to this thesis but it suffices to say that, without him, I would have never brought this work to an end. I hope I haven't caused him too much aggravation during these years and I will always consider it an honor to be counted among his students.

The Department of Mathematics and the Sloan School of Management have provided financial support for my graduate studies at MIT. Richard Dudley, Gordon Kaufman and Phyllis Ruby were instrumental in securing it.

Above all, however, I wish to thank my family for their support and love: my father for our lengthy discussions which influenced my way of thinking; my mother for taking care of every problem that occurred; my sister for keeping me in touch with reality over the years. These are the people that made this possible and to whom this work is dedicated.

Evangelos Tabakis

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The clustering problem

The main problem of cluster analysis is summarized in [MKB79], page 360:

> Let $x_1, \ldots, x_n$ be measurements of p variables on each of n objects which are believed to be heterogeneous. Then the aim of cluster analysis is to group these objects into g homogeneous classes where g is also unknown (but usually assumed to be much smaller than n).

There is no shortage of proposed methods to tackle this problem. Detailed listings have been included in books and review papers such as, e.g., [Eve74], [Har75], [Gor81], [Gor87], [JD88] and [KR90]. Very often, these methods are described by means of an algorithm. As it often happens with other nonparametric multivariate problems (see [Hub91]), the goal that the algorithm is trying to attain is not specified explicitly. This is partly due to the lack of a universally accepted interpretation of the term ***homogeneous*** as used in the quote from [MKB79]. Such an intepretation would also amount to a description of the properties of *clusters* and is, therefore, central to clustering analysis.

There are at least two widely used interpretations in the clustering literature (see e.g. [Boc85] and [Gor87]). One describes homogeneity as uniformity on a compact and connected set G. Tests of this hypothesis can be based on the work of N. Henze ([Hen83]). A different approach has been taken by D. Strauss ([Str75]). The most important drawback is that these tests assume that the set G is known. Without knowledge of G, we cannot account for the effect of the shape of G on the statistics used. A similar *edge* effect is recorded in the use of spatial processes in image processing (see e.g. [Rip88], chapter 3).

The other interpretation assumes the existence of a density f (with respect to Lebesgue measure) and equates homogeneity with unimodality of f. This leads us to the use of mode-seeking methods in order to specify the location of clusters (see e.g. [JD88], page 118). Note, however, that it is very difficult to find the modes of a density in d-dimensional space. In the one dimensional case, there has been some progress([Sil81], [HH85]). A suggestion for an extension of the one-dimensional method of [HH85] to higher dimensions is contained in [Har88].

A certain compromise between the two interpretations can be reached through the suggestion of J. Hartigan ([Har85]) to take clusters to be *maximally connected high-density sets*, i.e. the connected components of the region $\{x \in \mathbf{R}^d : f(x) > c\}$ for an appropriate c. It seems, therefore, that a search for clusters must use information about:

- where the observations lie and

- how densely they are packed together.

In fact, there have been suggestions ([Gil80]) which assume that a preliminary estimate of the location of each cluster is available (together with an estimate of the probability of the cluster) and proceed to find the observations that belong to that cluster through iterations. This, however, leaves the question of the global search for the location of the clusters open.

In the next two sections we intoduce the two main groups of clustering methods.

1.2 Partitional methods

The ambiguity of the terms *homogeneous group* or *cluster* makes it even more difficult to develop statistical inference for clustering. Some progress has been made in the area of partitional methods. These attempt to find a partition of the observations that optimizes a certain criterion. The main idea is to decide on the number of clusters *before* looking at the observations and then try to minimize the within-cluster distances of these observations. Such methods (and related algorithms) go back to the work of Friedman and Rubin (see [FRb67]).

The most popular among them is the *k-means* method where the partition $\mathcal{P} = (C_1, C_2, \dots, C_k)$ chosen is the one that minimizes:

$$T(\mathcal{P}) = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_j - \bar{x}_i)^2 1_{C_i}(x_j)$$

where:

$$\bar{x}_i = \frac{\sum_{j=1}^{n} 1_{C_i}(x_j) x_j}{|C_i|}.$$

Since the number of partitions of n observations into k clusters is:

$$S(n,k) = \frac{\sum_{i=1}^{k} (-1)^{k-i} \left(\begin{array}{c} k \\ i \end{array} \right) i^n}{k!}$$

(the Stirling numbers of the second kind, see [Sta86], pages 33-34) an exhaustive search is out of the question. Instead, iterative algorithms have been devised (see, e.g., [JD88], page 96 and [KR90], page 102).

Consistency of the k-means method is treated in [Har78], [Pol81] and, in a more general setting, in [CM88]. The asymptotic normality of the centers of the k-means clusters is proved in [Pol82]. Another interesting question is the estimation of k. [But86] and [But88] treat this on the real line. More recently, [PFvN89] addressed the same problem in the multivariate case.

1.3 Hierarchical methods

These methods use the observations to produce a sequence of n partitions $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_n$ (often refered to as a *hierarchy* of partitions) with the properties:

- $\mathcal{P}_1$ is the partition into n one-element clusters.

- $\mathcal{P}_i$ has $n - i + 1$ clusters of which $n - i$ are the same as $n - i$ clusters in $\mathcal{P}_{i-1}$ and the $(n - i + 1)st$ cluster is formed by joining the remaining two clusters of $\mathcal{P}_{i-1}$ into one ($i = 2, 3, \dots, n$).

A class of such methods is based on defining a distance d_C between clusters. Then a general algorithm that produces the sequence $\{\mathcal{P}_i, i = 1, \ldots, n\}$ is the following:

- $\mathcal{P}_1$ is the partition: $\{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$.
- Given $\mathcal{P}_{i-1}$, $\mathcal{P}_i$ is formed by finding the two clusters C_1 and C_2 for which: $d_C(C_1, C_2) = \min\{d_C(A, B),\ A, B \in \mathcal{P}_{i-1}\}$ and join them into one cluster.

Popular choices for d_C are:

$$d_C(A, B) = \min\{d(x, y),\ x \in A,\ y \in B\}$$

and

$$d_C(A, B) = \max\{d(x, y),\ x \in A,\ y \in B\}$$

resulting into the *single link* and *complete link* methods respectfully (see, e.g., [KR90], page 47).

Hierarchical methods have certain advantages that make them popular. Some of them are:

- They describe the *clustering structure* of the data set without the need to *prespecify* the number of clusters we must look for. Choosing the number of clusters can be then based on inspection of the hierarchy of partitions. Note, however, that inspecting the partitions is not a trivial task for large data sets in high dimensions.
- The algorithm we just described needs $O(n^3)$ steps to form the hierarchy of partitions[1] compared to partitional methods that need iterarive algorithms to produce a single partition. Even worse, the work done to compute a partition into, say, three clusters cannot be used in calculating a partition into four or two clusters when using a partitional method.
- Identifying clusters is often a subjective decision. What some people may see as one cluster, some others might consider as two or more. It is

[1] Using the concept of reciprocal neighbors it is possible to form the hierarchy in $O(n^2)$ steps (see [LMW84], pages 128-129).

often a question of how fine a partition we want to find, that determines the answer. This feature of the clustering problem is best captured by hierarchical methods.

The hierarchical structure involved in these methods explains why there is so little work done on the asymptotics of hierarchical methods. The problem of consistency of single-link has been addressed in [Har81].

1.4 Minimal spanning trees in clustering

Tree methods are often used in nonparametric multivariate statistics (see e.g. [BFOS84] for classification and regression and [FRf79, FRf81] for the two-sample problem). In this thesis, we will make ample use of the minimal spanning tree (MST) on n points. This is simply any tree with vertices these n points that attains the smallest possible total length. Complete definitions of all the graph-theoretic terms involved will be given in Chapter 2. In general, an MST can be computed (by a variety of algorithms) in $O(n^2)$ time (see [Hu82] pages 28-29, [PS82] pages 271-279 or [NW88] pages 60-61).

The close connection of the MST to clustering was pointed out in [GR69] and since then it is practically impossible to talk about single-link clustering without also talking about the MST. In Chapters 2,3,4 and 5, we will build on this connection to establish several asymptotic results about single-link. The connection is shown in the next examples. In Figure 1.1, we draw the MST for 160 observations drawn from the uniform distribution on the unit square and a boxplot for the edge lengths of this tree. As expected in this case, no edge stands out as significantly larger than the others. Compare that with Figure 1.2, where the MST and the corresponding boxplot is shown for a sample drawn from a mixture of two uniform distributions on disjoint squares. This time, the edge that connects the two squares is significantly longer than all others, indicating the existence of cluster structure. Removing this longest edge reveals the two clusters.

It may seem at this point that the use of the MST is all we need to solve the clustering problem described in Section 1.1. The next example shows that this is not the case at all. In Figure 1.3 we have the MST and boxplot for the same observations as in Figure 1.2, this time adding another 40 observations from a bivariate normal centered between the two squares.

Although the clustering structure is still clear to the human eye, the boxplot gives a very confusing picture. The additional observations form chains of observations through which the MST joins the two clusters without having to use a long edge. So, there is no significantly longest edge in this MST. In fact the longest edge is not connecting the two clusters but is rather caused by an outlier. This problem (appearing very often in real data) is called *chaining.* It seems, therefore, that some adjustments have to be made in order to be able to detect the cluster structure and discover the clusters in cases such as in Figure 1.3. This problem will be the object of Chapters 6 and 7.

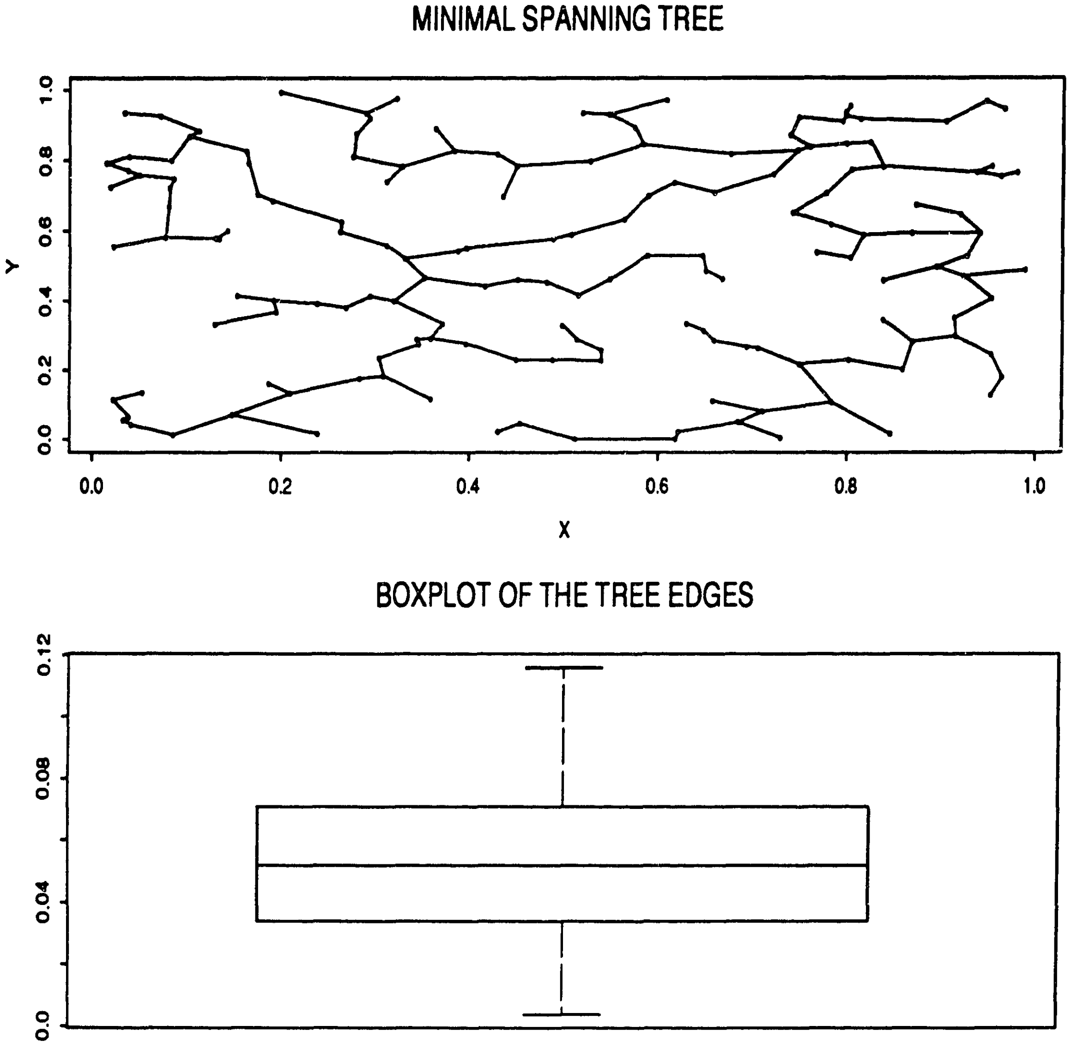
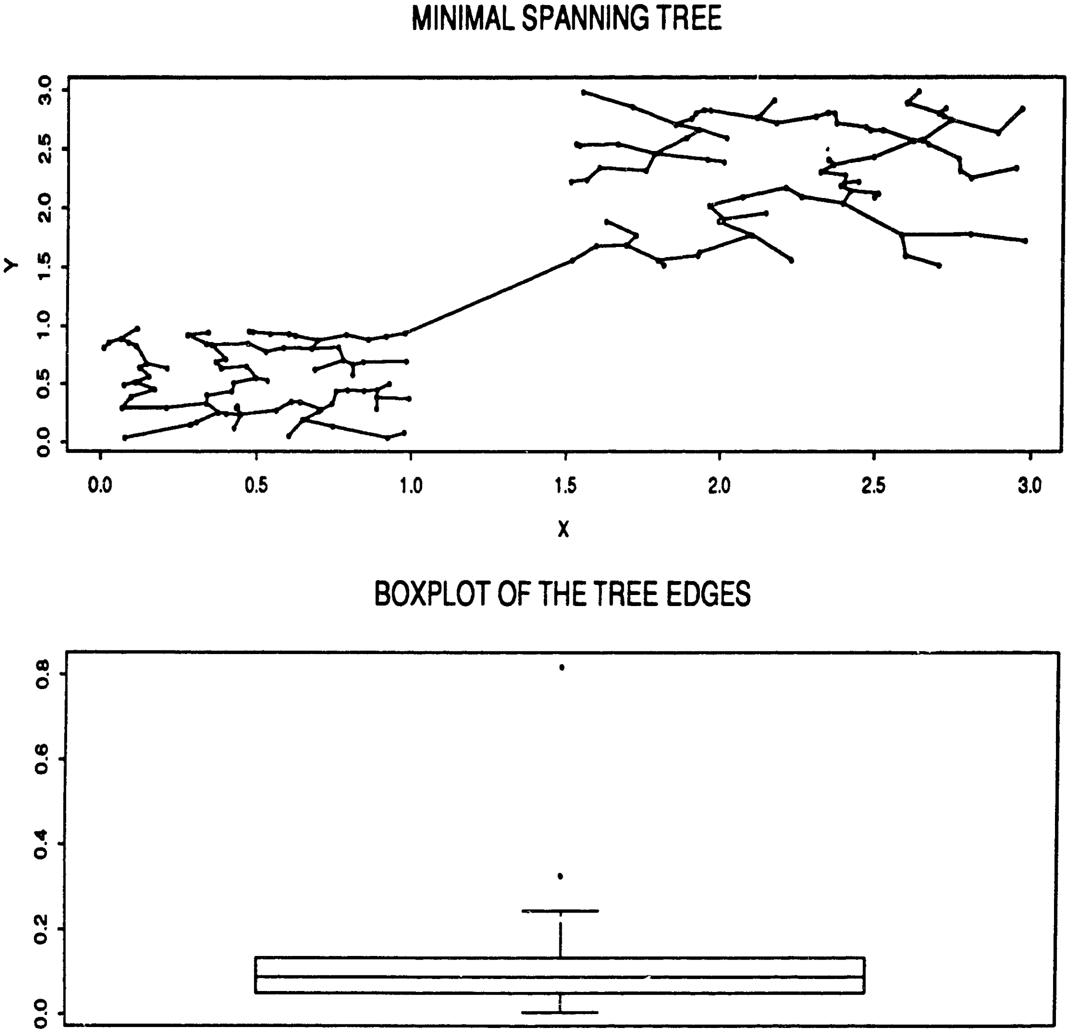


Figure 1.1: The MST for a uniform on the unit square.



MINIMAL SPANNING TREE

BOXPLOT OF THE TREE EDGES

Figure 1.2: The MST for a mixture of two uniforms with disjoint support.

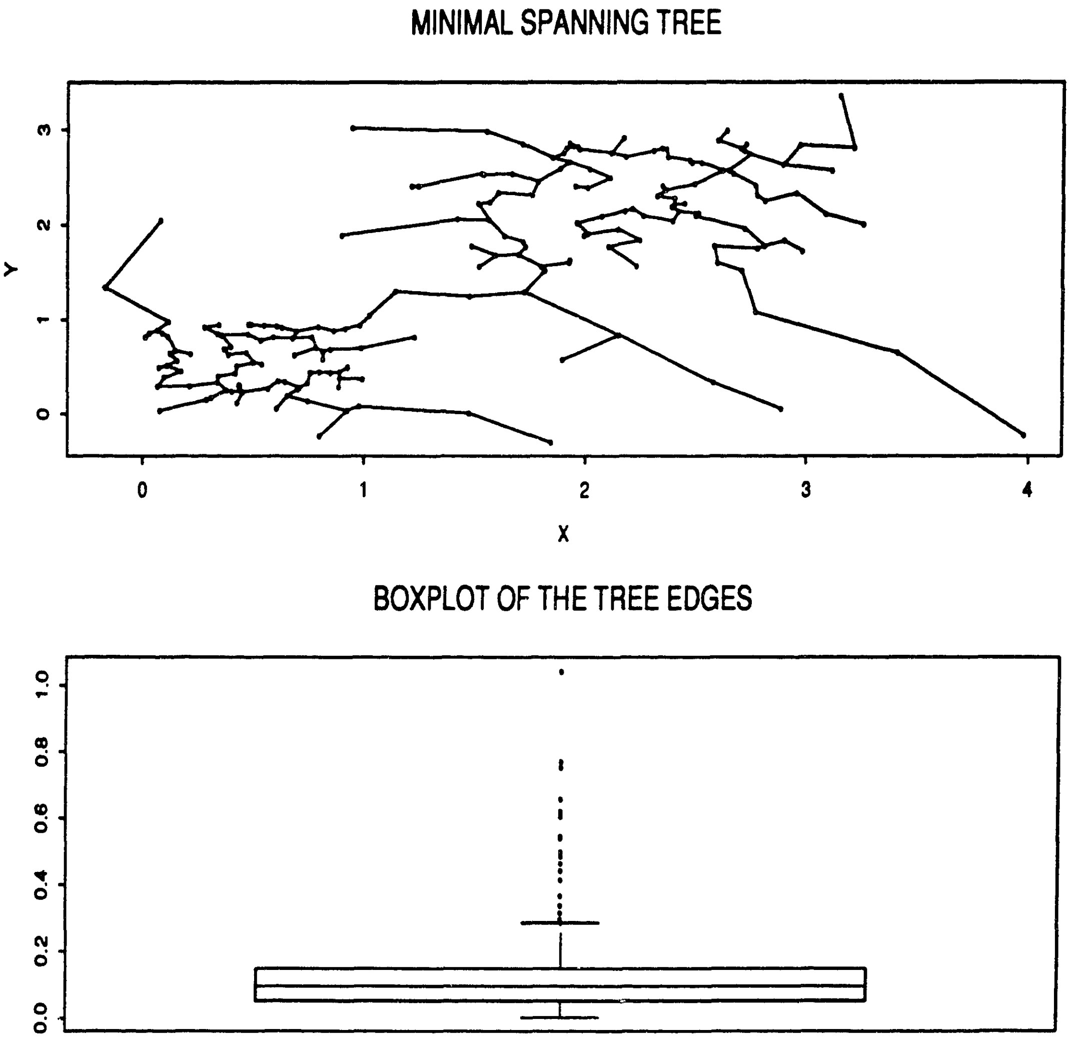


Figure 1.3: The MST for a contaminated mixture of two uniforms.

Chapter 2

Describing Hierarchies

2.1 $\mathcal{A}$-hierarchies

Let $\mathcal{A}$ be a family of subsets of $\mathbf{R}^d$.

Definition 2.1.1 *A partition $\mathcal{P}$ of a set $S \subset \mathbf{R}^d$ is a finite family $\{A_1, A_2, \ldots, A_r\}$ of non-empty subsets of R^d such that $S = A_1 \cup A_2 \cup \ldots \cup A_r$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, $1 \leq i, j \leq r$.*

Definition 2.1.2 *A partition $\mathcal{P}_1$ of S is finer than a partition $\mathcal{P}_2$ of S ($\mathcal{P}_1 \prec \mathcal{P}_2$) if:*

$$\forall A \in \mathcal{P}_2,\ \exists r \in \mathbf{N}\ and\ B_1, B_2, \ldots, B_r \in \mathcal{P}_1$$

such that:

$$A = B_1 \cup B_2 \cup \ldots \cup B_r.$$

Definition 2.1.3 *An $\mathcal{A}$-clustering of a set $S \in R^d$ is a partition:*

$$\mathcal{C} = \{C_1, C_2, \ldots, C_r\}$$

where:

$$C_i \in \mathcal{A}$$

and

$$\bar{C}_i \cap \bar{C}_j = \emptyset$$

for $1 \leq i < j \leq r$.

Definition 2.1.4 *An $\mathcal{A}$-hierarchy of a set $S \subset \mathbf{R}^d$ is a triple $(\mathcal{H}, \{\mathcal{P}_i\}_{i=0}^k, h)$ (or simply $\mathcal{H}$ when there is no danger of confusion) where:*

1. *$\mathcal{H} = \mathcal{P}_0 \cup \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_k$ where*
   - *$\mathcal{P}_i$ is a partition of S, for $0 \leq i \leq k$,*
   - *$\mathcal{P}_0$ is an $\mathcal{A}$-clustering of S,*
   - *$\mathcal{P}_k = \{S\}$ and*
   - *$\mathcal{P}_{i-1} \prec \mathcal{P}_i$ for $1 \leq i \leq k$.*
2. *$h : \mathcal{H} \mapsto \mathbf{R}^+$ such that:*
   - *$\forall A, B \in \mathcal{H} : A \subset B \Rightarrow h(A) \leq h(B)$,*
   - *$h(A) = 0 \Leftrightarrow A \in \mathcal{P}_0, \forall A \in \mathcal{H}$.*

Remark 2.1.1 $\forall A \in \mathcal{H}$, $\exists r \in \mathbf{N}$ and $C_1, C_2, \ldots, C_r \in \mathcal{C}$ such that:

$$A = C_1 \cup C_2 \cup \ldots \cup C_r.$$

Remark 2.1.2 Let $G(\mathcal{H}, E)$ be the graph with vertices the sets of $\mathcal{H}$ and edges E, where:

$$E = \{(A, B) : A, B \in \mathcal{H},\ \exists i : A \in \mathcal{P}_i,\ B \in \mathcal{P}_{i+1}, \text{ and } A \subset B\}.$$

Then $G(\mathcal{H}, E)$ is a tree with root S and leaves the sets of $\mathcal{C}$.

Remark 2.1.3 $\forall A \neq B \in \mathcal{H}$ one and only one of the following is true:

1. $A \cap B = \emptyset$,
2. $A \subset B$ or
3. $B \subset A$.

2.2 $\mathcal{A}$-ultra-pseudometrics

Definition 2.2.1 *An ultra-pseudo-metric (UPM) d on $S \subset \mathbf{R}^d$ is a pseudo-metric on S which, in addition, satisfies the inequality:*

$$d(x,y) \leq \max\{d(x,z), d(z,y)\}, \ \forall x,y,z \in S.$$

Definition 2.2.2 *An $\mathcal{A}$-ultra-pseudometric ($\mathcal{A}$-UPM) d on $S \subset \mathbf{R}^d$ is a UPM for which the family of sets:*

$$\{d^{-1}(x,.)(\{0\}),\ x \in S\}$$

forms an $\mathcal{A}$-clustering of S.

Lemma 2.2.1 *Let d be an $\mathcal{A}$-UPM on $S \subset R^d$ and let:*

$$\mathcal{C} = \{d^{-1}(x,.)(\{0\}),\ x \in S\}.$$

Then, $\forall C_1, C_2 \in \mathcal{C}$ $(C_1 \neq C_2)$ and $\forall x_1, y_1 \in C_1$ and $x_2, y_2 \in C_2$:

$$d(x_1,x_2) = d(y_1,y_2) = d(C_1,C_2) > 0$$

and

$$d(x_1,y_1) = d(x_2,y_2) = 0.$$

Proof: Since $x_1, y_1 \in C_1 = d^{-1}(x,.)(\{0\})$ for some $x \in C_1$:

$$d(x_1,x) = d(y_1,x) = 0 \Rightarrow$$

$$d(x_1,y_1) \leq d(x_1,x) + d(x,y_1) = 0 \Rightarrow$$

$$d(x_1,y_1) = 0.$$

Similarly: $d(x_2,y_2) = 0$.
Then:

$$\begin{aligned} d(x_1,x_2) &\leq d(x_1,y_1) + d(y_1,x_2) \\ &\leq d(x_1,y_1) + d(y_1,y_2) + d(y_2,x_2) \\ &= d(y_1,y_2). \end{aligned}$$

Similarly: $d(y_1, y_2) \leq d(x_1, x_2)$. So:

$$d(x_1, x_2) = d(y_1, y_2) = \inf\{d(z_1, z_2),\ z_1 \in C_1,\ z_2 \in C_2\} = d(C_1, C_2).$$

If $d(C_1, C_2) = 0$, then $\forall \epsilon > 0$, $\exists x_1 \in C_1$, $x_2 \in C_2 : d(x_1, x_2) < \epsilon$. Suppose $C_1 = d^{-1}(x, .)(\{0\})$ and $C_2 = d^{-1}(y, .)(\{0\})$. Then:

$$\begin{aligned} d(x, y) &\leq d(x, x_1) + d(x_1, x_2) + d(x_2, y) \\ &= d(x_1, x_2) < \epsilon. \end{aligned}$$

So:

$$\begin{aligned} d(x, y) = 0 &\Rightarrow y \in d^{-1}(x, .)(\{0\}) = C_1 \\ &\Rightarrow C_1 \cap C_2 \neq \emptyset, \end{aligned}$$

a contradiction. So $d(C_1, C_2) > 0$. □

As it turns out, $\mathcal{A}$-hierarchies and $\mathcal{A}$-UPMs are equivalent in describing hierarchical structure. The following theorem proves this in detail. The main idea used is taken from [Joh67].

Theorem 2.2.1 *Let $S \subset \mathbf{R}^d$, let $\mathbf{H}$ be the set of all $\mathcal{A}$-hierarchies of S and $\mathbf{U}$ the set of all $\mathcal{A}$-UPMs of S. Then, there is a map $m : \mathbf{H} \mapsto \mathbf{U}$ which is 1-1 and onto.*

Proof: Let $(\mathcal{H}, \{\mathcal{P}_i\}_{i=1}^k, h) \in \mathbf{H}$. Let $\mathcal{C} := \mathcal{P}_0$. Consider the function:

$$d_{\mathcal{H}} : \mathbf{R}^d \times \mathbf{R}^d \mapsto \mathbf{R}_+$$

defined as follows: For every pair $(x, y) \in \mathbf{R}^d \times \mathbf{R}^d$ let:

$$L_{x,y} := \{A \in \mathcal{H} : \{x, y\} \subset A\}.$$

Since $S \in L_{x,y}$, $L_{x,y} \neq \emptyset$. Let $A_{x,y} = \cap_{A \in L_{x,y}} A$. Because of Remark 2.1.3, $A_{x,y} \in \mathcal{H}$ so we can define:

$$d_{\mathcal{H}}(x, y) := h(A_{x,y}).$$

We must check that $d_{\mathcal{H}} \in \mathbf{U}$.

- Let $x \in S$. Then $\exists C_x \in \mathcal{C}$ so that $x \in C_x$. Since $A_{x,x} \in \mathcal{H}$, we have (using Remark 2.1.1): $C_x \subset A_{x,x}$. Also, by the definition of $A_{x,x}$:

  $$A_{x,x} \subset C_x.$$

  So:

  $$A_{x,x} = C_x \Rightarrow d_{\mathcal{H}}(x,x) = h(A_{x,x}) = h(C_x) = 0$$

  (by Definition 2.1.4).
- $d_{\mathcal{H}}(x,y) = h(A_{x,y}) = d_{\mathcal{H}}(y,x)$.
- Let $x, y, z \in S$. Then, $z \in A_{x,z} \cap A_{y,z}$. Because of Remark 2.1.3,

  $$A_{x,z} \subset A_{y,z} \text{ or } A_{y,z} \subset A_{x,z}.$$

  Let us assume that $A_{y,z} \subset A_{x,z}$. Then:

  $$h(A_{y,z}) \leq h(A_{x,z}) \Rightarrow d_{\mathcal{H}}(y,z) \leq d_{\mathcal{H}}(x,z).$$

  Also: $\{x,y\} \subset A_{x,z}$ so:

  $$A_{x,y} \subset A_{x,z} \Rightarrow h(A_{x,y}) \leq h(A_{x,z})$$

  $$\begin{aligned} &\Rightarrow d_{\mathcal{H}}(x,y) \leq d_{\mathcal{H}}(x,z) = \max\{d_{\mathcal{H}}(x,z),\, d_{\mathcal{H}}(y,z)\} \\ &\qquad \leq d_{\mathcal{H}}(x,z) + d_{\mathcal{H}}(y,z). \end{aligned}$$
- Let $x \in S$. Then again, let $C_x \in \mathcal{C}$ so that $x \in C_x$.

  $$\forall y \in C_x : A_{x,y} = C_x \Rightarrow d_{\mathcal{H}}(x,y) = h(C_x) = 0.$$

  $$\forall y \in S \setminus C_x : \ C_x \subset A_{x,y} \text{ but } C_x \neq A_{x,y} \Rightarrow d_{\mathcal{H}}(x,y) = h(A_{x,y}) > 0.$$

  So: $d_{\mathcal{H}}^{-1}(x,.)(\{0\}) = C_x$ and

  $$\{d_{\mathcal{H}}^{-1}(x,.)0,\, x \in S\} = \mathcal{C},$$

  an $\mathcal{A}$-clustering.

So, $d_{\mathcal{H}} \in \mathbf{U}$.

Conversely: Let $d \in \mathbf{U}$. Then:

$$\mathcal{C} = \{d^{-1}(x,.)(\{0\}),\, x \in S\}$$

is an $\mathcal{A}$-clustering of S (Definition 2.2.2). We now define the following partitions of S:

- $\mathcal{P}_0 := \mathcal{C}$ and $\forall C \in \mathcal{C} :\ h(C) := 0$.
- Suppose $\mathcal{P}_i$ is defined and is equal to $\{A_1, A_2, \ldots, A_{r_i}\}$. Let:

  $$s_i := \min_{1 \le l < j \le r_i} d(A_l, A_j).$$

  Let

  $$J_l := \{j : d(A_l, A_j) \le s_i\}$$

  and

  $$B_l := \cup_{j \in J_l} A_j,\ 1 \le l \le r_i.$$

  Let

  $$\mathcal{P}_{i+1} := \{B_l,\ 1 \le l \le r_i\}$$

  and

  $$r_{i+1} := \mathrm{card}(\mathcal{P}_{i+1}).$$

  Since at least two sets in $\mathcal{P}_i$ were joined into one in $\mathcal{P}_{i+1}$, we have $r_{i+1} < r_i$. Finally, $\forall B \in \mathcal{P}_{i+1} \setminus \mathcal{P}_i$, let $h(B) := s_i$.

Since $r_{i+1} < r_i$, we will eventually reach a $\mathcal{P}_k$ with $\mathrm{card}(\mathcal{P}_k) = 1$. At this point the construction terminates and we let $\mathcal{H} := \mathcal{P}_0 \cup \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_k$.

We first need to show that $\mathcal{P}_i$, $0 \le i \le k$ are partitions of S. If $k = 0$, this is obvious. Assume $k > 0$. In fact, we can show, by induction, that:

$$\mathcal{P}_i = \{A_1, A_2, \ldots, A_{r_i}\}$$

is a partition *and*

$$\mathrm{diam}(A_l) < s_i,\ 1 \le l \le r_i$$

where $\mathrm{diam}(A) := \sup\{d(x,y),\ x, y \in A\}$, $A \in \mathcal{H}$.

- For $i = 0$, $\mathcal{P}_0 = \mathcal{C}$ is a partition and, because of Lemma 2.2.1, $\text{diam}(C) = 0$, $\forall C \in \mathcal{C}$ and $s_0 = \min_{1 \leq l < j \leq r_0} d(A_l, A_j) > 0$.

- Suppose $\mathcal{P}_i$ is a partition with $\text{diam}(A_l) < s_i$, $1 \leq l \leq r_i$.
  Let B_l, $1 \leq l \leq r_i$, be defined as above. Suppose $\exists l_1, l_2$ such that $B_{l_1} \neq B_{l_2}$ and $B_{l_1} \cap B_{l_2} \neq \emptyset$. That would mean $\exists A_j$, $1 \leq j \leq r_i$ such that:

$$d(A_{l_1}, A_j) \leq s_i,\ d(A_{l_2}, A_j) \leq s_i$$

  but

$$d(A_{l_1}, A_{l_2}) > s_i.$$

  Let $\epsilon < d(A_{l_1}, A_{l_2}) - s_i$. Then:

$$\exists x_{l_1} \in A_{l_1},\ x_j \in A_j : d(x_{l_1}, x_j) < s_i + \epsilon/2$$

  and

$$\exists x_{l_2} \in A_{l_2},\ y_j \in A_j : d(x_{l_2}, y_j) < s_i + \epsilon/2.$$

  Then:

$$d(x_{l_1}, x_{l_2}) \geq d(A_{l_1}, A_{l_2}) > s_i + \epsilon/2.$$

  By the induction hypothesis: $d(x_j, y_j) < s_i$. Then, applying the ultrametric inequality:

$$\begin{aligned} d(x_{l_1}, x_{l_2}) &\leq \max\{d(x_{l_1}, x_j), d(x_j, x_{l_2})\} \\ &\leq \max\{d(x_{l_1}, x_j), d(x_j, y_j), d(y_j, x_{l_2})\} \\ &< \max\{s_i + \epsilon/2, s_i, s_i + \epsilon/2\} \\ &= s_i + \epsilon/2, \end{aligned}$$

  contradicting $d(x_{l_1}, x_{l_2}) > s_i + \epsilon/2$. So,

$$\forall l_1, l_2 : B_{l_1} = B_{l_2} \text{ or } B_{l_1} \cap B_{l_2} \neq \emptyset.$$

  In addition: $A_l \subset B_l$ for $1 \leq l \leq r_i$ so:

$$S = \cup_l A_l \subset \cup_l B_l \subset S \Rightarrow \cup_l B_l = S.$$

  So: $\mathcal{P}_{i+1} = \{B_l,\ 1 \leq l \leq r_i\}$ is a partition.

Now, clearly, $s_{i+1} = \min_{1 \leq l < j \leq r_{i+1}} d(B_l, B_j) > s_i$. As we just proved:

$$\forall j_1, j_2 \in J_l,\ (1 \leq l \leq r_i) : d(A_{j_1}, A_{j_2}) \leq s_i.$$

By the induction hypothesis: $\mathrm{diam}(A_l) < s_i$. So:

$$\mathrm{diam}(B_l) = \mathrm{diam}(\cup_{j \in J_l} A_j)$$
$$= \max\{\max_{j \in J_l} \mathrm{diam}(A_j), \max_{j_1 \neq j_2,\, j_1, j_2 \in J_l} d(A_{j_1}, A_{j_2})\}$$
$$\leq s_i < s_{i+1}.$$

This completes the induction.

Using the fact that $s_{i+1} > s_i$, it becomes obvious that the properties of the h function in Definition 2.1.4 also hold. So $(\mathcal{H}, \{\mathcal{P}_i\}_{i=0}^{k}, h)$, as defined, is an $\mathcal{A}$-hierarchy.

It remains to be proved that when the map m is applied to the $\mathcal{A}$-hierarchy $(\mathcal{H}, \{\mathcal{P}_i\}_{i=0}^{k}, h)$ we just obtained, we get the original d back, i. e. :

$$d_{\mathcal{H}} := m(\mathcal{H}, \{\mathcal{P}_i\}_{i=0}^{k}, h) = d.$$

We will prove that:

$$\forall x, y \in S : d_{\mathcal{H}}(x, y) = d(x, y).$$

By definition: $d_{\mathcal{H}}(x, y) = h(A_{x,y})$, where $A_{x,y}$ is the smallest set in $\mathcal{H}$ that contains both x and y. Let $\mathcal{P}_i$ be the finest partition that contains $A_{x,y}$.

We proceed by induction on i:

- For $i = 0$,
$$A_{x,y} \in \mathcal{P}_0 = \mathcal{C} = \{d^{-1}(x, .)(\{0\}),\ x \in S\}.$$
Because $A_{x,y} \in \mathcal{C}$:
$$h(A_{x,y}) = 0 \Rightarrow d_{\mathcal{H}}(x, y) = 0.$$
Because $A_{x,y} \in \{d^{-1}(x, .)(\{0\}),\ x \in S\}$:
$$d(x, y) = 0.$$
So: $d_{\mathcal{H}}(x, y) = d(x, y)$.

- Suppose $d_{\mathcal{H}}(x,y) = d(x,y)$ for all $x, y \in S$ such that:

$$A_{x,y} \in \mathcal{P}_0 \cup \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_i.$$

We will prove the same for $A_{x,y} \in \mathcal{P}_{i+1} \setminus \mathcal{P}_i$. Let $\mathcal{P}_i = \{A_1, A_2, \ldots, A_{r_i}\}$. Then, for some l, $1 \leq l \leq r_i : A_{x,y} = \cup_{j \in J_l} A_j$. Suppose:

$$x \in A_{j_x},\ y \in A_{j_y},\ j_x, j_y \in J_l.$$

By the definition of $A_{x,y}$ as the smallest set in $\mathcal{H}$ containing both x and y:

$$A_{j_x} \cap A_{j_y} = \emptyset.$$

By the definition of s_i:

$$s_i := \min_{1 \leq l < j \leq r_i} d(A_l, A_j)$$

and that of J_l:

$$J_l := \{j : d(A_l, A_j) \leq s_i\}$$

we have $d(A_{j_x}, A_{j_y}) = s_i$. Since $A_{x,y} \in \mathcal{P}_{i+1} \setminus \mathcal{P}_i$, $h(A_{x,y}) = s_i$. Let $z \in A_{j_x}$. Then:

$$A_{x,z} \subset A_{j_x} \subset A_{x,y}$$
$$\Rightarrow h(A_{x,z}) \leq h(A_{j_x}) \leq h(A_{x,y}).$$

Also $A_{x,z} \subset A_{j_x}$ implies that:

$$A_{x,z} \in \mathcal{P}_0 \cup \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_i$$

and so, by the induction hypothesis $d_{\mathcal{H}}(x,z) = d(x,z)$. So:

$$d(x,z) = d_{\mathcal{H}}(x,z) = h(A_{x,z})$$
$$\leq h(A_{x,y}) = s_i = d(A_{j_x}, A_{j_y})$$
$$= \inf_{u \in A_{j_x}, w \in A_{j_y}} d(u,w) \begin{cases} \leq d(x,y) \\ \leq d(z,y). \end{cases}$$

But then $d(x,y) = d(z,y)$. Otherwise, if, e.g.,

$$d(x,y) > d(z,y) \geq d(x,z)$$

then the ultrametric inequality:

$$d(x,y) \leq \max\{d(z,y), d(x,z)\}$$

would be violated.

Similarly, for $w \in A_{j_y}$:

$$A_{w,y} \subset A_{j_y} \subset A_{x,y}$$

$$\Rightarrow h(A_{w,y}) \leq h(A_{j_y}) \leq h(A_{x,y})$$

and $A_{w,y} \subset A_{j_y}$ implies that:

$$A_{w,y} \in \mathcal{P}_0 \cup \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_i.$$

Then, by the induction hypothesis:

$$d(w,y) = d_{\mathcal{H}}(w,y) = h(A_{w,y}) \leq h(A_{x,y}) = s_i$$

$$= \inf_{u \in A_{j_x}, v \in A_{j_y}} d(u,v) \begin{cases} \leq d(z,y) \\ \leq d(z,w). \end{cases}$$

Then, again: $d(z,y) = d(z,w)$. So:

$$d(x,y) = d(z,y) = d(z,w), \ \forall z \in A_{j_x}, \ w \in A_{j_y}$$

$$\Rightarrow d_{\mathcal{H}}(x,y) = h(A_{x,y}) = s_i = d(A_{j_x}, A_{j_y})$$

$$= \inf_{z \in A_{j_x}, w \in A_{j_y}} d(z,w) = d(x,y).$$

This concludes the proof of $d_{\mathcal{H}} = d$ and the proof of the theorem. □

2.3 Single-link hierarchies

In what follows, we choose and fix a metric ρ on $\mathbf{R}^d$ that metrizes the usual topology. Definition 2.3.3 and Theorems 2.3.1 and 2.3.2 below are based on ideas in [LMW84], pages 143-144.

Definition 2.3.1 *Let $\mathcal{P}$ be a partition of $S \subset \mathbf{R}^d$. For $A, B \in \mathcal{P}$, we define a* **path** *from A to B on $\mathcal{P}$ to be a finite sequence:*

$$(A \equiv C_0, C_1, \ldots, C_k \equiv B)$$

of sets $C_i \in \mathcal{P}$, $1 \leq i \leq k$.

Definition 2.3.2 *The* **size** *of a path $(A \equiv C_0, C_1, ..., C_k \equiv B)$ from A to B on $\mathcal{P}$ is defined to be:*

$$\max_{1 \leq i \leq k} \rho(C_{i-1}, C_i).$$

Definition 2.3.3 *Let $\mathcal{C}$ be an $\mathcal{A}$-clustering of $S \subset \mathbf{R}^d$. On $S \times S$ we define the function:*

$$d_{SL}^{\mathcal{C}} : S \times S \mapsto \mathbf{R}^+$$

as follows: For $x, y \in S$ let $x \in C_x \in \mathcal{C}$, $y \in C_y \in \mathcal{C}$. Then:

$$d_{SL}^{\mathcal{C}}(x, y) := \min\{s : \exists \textit{ path from } C_x \textit{ to } C_y \textit{ on } \mathcal{C} \textit{ of size } s\}$$

will be called the **single-link** *distance of x and y with respect to $\mathcal{C}$.*

Theorem 2.3.1 *For any $S \subset \mathbf{R}^d$ and any $\mathcal{A}$-clustering $\mathcal{C}$ of S, $d_{SL}^{\mathcal{C}}$ is an $\mathcal{A}$-UPM.*

Proof:

1. For any $C \in \mathcal{C}$, the path (C, C) has size 0. So:

$$\forall x \in S : d_{SL}^{\mathcal{C}}(x, x) = 0.$$

2. To any path $(C_x \equiv C_0, C_1, \ldots, C_k \equiv C_y)$, there corresponds a path $(C_y \equiv C_k, C_{k-1}, \ldots, C_0 \equiv C_x)$ of the same size. So:

$$d_{SL}^{\mathcal{C}}(x, y) = d_{SL}^{\mathcal{C}}(y, x).$$

3. Let $x, y, z \in S$. Let us fix, for the moment, a path $(C_x, \dots, C_z)$ of size s_1 and a path $(C_z, \dots, C_y)$ of size s_2. Then, the path:

   $$(C_x, \dots, C_z, \dots, C_y)$$

   has size $\max\{s_1, s_2\}$. So $d^{\mathcal{C}}_{SL}(x, y) \leq \max\{s_1, s_2\}$. Taking the minimum over all paths from C_x to C_z and all paths from C_z to C_y we get:

   $$d^{\mathcal{C}}_{SL}(x, y) \leq \max\{d^{\mathcal{C}}_{SL}(x, z), d^{\mathcal{C}}_{SL}(z, y)\},$$

   the ultrametric inequality.

4. Finally:

   $$\forall y \in C_x : d^{\mathcal{C}}_{SL}(x, y) \leq \rho(C_x, C_x) = 0$$

   so $C_x \subset {d^{\mathcal{C}}_{SL}}^{-1}(x, .)(\{0\})$. If, however, $y \in S \setminus C_x$, then all paths from C_x to C_y include a first step from C_x to some $C \in \mathcal{C}$, $C \neq C_x$. But then:

   $$\bar{C} \cap \bar{C}_x = \emptyset \Rightarrow \rho(C, C_x) \geq \rho(\bar{C}, \bar{C}_x) > 0 \Rightarrow d^{\mathcal{C}}_{SL}(x, y) > 0.$$

   So:

   $${d^{\mathcal{C}}_{SL}}^{-1}(x, .)(\{0\}) = C_x \Rightarrow \{{d^{\mathcal{C}}_{SL}}^{-1}(x, .)(\{0\}),\ x \in S\} = \mathcal{C},$$

   an $\mathcal{A}$-clustering of S.

□

Definition 2.3.4 *The $\mathcal{A}$-hierarchy $(\mathcal{H}, \{\mathcal{P}_i\}_{i=0}^{k}, h) = m^{-1}(d^{\mathcal{C}}_{SL})$ corresponding to $d^{\mathcal{C}}_{SL}$ through the map m of Theorem 2.2.1 is called the* **single-link** *hierarchy with respect to $\mathcal{C}$.*

The choice of $d^{\mathcal{C}}_{SL}$ (among other UPM that can be based on the same $\mathcal{A}$-clustering $\mathcal{C}$) might seem arbitrary. The following theorem gives a reason.

Theorem 2.3.2 *Let $S \subset \mathbf{R}^d$ and $\mathcal{C}$ an $\mathcal{A}$-clustering of S. Let $\mathbf{D}(\mathcal{C})$ be the set of all $\mathcal{A}$-UPM d such that:*

- *$\forall x, y \in S : d(x, y) \leq \rho(x, y)$ and*

- $\{d^{-1}(x,.)(\{0\}),\ x \in S\} = \mathcal{C}$.

Then:

$$\forall d \in \mathbf{D}(\mathcal{C}),\ \forall x,y \in S : d(x,y) \leq d_{SL}^{\mathcal{C}}(x,y) \leq \rho(x,y).$$

Proof:

- An obvious path from C_x to C_y is just (C_x, C_y) with size $\rho(C_x, C_y) \leq \rho(x,y)$. Taking the minimum over all paths from C_x to C_y:

$$d_{SL}^{\mathcal{C}}(x,y) \leq \rho(x,y).$$

- Let $r = \text{card}(\mathcal{C})$. Fix $x, y \in S$. Let $x \in C_x \in \mathcal{C}$ and $y \in C_y \in \mathcal{C}$. A path from C_x to C_y of the form $(C_x, \ldots, C_p, \ldots, C_p, \ldots, C_y)$ cannot have size less than the same path without the inner cycle $(C_p, \ldots, C_p)$. So it is safe to assume that the optimal path from C_x to C_y has, at most, r vertices. Let

$$(C_x \equiv C_0, C_1, C_2, \ldots, C_k \equiv C_y)$$

be such a path. Choose $\epsilon > 0$.

For $0 \leq i \leq k-1$, choose $y_i \in C_i$ and $x_{i+1} \in C_{i+1}$ so that:

$$\rho(x_{i+1}, y_i) < \rho(C_i, C_{i+1}) + \epsilon.$$

For any $d \in \mathbf{D}(\mathcal{C})$ the ultrametric inequality implies that:

$$\begin{aligned} d(x,y) &\leq \max\{d(x,x_k), d(x_k,y)\} \\ &\leq \max\{d(x,y_{k-1}), d(y_{k-1},x_k), d(x_k,y)\} \\ &\leq \ldots \\ &\leq \max\{d(x,y_0), d(y_0,x_1), d(x_1,y_1), \ldots, d(x_k,y)\} \\ &= \max_{1 \leq i \leq k} d(y_{i-1}, x_i) \end{aligned}$$

because distances within the clusters $C_0, \ldots, C_k$ are 0 (Lemma 2.2.1). Then, by assumption:

$$\begin{aligned} d(x,y) &\leq \max_{1 \leq i \leq k} \rho(y_{i-1}, x_i) \\ &\leq \max_{1 \leq i \leq k} \rho(C_{i-1}, C_i) + k\epsilon \\ &\leq \max_{1 \leq i \leq k} \rho(C_{i-1}, C_i) + r\epsilon. \end{aligned}$$

Letting $\epsilon \downarrow 0$:

$$d(x,y) \leq \max_{1\leq i\leq k} \rho(C_{i-1},C_i) = \text{size}(C_0,C_1,\ldots,C_k).$$

Since this is true for any path from C_x to C_y with $\leq r$ vertices, it is also true for the optimal path. So $d(x,y) \leq d^{\mathcal{C}}_{SL}(x,y)$.

This completes the proof. □

2.4 Single-link algorithms

The definition and treatment of single-link hierarchies and distances in the previous section is somewhat different from the traditional approach found in the literature. In that traditional treatment, a finite set $S = \{x_1,x_2,\ldots,x_n\}$ of observations is specified (possibly the values $X_1(\omega),X_2(\omega),\ldots,X_n(\omega)$ of iid random variables) and distances $d^S_{SL}(x_i,x_j)$ are defined only on the finite set S. Notice, however, that this can now be considered a special case of single-link distances.

Definition 2.4.1 *Let $\mathcal{A}_s$ be the class of singletons $\{x\}$, $x \in \mathbf{R}^d$. Let $S = \{x_1,x_2,\ldots,x_n\}$ be a finite subset of $\mathbf{R}^d$. Then $C_S = \{\{x_i\},\ 1 \leq i \leq n\}$ is an $\mathcal{A}_s$-clustering of S.*

We define the (classical) single-link distance on $S \times S$ as:

$$d^S_{SL}(x_i,x_j) = d^{C_S}_{SL}(x_i,x_j),\ 1 \leq i,j \leq n.$$

Remark 2.4.1 As the following example shows, $d^S_{SL}(x,y)$ does not depend only on x and y but on the whole set S:
On the real line, let ρ be the usual metric. Then:

$$d^{\{0,5\}}_{SL}(0,5) = \rho(0,5) = 5$$

but

$$d^{\{0,2,5\}}_{SL}(0,5) = \min\{\rho(0,5), \max\{\rho(0,2),\rho(2,5)\}\} = 3.$$

Finding efficient algorithms to compute single-link distances (and thus form single-link hierarchies as in Theorem 2.2.1) is going to be our next priority. A very popular method (providing an algorithm that computes the matrix $\{d^{\mathcal{S}}_{SL}(x_i, x_j)\}_{i,j=1}^{n}$ in just $O(n^2)$ steps is based on the *minimal spanning tree*. As in the previous section, we will present this concept in the more general setting of $d^{\mathcal{C}}_{SL}$ distances on $S \times S$, with respect to a certain clustering $\mathcal{C}$ of S.

Remark 2.4.2 Because of Lemma 2.2.1, computing $d^{\mathcal{C}}_{SL}$ on $S \times S$ is reduced to computing the matrix:

$$\{d^{\mathcal{C}}_{SL}(C_i, C_j)\}_{i,j=1}^{n}$$

where $C_i \in \mathcal{C}$, $1 \leq i \leq n$.

We will now need the following elementary terminology from graph theory.

Definition 2.4.2 *Given a finite graph $G = (V, E)$ with vertices in V and edges in E:*

1. *a tree $T = (V_T, E_T)$ is a subgraph of G (i.e. $V_T \subset V$ and $E_T \subset E$) which is connected and contains no cycles,*

2. *a spanning tree is a tree for which $V_T = V$,*

3. *a weight function on G is a function $w : E \mapsto \mathbf{R}^+$,*

4. *the weight of a tree T is $w(T) = \sum_{e \in E_T} w(e)$, and*

5. *a minimal spanning tree is any tree T_0 for which:*

$$w(T_0) = \min\{w(T),\ T \textit{ is a spanning tree of } G\}.$$

Remark 2.4.3 In general, there can be several minimal spanning trees as in the following example:

$$G = (V, E),\ V = \{1, 2, 3\},\ E = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$$

and

$$w : E \mapsto \mathbf{R}^+ : \; w(\{1,2\}) = w(\{2,3\}) = w(\{3,1\}) = 1.$$

Let:

$$E_1 = \{\{1,2\},\{3,1\}\},$$
$$E_2 = \{\{1,2\},\{2,3\}\},$$
$$E_3 = \{\{2,3\},\{3,1\}\}.$$

Then $T_1 = (V, E_1)$, $T_2 = (V, E_2)$ and $T_3 = (V, E_3)$ are all minimal spanning trees of G.

Remark 2.4.4 If T is a spanning tree of $G = (V, E)$, then for every $v, u \in V$, there exists a unique path from v to u along edges of T that does not use any edge more than once.

Proposition 2.4.1 *Let $\mathcal{C}$ be an $\mathcal{A}$-clustering of $S \subset \mathbf{R}^d$. Consider the complete graph G with vertices in $V = \mathcal{C}$. define:*

$$w : E \mapsto \mathbf{R}^+ : \; e = \{C_i, C_j\} \mapsto \rho(C_i, C_j),$$

for $C_i, C_j \in \mathcal{C}$. Let T be a minimal spanning tree of G. Let $p_{T'}(C_i, C_j)$ be the unique path from C_i to C_j along edges of T that does not use any edge more than once. Define:

$$d_T(C_i, C_j) = size(p_T(C_i, C_j).$$

Then:

$$d_T(C_i, C_j) = d_{SL}^{\mathcal{C}}(C_i, C_j).$$

Proof: Clearly $d_T(C_i, C_j) \geq d_{SL}^{\mathcal{C}}(C_i, C_j)$ (see Definition 2.3.3). Suppose $d_T(C_i, C_j) > d_{SL}^{\mathcal{C}}(C_i, C_j)$ for some $C_i, C_j \in \mathcal{C}$. Then, there is a path from C_i to C_j whose size is less than size$(p_T(C_i, C_j))$. Let:

$$\text{size}(p_T(C_i, C_j)) = \rho(C_k, C_l), \; C_k, C_l \in \mathcal{C}.$$

Let:

$$(C_i \equiv D_0, D_1, \ldots, D_r \equiv C_j)$$

be the path with size $< \rho(C_k, C_l)$. Then $\rho(D_{i-1}, D_i) < \rho(C_k, C_l)$, $1 \leq i \leq r$. Removing the edge (C_k, C_l) from the tree T, divides T into 2 trees $T_i = (V_i, E_i)$ and $T_j = (V_j, E_j)$ with $V_i \cap V_j = \emptyset$ and $E_i \cap E_j = \emptyset$, such that $C_i \in V_i$ and $C_j \in V_j$ (see Figure 2.1). However, the path

$$(C_i \equiv D_0, D_1, \ldots, D_r \equiv C_j)$$

connects C_i to C_j, so $\exists D_k$, with $D_k \in T_i$, $D_{k+1} \in T_j$ and $\rho(D_k, D_{k+1}) < \rho(C_k, C_l)$. So, substituting the edge (C_k, C_l) with (D_k, D_{k+1}) gives a new spanning tree T' with $w(T') < w(T)$, a contradiction. So $d_T(C_i, C_j) = d_{SL}^{\mathcal{C}}(C_i, C_j)$. □

The last proposition implies that the computation of the matrix:

$$\{d_{SL}^{\mathcal{C}}(C_i, C_j)\},\ 1 \leq i, j \leq n$$

reduces to the computation of:

$$\{d_T(C_i, C_j)\},\ 1 \leq i, j \leq n$$

for some minimal spanning tree T. The next step now is to provide an efficient algorithm for computing minimal spanning trees. Several such algorithms exist, proposed by Florek et al., Kruskal and Prim. Details are provided in [LMW84]. Here we will give a version of Prim's algorithm as found in [Hu82].

ALGORITHM: Let $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ be an $\mathcal{A}$-clustering of S. Let $r_{i,j} = \rho(C_i, C_j)$, $1 \leq i, j \leq k$.

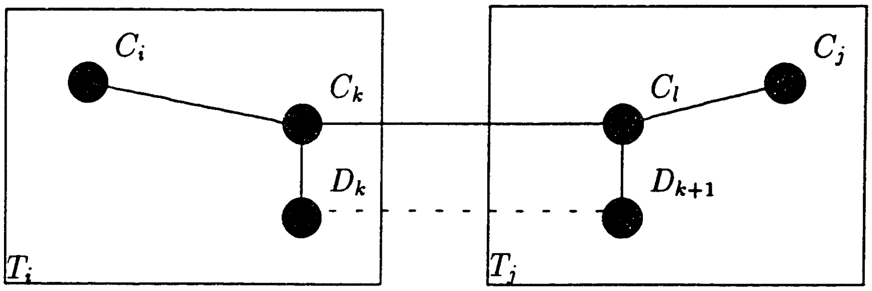


Figure 2.1: Single-link distances and the MST.

Step 0: Let $V := \{C_1\}$, $E := \emptyset$. Let $t_j := r_{1,j}$, $2 \leq j \leq k$.

Step 1: If $V = \{C_{i_1}, C_{i_2}, \dots, C_{i_l}\}$, $1 \leq l < k$, let $t_{i_{l+1}} = \min\{t_j,\ C_j \notin V\}$.
Find h, $1 \leq h \leq l$ such that $t_{i_{l+1}} = r_{i_h, i_{l+1}}$.
Include $C_{i_{l+1}}$ in V and $(C_{i_h}, C_{i_{l+1}})$ in E.

Step 2: If $l + 1 = k$, let $T = (V, E)$ and **stop**.
Otherwise, $\forall C_j \notin V$, $t_j := \min\{t_j, r_{j,i_{l+1}}\}$. Continue with step 1.

The fact that the resulting graph T is a minimal spanning tree is a consequence of the following two lemmas proved in [Hu82] (page 28).

Lemma 2.4.1 *If $\rho(C_i, C_{j_i}) = min_{j \neq i} \rho(C_i, C_j)$ then there exists a minimal spanning tree containing the edge (C_i, C_{j_i}).*

Lemma 2.4.2 *If $T = (V, E)$ is a tree, known to be part of a minimal spanning tree, and:*

$$\exists C_1 \in V,\ C_2 \in \mathcal{C} \setminus V : \ \rho(C_1, C_2) = \min_{C \in V,\, D \in \mathcal{C} \setminus V} \rho(C, D)$$

then there exists a minimal spanning tree including the edge (C_1, C_2) and having T as a subtree.

Remark 2.4.5 Both **step 1** and **step 2** of the algorithm described require $O(k^2)$ operations, so this is an $O(k^2)$-complexity algorithm. (More details on that can be found in [Hu82], as above).

We now have the necessary tools to treat single-link hierarchies defined on a set S, not necessarily finite. In the next chapter, we will use these tools to explore the asymptotic behavior of single-link hierarchies, based on an iid sample.

Chapter 3

Consistency

3.1 Distances as kernels of U-statistics

Up to now, we have treated clustering as a data analytic problem, we have not introduced probability measures and we have not made any distributional assumptions concerning the observed data. In this chapter, we will be introducing a model, appropriate for the study of hierarchical clustering methods, so that we can study consistency of the method described in the previous chapter. To achieve this goal we will make use of the equivalence of hierarchies and ultrametrics that we have proved (Theorem 2.2.1). Because of this result, we can rely exclusively on the study of ultrametrics for a complete description of the behavior of the corresponding hierarchies.

Suppose that d is a distance on $\mathbf{R}^d$. Let $\mathbf{P}$ be a Borel probability measure on $\mathbf{R}^d$ and let:

$$X_1, X_2, \ldots, X_n \, iid \sim \mathbf{P}.$$

We will soon need to study statistics of the form:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(X_i, X_j).$$

Fortunately, such statistics are special cases of U-statistics (of order 2) with kernel d and their asymptotic properties are well understood.

Definition 3.1.1 *Let* $\mathbf{P}$ *be a Borel probability measure on* $\mathbf{R}^d$ *and let:*

$$h : \mathbf{R}^d \times \mathbf{R}^d \mapsto \mathbf{R}$$

be a measurable, symmetric real function. Let $X_1, X_2, \ldots, X_n$ iid $\sim$ $\mathbf{P}$ and define:

$$U_n := U_n(X_1, X_2, \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} h(X_i, X_j).$$

We call U_n the U-statistic (of order 2) with kernel h.

U-statistics have been studied by Hoeffding (see [Hoe48]) and Halmos (see [Hal46]). They generalize the sample mean and in fact, Halmos has proved that when $\mathrm{E}_{\mathbf{P}}(h^2) < \infty$ then U_n is the unbiased estimator of $\mathrm{E}_{\mathbf{P}}(h)$ with smallest variance. However, here we are more interested in the asymptotic properties of U_n. A detailed list of asymptotic results on U-statistics is included in [Ser80], chapter 5. A law of large numbers for U_n is provided by the following:

Theorem 3.1.1 *Let $\mathbf{P}, h, U_n$ be defined as in Definition 3.1.1. Then, if $E_{\mathbf{P}}(|h|) < \infty$ we have :*

$$U_n(X_1, X_2, \ldots, X_n) \overset{n \to \infty}{\longrightarrow} E_{\mathbf{P}}(h)$$

almost surely.

Proof: The result was established by Hoeffding but an easier proof based on the reversed martingale structure of U_n is given by Berk in [Ber66]. □

In addition, we have the following central limit theorem for U-statistics:

Theorem 3.1.2 *Let $\mathbf{P}, h, U_n$ be defined as in Definition 3.1.1 and define:*

$$h_1 : \mathbf{R}^d \mapsto \mathbf{R} \; : \; x \mapsto \int h(x,y) P(dy)$$

and $V(h_1) := Var_{\mathbf{P}}(h_1(X_1))$. Then, if $E_{\mathbf{P}}(h^2) < \infty$:

$$\mathcal{L}(\sqrt{n}(U_n - E_{\mathbf{P}}(h))) \overset{n \to \infty}{\longrightarrow} N(0, 4V(h_1)).$$

Proof: The original proof was given by Hoeffding in [Hoe48]. In this form, the theorem is proved in [Dud89], pages 337-339. □

Remark 3.1.1 Notice that a distance d is a symmetric function which is equal to 0 on the diagonal. Therefore, the difference between:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(X_i, X_j) \tag{3.1}$$

and:

$$U_n = \frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d(X_i, X_j) \tag{3.2}$$

is just the difference between the scaling factors:

$$\frac{1}{n^2} \text{ and } \frac{1}{n(n-1)}.$$

Because of this, we will have no difficulty applying the asymptotic results for (3.2) (such as Theorem 3.1.1 and Theorem 3.1.2) to (3.1).

Remark 3.1.2 There is also no difficulty in extending Theorem 3.1.1 and Theorem 3.1.2 to functions $h : \mathbf{R}^d \mapsto \mathbf{R}^k$. In that case $V(h_1)$ of Theorem 3.1.2 is the $k \times k$ covariance matrix:

$$V(h_1) := \mathrm{Cov}_{\mathbf{P}}(h_{11}(X_1), \ldots, h_{1k}(X_1)).$$

3.2 Clustered measures

Let us begin our study of the asymptotics of single-link distances by examining consistency. In this and the following chapters, ρ will denote the euclidean distance on $\mathbf{R}^d$ unless otherwise noted.

Definition 3.2.1 *A Borel probability measure* $\mathbf{P}$ *on* $\mathbf{R}^d$ *will be called* $\mathcal{A}$*-clustered if there exists an* $\mathcal{A}$*-clustering of supp*$(\mathbf{P})$*. If* $\mathcal{A}$ *is the class of compact and connected sets, then* $\mathbf{P}$ *will be simply called clustered.*

Remark 3.2.1 For a general class of sets $\mathcal{A}$ and an $\mathcal{A}$-clustered measure $\mathbf{P}$, there might exist more than one $\mathcal{A}$-clustering of supp($\mathbf{P}$). For an example, consider $\mathcal{A} = \{$ finite subsets of $\mathbf{R}^d\}$ and take $\mathbf{P}$ to be any probability measure with $1 < \text{card}(\text{supp}(\mathbf{P})) < \infty$.

The next proposition shows that such ambiguity is avoided in the case of clustered measures:

Proposition 3.2.1 *Let* $\mathbf{P}$ *be a clustered measure in* $\mathbf{R}^d$. *If* $\mathcal{C}$ *is a clustering of supp*($\mathbf{P}$)*, then:*

$$\mathcal{C} = \{C_x,\ x \in \mathit{supp}(\mathbf{P})\}$$

where C_x *is the unique connected component of supp*($\mathbf{P}$) *containing* x.

Proof: Let $\mathcal{C} = \{A_1, A_2, \ldots, A_r\}$. Take $x \in A_i$ for some $i : 1 \leq i \leq r$. Since A_i is connected:

$$A_i \subset C_x.$$

By Definition 2.1.3: $\bar{A}_k \cap \bar{A}_m = A_k \cap A_m = \emptyset$ for $1 \leq k < m \leq r$. So, there exist open sets U_k, U_m:

$$A_k \subset U_k,\ A_m \subset U_m,\ U_k \cap U_m = \emptyset.$$

Then:

$$C_x \subset \text{supp}(\mathbf{P}) = A_1 \cup A_2 \cup \ldots \cup A_r \subset U_1 \cup U_2 \cup \ldots \cup U_r.$$

Since C_x is connected, there is a unique j, $1 \leq j \leq r$ such that $C_x \subset U_j$. Since $C_x \cap A_i \neq \emptyset$, $j = i$. So, $C_x \subset U_i$ and since $C_x \subset \text{supp}(\mathbf{P})$:

$$C_x \subset U_i \cap \text{supp}(\mathbf{P}) = A_i.$$

Therefore, $C_x = A_i$. We conclude that:

$$\mathcal{C} = \{C_x,\ x \in \text{supp}(\mathbf{P})\}.$$

□

Remark 3.2.2 Proposition 3.2.1 shows, in particular, that supp($\mathbf{P}$) has a finite number of connected components.

We can now give the following definition:

Definition 3.2.2 *Let* $\mathbf{P}$ *be a clustered measure. Then,* $\mathcal{C}(\mathbf{P})$ *will denote the unique clustering of supp*($\mathbf{P}$).

3.3 Consistency of single-link distances

Let $\mathbf{P}$ be a clustered measure in $\mathbf{R}^d$ and let $X_1, X_2, \ldots, X_n$ *iid* $\sim \mathbf{P}$. If $\mathbf{P}_n$ denotes the empirical measure:

$$\mathbf{P}_n(\omega) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}$$

then $\mathbf{P}_n$ is a clustered measure with:

$$\mathcal{C}(\mathbf{P}_n) = S_n := \{\{X_i\},\ 1 \leq i \leq n\}$$

for all ω.

Therefore, given $X_1, X_2, \ldots, X_n$ *iid* $\sim \mathbf{P}$, we can define two different *single-link* distances. One is $d_{SL}^{\mathcal{C}(\mathbf{P})}$ that is defined on supp($\mathbf{P}$) × supp($\mathbf{P}$) with respect to the clustering $\mathcal{C}(\mathbf{P})$ (Definition 2.3.3). The other is $d_{SL}^{\mathcal{C}(\mathbf{P}_n)}$ defined on $S_n \times S_n$ with respect to the clustering $\mathcal{C}(\mathbf{P}_n)$ (Definition 2.3.3 but also Definition 2.4.1). Therefore, both distances are defined on $S_n \times S_n$, although only $d_{SL}^{\mathcal{C}(\mathbf{P}_n)}$ is observable.

What follows is the main result of this chapter. It shows that, as n goes to ∞, $d_{SL}^{\mathcal{C}(\mathbf{P}_n)}$ converges (uniformly on the sample) to $d_{SL}^{\mathcal{C}(\mathbf{P})}$.

Theorem 3.3.1 *Let $\mathbf{P}$ be a clustered measure in $\mathbf{R}^d$ and let:*

$$X_1, X_2, \ldots, X_n \; iid \sim \mathbf{P}.$$

If $\mathbf{P}_n$ denotes the empirical measure, then, for $n \in \mathbf{N}$ and $1 \leq i,j \leq n$, we define:

$$\Delta_n(X_i, X_j) = d_{SL}^{\mathcal{C}(\mathbf{P}_n)}(X_i, X_j) - d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j).$$

Then:

- $\Delta_n(X_i, X_j) \geq 0$ *a.s. for* $1 \leq i,j \leq n$ *and*
- $\lim_{n\to\infty} \max_{1\leq i,j\leq n} \Delta_n(X_i, X_j) = 0$ *a.s.*.

Proof: Let $\mathcal{C}(\mathbf{P}) = \{C_1, C_2, \ldots, C_r\}$. If $r = 1$ then let $\delta := \infty$, otherwise:

$$\delta := \min_{1\leq i<j\leq r} \rho(C_i, C_j).$$

Because of the definition of a clustering (Definition 2.1.3) and the fact that ρ metrizes the usual topology: $\delta > 0$. Choose and fix ϵ such that $0 < \epsilon < \delta$. Since $\mathrm{supp}(\mathbf{P}) = \cup_{i=1}^{r} C_i$ is compact, there exists a finite number (say $k(\epsilon)$) of open balls with radius $\epsilon/4$ covering $\mathrm{supp}(\mathbf{P})$. Let $\mathcal{B}$ be such a cover. Consider now the following lemma:

Lemma 3.3.1 *Let $\mathbf{P}, \mathcal{C}, \Delta_n, \epsilon, \mathcal{B}$ and $k(\epsilon)$ be defined as above. If, for some n, all $k(\epsilon)$ balls in $\mathcal{B}$ contain at least one observation each, then:*

$$\max_{1 \leq i,j \leq n} \Delta_n(X_i, X_j) < \epsilon.$$

Proof: (of Lemma 3.3.1)
First note that:

$$d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j) \leq d_{SL}^{\mathcal{C}(\mathbf{P}_n)}(X_i, X_j),\ 1 \leq i,j \leq n.$$

Indeed, if $X_i \in C_i$, $X_j \in C_j$, then to every path from X_i to X_j on $\mathcal{C}(\mathbf{P}_n)$, there corresponds a path from C_i to C_j on $\mathcal{C}(\mathbf{P})$ having edges smaller or equal to the corresponding edges of the path on $\mathcal{C}(\mathbf{P}_n)$. So:

$$\Delta_n(X_i, X_j) \geq 0.$$

On the other hand:

$$d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j) = \min\{s : \exists \text{ path from } C_i \text{ to } C_j \text{ with size s}\}.$$

Let:

$$(C_i \equiv D_0, D_1, \ldots, D_k \equiv C_j)$$

be one of the paths that achieve the above minimum. We can now construct a path from X_i to X_j on $\mathcal{C}(\mathbf{P}_n)$ with size $< d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j) + \epsilon$, using the assumption of the lemma.

To do that, we begin by choosing observations:

$$X_i \equiv X_{i_0}, X'_{i_0}, X_{i_1}, X'_{i_1}, \ldots, X_{i_k}, X'_{i_k} \equiv X_j$$

such that:

- $X_{i_l}, X'_{i_l} \in D_l$, $0 \leq l \leq k$ and
- $\rho(X'_{i_{l-1}}, X_{i_l}) \leq \rho(D_{l-1}, D_l) + \epsilon$.

We can do this as follows:

For l such that $1 \leq l \leq k$, $\exists a'_{l-1} \in D_{l-1}$, $a_l \in D_l$ such that $\rho(a'_{l-1}, a_l) = \rho(D_{l-1}, D_l)$ (because the sets D_l and D_{l-1} are compact). Then $\exists B'_{l-1}, B_l \in \mathcal{B}$ such that $a'_{l-1} \in B'_{l-1}$ and $a_l \in B_l$. Finally, using the assumption of the lemma:

$$\exists X'_{i_{l-1}} \in B'_{l-1} \text{ and } X_{i_l} \in B_l.$$

Then:

$$\begin{aligned}
\rho(X'_{i_{l-1}}, X_{i_l}) &\leq \rho(X'_{i_{l-1}}, a'_{l-1}) + \rho(a'_{l-1}, a_l) + \rho(a_l, X_{i_l}) \\
&< \epsilon/2 + \rho(D_{l-1}, D_l) + \epsilon/2 \\
&\leq \epsilon + d_{SL}^{C(\mathbf{P})}(X_i, X_j).
\end{aligned}$$

We must now complete the path from X_i to X_j by inserting paths from X_{i_l} to X'_{i_l} of size $< \epsilon$ for $0 \leq l \leq k$. Notice that, thanks to our initial choice of ϵ, no ball $B \in \mathcal{B}$ intersects more than one of the clusters $C_1, C_2, \ldots, C_r$. Concentrating now on the cluster D_l under consideration, we define:

- $\mathcal{B}_l$ to be the subcover of $\mathcal{B}$ covering D_l,
- S_l to be the set of observations contained in D_l (these include, in particular X_{i_l} and X'_{i_l}),
- S_X to be the set of observations from S_l reachable from X with paths of size $< \epsilon$ for $X \in S_l$,
- $\mathcal{E}_X$ to be the set of balls from $\mathcal{B}_l$ containing observations from S_X and
- $E_X := \cup_{B \in \mathcal{E}_X} B$ (an open set).

Note that:

- $S_l = \cup_{X \in S_l} S_X$,
- $\mathcal{B}_l = \cup_{X \in S_l} \mathcal{E}_X$ (since all balls contain observations) and
- $D_l \subset \cup_{X \in S_l} E_X$.

Also, for $X, Y \in S_l$, either:

$$S_X = S_Y \Leftrightarrow \mathcal{E}_X = \mathcal{E}_Y \Leftrightarrow E_X = E_Y$$

or

$$S_X \cap S_Y = \emptyset \Leftrightarrow \mathcal{E}_X \cap \mathcal{E}_Y = \emptyset \Leftrightarrow E_X \cap E_Y = \emptyset.$$

Then:

- $D_l \subset \cup_{X \in S_l} E_X$ and
- D_l is a connected set

together imply that:

$$\forall X, Y \in S_l : \; E_X = E_Y \Rightarrow S_X = S_Y.$$

In particular, $S_{X_{i_l}} = S_{X'_{i_l}}$ and so there exists a path from X_{i_l} to X'_{i_l} of size $< \epsilon$. These path segments (for $0 \leq l \leq k$) complete a path from X_i to X_j of size s with:

$$\begin{aligned} s &< \max\{\epsilon, \epsilon + d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j)\} \\ &= \epsilon + d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j) \end{aligned}$$

So:

$$\begin{aligned} & d_{SL}^{\mathcal{C}(\mathbf{P}_n)}(X_i, X_j) < \epsilon + d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j) \\ & \Rightarrow \Delta_n(X_i, X_j) < \epsilon, \; 1 \leq i, j \leq n \\ & \Rightarrow \max_{1 \leq i, j \leq n} \Delta_n(X_i, X_j) < \epsilon \end{aligned}$$

proving the lemma.

□

Now, to complete the proof of Theorem 3.3.1 note that we can always assume that all $k(\epsilon)$ balls of the cover have positive probability. If some of them have probability 0, we can discard them and still be able to cover supp($\mathbf{P}$) with the rest. Let $p := \min_{B \in \mathcal{B}} \mathbf{P}(B) > 0$.

Now, using Lemma 3.3.1:

$$\begin{aligned}\Pr(\max_{i,j}\delta_n(X_i,X_j)\geq\epsilon) &\leq \Pr(\exists\, B\in\mathcal{B}\text{ containing no observations })\\ &\leq \sum_{B\in\mathcal{B}}\Pr(B\text{ is empty })\\ &\leq \sum_{B\in\mathcal{B}}[1-\mathbf{P}(B)]^n\\ &\leq k(\epsilon)(1-p)^n.\end{aligned}$$

So, if:

$$A_n^\epsilon := [\max_{i,j}\Delta_n(X_i,X_j)\geq\epsilon]$$

then:

$$\Pr(A_n^\epsilon)\leq k(\epsilon)(1-p)^n$$

$$\Rightarrow \sum_{n=1}^{\infty}\Pr(A_n^\epsilon)\leq k(\epsilon)\sum_{n=1}^{\infty}(1-p)^n = k(\epsilon)\cdot\frac{1-p}{p}<\infty$$

since $p>0$. So, by the Borel-Cantelli lemma:

$$\Pr(\limsup A_n^\epsilon)=0\Rightarrow \max_{i,j}\Delta_n(X_i,X_j)\longrightarrow 0\text{ a.s.}$$

as $n\to\infty$.

□

The last theorem provides us with the means to study statistics based on single link distances:

Definition 3.3.1 *Let* $\mathbf{P}$ *be a clustered measure in* $\mathbf{R}^d$ *and let:*

$$\mathcal{C}(\mathbf{P}) := \{C_1, C_2, \ldots, C_n\}.$$

Since the length of the longest edge of any minimal spanning tree on $\mathcal{C}(\mathbf{P})$ *must be equal to* $\max_{1\leq i,j\leq n} d_{SL}^{\mathcal{C}(\mathbf{P})}(C_i,C_j)$ *(see Proposition 2.4.1), we can define:*

- $M(\mathbf{P})$:= *the length of the longest edge of any minimal spanning tree of* $\mathcal{C}(\mathbf{P})$ *and*
- $D(\mathbf{P}) := \int\int d_{SL}^{\mathcal{C}(\mathbf{P})}(x,y)\mathbf{P}(dx)\mathbf{P}(dy)$.

Remark 3.3.1 $M(\mathbf{P})$ is the largest single-link distance on $\mathcal{C}(\mathbf{P})$ while $D(\mathbf{P})$ is the average single-link distance on $\mathcal{C}(\mathbf{P})$.

Theorem 3.3.2 *Let* $\mathbf{P}$ *be a clustered measure on* $\mathbf{R}^d$ *and let*

$$X_1, X_2, \ldots, X_n \; iid \sim \mathbf{P}.$$

If $\mathbf{P}_n$ *is the empirical measure then, as* $n \to \infty$*:*

$$\begin{aligned} M(\mathbf{P}_n) &\longrightarrow M(\mathbf{P}) \; a.s. \; and \\ D(\mathbf{P}_n) &\longrightarrow D(\mathbf{P}) \; a.s. \end{aligned}$$

Proof: The result for $M(\mathbf{P})$ is obtained by direct application of Theorem 3.3.1:

$$M(\mathbf{P}_n) - M(\mathbf{P}) \leq \max_{i,j} \Delta_n(X_i, X_j) \longrightarrow 0 \text{ a.s.}$$

as $n \to \infty$.

The case of $D(\mathbf{P})$ needs some more work. Since $0 \leq d_{SL}^{\mathcal{C}(\mathbf{P})} \leq \rho$ on supp($\mathbf{P}$) (Theorem 2.3.2) and supp($\mathbf{P}$) is compact:

$$D(\mathbf{P}) = \int\int d_{SL}^{\mathcal{C}(\mathbf{P})}(x,y)\mathbf{P}(dx)\mathbf{P}(dy) < \infty.$$

Then, the U-statistic with kernel $d_{SL}^{\mathcal{C}(\mathbf{P})}$:

$$U_n := \frac{2}{n(n-1)} \cdot \sum_{i<j} d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j)$$

satisfies the strong law of large numbers (Theorem 3.1.1):

$$U_n \longrightarrow D(\mathbf{P}) \; a.s. \tag{3.3}$$

Then:

$$\begin{aligned} & |D(\mathbf{P}_n) - D(\mathbf{P})| \\ \leq \; & |\frac{1}{n^2} \sum_{i,j} d_{SL}^{\mathcal{C}(\mathbf{P}_n)}(X_i, X_j) - \frac{1}{n^2} \sum_{i,j} d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j)| \end{aligned}$$

$$
\begin{aligned}
+ \quad & |\frac{1}{n^2}\sum_{i,j} d_{SL}^{C(\mathbf{P})}(X_i, X_j) - U_n| \\
+ \quad & |U_n - D(\mathbf{P})| \\
\leq \quad & \max_{1\leq i,j\leq n} |d_{SL}^{C(\mathbf{P}_n)}(X_i, X_j) - d_{SL}^{C(\mathbf{P})}(X_i, X_j)| \\
+ \quad & |\frac{1}{n^2} - \frac{1}{n(n-1)}| \cdot |\sum_{i,j} d_{SL}^{C(\mathbf{P})}(X_i, X_j)| \\
+ \quad & |U_n - D(\mathbf{P})| \\
\leq \quad & \max_{1\leq i,j\leq n} \Delta_n(X_i, X_j) + \frac{1}{n} \cdot |U_n| + |U_n - D(\mathbf{P})| \\
\longrightarrow \quad & 0 \; a.s.
\end{aligned}
$$

by Theorem 3.3.1 and (3.3). □

Chapter 4

Asymptotics on the real line

4.1 Spacings

The fact that the statistic $D(\mathbf{P}_n)$ introduced in Chapter 3 exhibited similarities to a U-statistic, encourages us to attempt a more detailed study of the asymptotic behavior of $D(\mathbf{P}_n)$. More specifically, one would hope that $D(\mathbf{P}_n)$ behaves similarly to:

$$\int\int d_{SL}^{C(\mathbf{P})}(x,y)\mathbf{P}_n(dx)\mathbf{P}_n(dy)$$

which cannot be observed (since $\mathbf{P}$ is not known) but which is a U-statistic.

Notice that, to prove consistency for $D(\mathbf{P}_n)$, we used the fact that:

$$\lim_n \max_{i,j} \Delta_n(X_i, X_j) = 0 \text{ a.s.}$$

(Theorem 3.3.1). To prove asymptotic normality, we will need something stronger, namely:

$$\max_{i,j} \Delta_n(X_i, X_j) = o_p(1/\sqrt{n}).$$

As it will turn out, this is only true when $d = 1$.

Clustering on the real line has been studied in the literature: [Har78] has studied the asymptotics of k-means clustering on the real line and more recently [But86] and [But88] have dealt with optimal clustering on the real line. Hartigan has also proved *set-consistency* for the single-link method on the line ([Har81]) and showed that *set-consistency* fails in higher dimensions.

Altrenative approaches based on change-point detection (see also [KS89]) are reviewed in [Eub88]. The approach based on $D(\mathbf{P}_n)$ in this chapter is new but bears some resemblance to the central limit theorem for functions of the nearest neighbor distances proved in [BB83].

What makes things easier on the real line is the simple form of compact and connected sets on $\mathbf{R}$. These are simply the closed intervals. To understand the behavior of $D(\mathbf{P}_n)$, we must first study the spacings formed between observations on $\mathbf{R}$. We will adopt the following notation:

Definition 4.1.1 *Let* $\mathbf{P}$ *be such that* $supp(\mathbf{P}) = I = [a, b]$ *(where* $a \leq b$*).*
Let $X_1, X_2, \ldots, X_n$ *be iid* $\sim \mathbf{P}$ *and let* $X_0 := a$ *and* $X_{n+1} := b$*.*
The **spacings** $Y_0, Y_1, Y_2, \ldots, Y_n$ *of the* $X_1, X_2, \ldots, X_n$ *are defined by:*

$$Y_i := X_{(i+1)} - X_{(i)} \text{ for } 0 \leq i \leq n.$$

The **maximum spacing** Z_n *of the* $X_1, X_2, \ldots, X_n$ *is defined by:*

$$Z_n := \max_{0 \leq i \leq n} Y_i.$$

Several results are known about spacings. A detailed review of such results can be found in [Pyk65]. On the subject of uniform spacings, a standard reference is [Dar53]. Here, however, we will only make use of the oldest of these results which is due to Paul Lévy ([Lévy39]).

Lemma 4.1.1 (Paul Lévy 1939) *When* $X_1, X_2, \ldots, X_n$ *are iid* $\sim U[0, 1]$ *then, for every* $t \in \mathbf{R}$*:*

$$\lim_n \Pr(nZ_n - \log n < t) = e^{-e^{-t}}.$$

Remark 4.1.1 The previous lemma implies, in particular, that for uniform random variables:

$$Z_n = O_p(\log n / n).$$

It will be useful to generalize this result to other clustered measures on $\mathbf{R}$. In fact, we can prove the following:

Proposition 4.1.1 *Let* $\mathbf{P}$ *be a clustered measure on* $\mathbf{R}$ *having a density* f *(with respect to Lebesgue measure) that satisfies:*

$$\inf\{f(x), x \in supp(\mathbf{P})\} \geq \delta > 0.$$

Then:

$$\max_{1 \leq i,j \leq n} \Delta_n(X_i, X_j) = O_p(\log n/n).$$

Proof: There are two cases to consider:

Case I: card$(\mathcal{C}(\mathbf{P})) = 1$.
Let $F(x) = \int_{-\infty}^{x} f(t)dt$ be the distribution function of $\mathbf{P}$.
Since $d_{SL}^{\mathcal{C}(\mathbf{P})} \equiv 0$ we have

$$\begin{aligned}
\max_{1 \leq i,j \leq n} \Delta_n(X_i, X_j) &= \max_{1 \leq i,j \leq n} d_{SL}^{\mathcal{C}(\mathbf{P}_n)}(X_i, X_j) \\
&\leq \max_{0 \leq s \leq n} (X_{(s+1)} - X_{(s)}) \\
&= (1/\delta) \max_{0 \leq s \leq n} \delta(X_{(s+1)} - X_{(s)}) \\
&\leq (1/\delta) \max_{0 \leq s \leq n} \int_{X_{(s)}}^{X_{(s+1)}} f(x)dx \\
&= (1/\delta) \max_{0 \leq s \leq n} [F(X_{(s+1)}) - F(X_{(s)})] \\
&= O_p(\log n/n)
\end{aligned}$$

by Lemma 4.1.1, because $F(X_1), \ldots F(X_n) \sim U[0,1]$.

Case II: card$(\mathcal{C}(\mathbf{P})) > 1$.
Suppose $\mathcal{C}(\mathbf{P}) = \{I_1, I_2, \ldots, I_k\}$ where $k > 1$. Let $I_t = [a_t, b_t]$, with

$$a_t < b_t, \ 1 \leq t \leq k$$

and

$$b_t < a_{t+1}, \ 1 \leq t \leq k-1.$$

We can reduce this case to case I, by considering the function:

$$h(x) = x - \sum_{t=1}^{k-1}(a_{t+1} - b_t)1_{[a_{t+1},\infty)}(x)$$

which maps supp($\mathbf{P}$) onto the interval:

$$J = [a_1, b_k - \sum_{t=1}^{k-1}(a_{t+1} - b_t)]$$

and $\mathbf{P}$ to a measure covered by case I.

The reduction will be based on the following:

Lemma 4.1.2 *If all intervals I_r, $1 \leq r \leq k$ contain observations, then:*

$$\max_{1\leq i,j\leq n} \Delta_n(X_i, X_j) \leq \max_{1\leq i,j\leq n} \Delta_n(h(X_i), h(X_j)).$$

Proof: (of the lemma).
Assume that $X_{(i)} \in I_{t_i}$. Also, let:

$$\mathbf{P}_n^h = \frac{1}{n}\sum_{i=1}^{n}\delta_{h(X_i)}.$$

Note that:

$$d_{SL}^{C(\mathbf{P}_n)}(X_{(i)}, X_{(j)}) = \max_{i\leq m<j} \rho(X_{(m)}, X_{(m+1)}).$$

If $I_{t_i} = I_{t_j}$ then $d_{SL}^{C(\mathbf{P})}(X_{(i)}, X_{(j)}) = 0$ and also:

$$\rho(X_{(m)}, X_{(m+1)}) = \rho(h(X_{(m)}), h(X_{(m+1)})),\ i \leq m < j.$$

So:

$$\begin{aligned}
\Delta_n(X_{(i)}, X_{(j)}) &= d_{SL}^{C(\mathbf{P}_n)}(X_{(i)}, X_{(j)}) \\
&= \max_{i\leq m<j} \rho(X_{(m)}, X_{(m+1)}) \\
&= \max_{i\leq m<j} \rho(h(X_{(m)}), h(X_{(m+1)})) \\
&= d_{SL}^{C(\mathbf{P}_n^h)}(h(X_{(i)}), h(X_{(j)}))
\end{aligned}$$

If $I_{t_i} \neq I_{t_j}$ then, for $i \leq m < j$:

$$\rho(X_{(m)}, X_{(m+1)}) = \begin{cases} \rho(h(X_{(m)}), h(X_{(m+1)})) \\ \text{if} \\ X_{(m)}, X_{(m+1)} \text{ are in the same interval,} \\ \\ \rho(X_{(m)}, b_{t_m}) + \rho(I_{t_m}, I_{t_m+1}) + \rho(X_{(m+1)}, a_{t_{m+1}}) = \\ \rho(h(X_{(m)}), h(X_{(m+1)})) + \rho(I_{t_m}, I_{t_m+1}) \\ \text{if} \\ X_{(m)}, X_{(m+1)} \text{ are in the successive intervals } I_{t_m}, I_{t_m+1}. \end{cases}$$

So:

$$\begin{aligned} d_{SL}^{\mathcal{C}(\mathbf{P}_n)}(X_{(i)}, X_{(j)}) &= \max_{i \leq m < j} \rho(X_{(m)}, X_{(m+1)}) \\ &\leq \max_{1 \leq m < j} \rho(h(X_{(m)}), h(X_{(m+1)})) + \max_{t_i \leq t < t_j} \rho(I_t, I_{t+1}) \\ &= d_{SL}^{\mathcal{C}(\mathbf{P}_n^h)}(h(X_{(i)}), h(X_{(j)})) + d_{SL}^{\mathcal{C}(\mathbf{P})}(X_{(i)}, X_{(j)}). \end{aligned}$$

Therefore:

$$\Delta_n(X_i, X_j) \leq d_{SL}^{\mathcal{C}(\mathbf{P}_n^h)}(h(X_i), h(X_j)) = \Delta_n(h(X_i), h(X_j))$$

proving the lemma. □

Now, by applying case I:

$$\max_{1 \leq i,j \leq n} \Delta_n(h(X_i), h(X_j)) = O_p(\log n / n).$$

Therefore, $\exists M > 0$ such that:

$$\lim_n \Pr\left(\frac{n}{\log n} \max_{1 \leq i,j \leq n} \Delta_n(h(X_i), h(X_j)) > M\right) = 0. \qquad (4.1)$$

So:

$$\Pr\left(\frac{n}{\log n} \max \Delta_n(X_i, X_j) > M\right)$$

$$\begin{aligned}
&\leq && \Pr\left(\frac{n}{\log n}\max \Delta_n(X_i,X_j) > M \mid \text{no empty intervals}\right) \\
&+ && \Pr(\exists \text{ an empty interval}) \\
&\leq && \Pr\left(\frac{n}{\log n}\max \Delta_n(h(X_i),h(X_j)) > M\right) \\
&+ && \sum_{t=1}^{k}(1-\mathbf{P}(I_t))^n \\
&\longrightarrow && 0
\end{aligned}$$

as $n \longrightarrow \infty$ by (4.1).

□

4.2 A central limit theorem

We can now describe the asymptotic distribution of $D(\mathbf{P}_n)$:

Theorem 4.2.1 *Let* $\mathbf{P}$ *be a clustered measure on* $\mathbf{R}$ *having a density* f *(with respect to Lebesgue measure) that satisfies:*

$$\inf\{f(x),\ x \in supp(\mathbf{P})\} \geq \delta > 0.$$

Let $X_1, X_2, \ldots, X_n$ *iid* $\sim \mathbf{P}$ *and* $\mathbf{P}_n$ *the empirical measure. Then:*

$$\lim_n \mathcal{L}(\sqrt{n}(D(\mathbf{P}_n) - D(\mathbf{P})) = N(0, 4\sigma^2)$$

where:

$$\sigma^2 = Var_{\mathbf{P}}(\int d_{SL}^{C(\mathbf{P})}(X_1, x)\mathbf{P}(dx)).$$

Proof: The U-statistic:

$$U_n = \frac{2}{n(n-1)} \cdot \sum_{i>j} d_{SL}^{C(\mathbf{P})}(X_i, X_j)$$

satisfies:

$$\lim_n \mathcal{L}(\sqrt{n}(U_n - D(\mathbf{P}))) = N(0, 4\sigma^2).$$

(Theorem 3.1.2). In addition:

$$\begin{aligned}
& \left|\sqrt{n}\,(D(\mathbf{P}_n) - D(\mathbf{P})) - \sqrt{n}\,(U_n - D(\mathbf{P}))\right| \\
= \;& \sqrt{n}\,|D(\mathbf{P}_n) - U_n| \\
\leq \;& \frac{\sqrt{n}}{n^2} \cdot \sum_{i,j} \left| d_{SL}^{\mathcal{C}(\mathbf{P}_n)}(X_i, X_j) - d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j) \right| \\
+ \;& \sqrt{n} \cdot \left| \frac{1}{n^2} - \frac{1}{n(n-1)} \right| \cdot \sum_{i,j} d_{SL}^{\mathcal{C}(\mathbf{P})}(X_i, X_j) \\
\leq \;& \sqrt{n} \cdot \max_{1 \leq i,j \leq n} \Delta_n(X_i, X_j) + \frac{\sqrt{n}}{n} \cdot |U_n| \\
\xrightarrow{\mathbf{P}} \;& 0
\end{aligned}$$

by Proposition 4.1.1 and the fact that $U_n \xrightarrow{\text{a.s.}} D(\mathbf{P})$ (Theorem 3.1.1). So, $\sqrt{n}(D(\mathbf{P}_n) - D(\mathbf{P}))$ has the same asymptotic distribution as $\sqrt{n}(U_n - D(\mathbf{P}))$. The result follows.

□

Remark 4.2.1 To use Theorem 4.2.1, we need to know whether $\sigma^2 > 0$. There are cases where this is not true, as in the following examples:

Example 1:

Let $\mathbf{P} = U([0,1])$. Then $\mathcal{C}(\mathbf{P}) = \{[0,1]\}$ and therefore:

$$d_{SL}^{\mathcal{C}(\mathbf{P})} \equiv 0.$$

A fortiori, $\mathrm{Var}_{\mathbf{P}}(\int d_{SL}^{\mathcal{C}(\mathbf{P})}(X_1, x)\mathbf{P}(dx)) = 0$. In this case, the decomposition:

$$d_{SL}^{\mathcal{C}(\mathbf{P}_n)} = d_{SL}^{\mathcal{C}(\mathbf{P})} + (d_{SL}^{\mathcal{C}(\mathbf{P}_n)} - d_{SL}^{\mathcal{C}(\mathbf{P})})$$

will not be useful. However, we know the asymptotic distribution of $M(\mathbf{P}_n)$ (see Definition 3.3.1) by Lemma 4.1.1 and that would be more appropriate to use. For other distributions with connected support, one could use the results of [Deh84].

Example 2:

Let $\mathcal{C}(\mathbf{P}) = \{I_1, I_2\}$. Then:

- with probability $\mathbf{P}(I_1)$, $X_1 \in I_1$ and:

$$\int d_{SL}^{C(\mathbf{P})}(X_1,x)\mathbf{P}(dx) = \mathbf{P}(I_1)\cdot 0 + \mathbf{P}(I_2)\cdot \rho(I_1,I_2)$$

- with probability $\mathbf{P}(I_2)$, $X_1 \in I_2$ and:

$$\int d_{SL}^{C(\mathbf{P})}(X_1,x)\mathbf{P}(dx) = \mathbf{P}(I_1)\cdot \rho(I_1,I_2) + \mathbf{P}(I_2)\cdot 0.$$

Therefore:

$$\sigma^2 = \mathrm{Var}(\mathbf{P})\int d_{SL}^{C(\mathbf{P})}(X_1,x)\mathbf{P}(dx) = 0 \Leftrightarrow \mathbf{P}(I_1) = \mathbf{P}(I_2).$$

Remark 4.2.2 Under the assumption $\sigma^2 > 0$, we can construct approximate confidence intervals for $D(\mathbf{P})$. At level α, these intervals would be of the form:

$$(D(\mathbf{P}_n) - 2n^{-1/2}\sigma_n z(\alpha/2), D(\mathbf{P}_n) + 2n^{-1/2}\sigma_n z(\alpha/2))$$

where σ_n^2 is an estimate of σ^2. E.g. we can use:

$$\sigma_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n} d_{SL}^{C(\mathbf{P}_n)}(X_i,X_j) - D(\mathbf{P}_n)\right]^2 .$$

Remark 4.2.3 The general problem of *homogeneity* in a sample can be formulated in the following way:

Let $\mathcal{F}$ be a class of probability distributions on $\mathbf{R}$. Given independent random variables $X_1, X_2, \ldots, X_n$ we want to test the hypothesis:

$$H_0 : \exists F \in \mathcal{F},\ X_1, X_2, \ldots, X_n\ iid \sim F$$

versus the alternative:

$$H_A : \exists i_1 \neq i_2,\ F_{i_1} \neq F_{i_2} :\ X_{i_1} \sim F_{i_1},\ X_{i_2} \sim F_{i_2}.$$

We have just given a solution to this problem when $\mathcal{F}$ is the class of distributions with compact and connected support.

4.3 Measuring hierarchical structure

In Chapter 2 we showed that $d_{SL}^{C(\mathbf{P})}$ was the ultrametric that minimized $\rho - d$ among all ultrametrics d for which $\rho \geq d$ (Theorem 2.3.2). Therefore, we can measure hierarchical structure by how closely $d_{SL}^{C(\mathbf{P})}$ approximates ρ. For this purpose we will define an affine invariant functional CR(**P**)) which takes values in the $[0,1]$ interval so that 0 corresponds to a clustered measure **P** with connected support and 1 corresponds to *perfect* hierarchical structure.

Definition 4.3.1 *Let* **P** *be a clustered measure in* **R** *and let*

$$\rho(x,y) = |x-y|, \ x,y \in \mathbf{R}.$$

First we let:

$$R(\mathbf{P}) := \int\int \rho(x,y)\mathbf{P}(dx)\mathbf{P}(dy).$$

Then we define CR(**P**)*, the cluster ratio of* **P** *to be:*

$$CR(\mathbf{P}) = \begin{cases} 1 & \textit{if } \mathbf{P} \textit{ is a point mass,} \\ D(\mathbf{P})/R(\mathbf{P}) & \textit{otherwise.} \end{cases}$$

The following proposition gives some properties of CR(**P**):

Proposition 4.3.1 *If* **P** *is a clustered measure in* **R***, then:*

1. $0 \leq CR(\mathbf{P}) \leq 1$,
2. $CR(\mathbf{P}) = 0 \Leftrightarrow card(C(\mathbf{P})) = 1$ *and* **P** *is not a point mass,*
3. $CR(\mathbf{P}) = 1 \Leftrightarrow card(supp(\mathbf{P})) = card(C(\mathbf{P})) \leq 2.$

Proof:

1. This follows from the fact that:

$$0 \leq d_{SL}^{C(\mathbf{P})} \leq \rho \text{ on } \mathrm{supp}(\mathbf{P})$$

(see Theorem 2.3.2).

2. If $\mathbf{P}$ is a point mass then $\mathrm{CR}(\mathbf{P}) = 1$. When $\mathbf{P}$ is not a point mass the following are equivalent:

$$\begin{aligned} & \mathrm{CR}(\mathbf{P}) = 0 \\ \Leftrightarrow\ & \int\!\!\int d_{SL}^{\mathcal{C}(\mathbf{P})}(x,y)\mathbf{P}(dx)\mathbf{P}(dy) = 0 \\ \Leftrightarrow\ & d_{SL}^{\mathcal{C}(\mathbf{P})} = 0,\ \mathbf{P}^2\text{-a.s.} \\ \Leftrightarrow\ & \mathrm{card}(\mathcal{C}(\mathbf{P})) = 1. \end{aligned}$$

3. For any clustered measure $\mathbf{P}$:

$$\begin{aligned} & \mathrm{CR}(\mathbf{P}) = 1 \\ \Leftrightarrow\ & R(\mathbf{P}) = D(\mathbf{P}) \\ \Leftrightarrow\ & \int\!\!\int [\rho(x,y) - d_{SL}^{\mathcal{C}(\mathbf{P})}(x,y)]\mathbf{P}(dx)\mathbf{P}(dy) = 0 \\ \Leftrightarrow\ & \rho(x,y) = d_{SL}^{\mathcal{C}(\mathbf{P})}(x,y),\ \mathbf{P}^2\text{-a.s.} \end{aligned}$$

So, if $\mathcal{C}(\mathbf{P}) = \{C_1, C_2, \ldots, C_k\}$, then, for any $1 \le i \le k$ and for any $x, y \in C_i$:

$$\rho(x,y) = d_{SL}^{\mathcal{C}(\mathbf{P})}(x,y) = 0,\ \mathbf{P}^2\text{-a.s.}$$

$$\Rightarrow \mathrm{diam}(C_i) = 0 \Rightarrow \mathrm{card}(\mathrm{supp}(\mathbf{P})) = \mathrm{card}(\mathcal{C}(\mathbf{P})).$$

Let $C_i = \{c_i\}$, $1 \le i \le k$.

- If $k = 1$ then $\mathrm{CR}(\mathbf{P}) = 1$ by definition.
- If $k = 2$, i.e. $\mathrm{supp}(\mathbf{P}) = \{c_1, c_2\}$ then:

$$\begin{aligned} \rho(c_1,c_1) &= d_{SL}^{\mathcal{C}(\mathbf{P})}(c_1,c_1) = 0 \\ \rho(c_2,c_2) &= d_{SL}^{\mathcal{C}(\mathbf{P})}(c_2,c_2) = 0 \\ \rho(c_1,c_2) &= d_{SL}^{\mathcal{C}(\mathbf{P})}(c_1,c_2) \end{aligned}$$

  So: $\rho = d_{SL}^{\mathcal{C}(\mathbf{P})}$, $\mathbf{P}^2$-a.s..

- If $k \ge 3$ and assuming that $c_1 < c_2 < c_3$, we have:

$$d_{SL}^{\mathcal{C}(\mathbf{P})}(c_1,c_3) = \max\{\rho(c_1,c_2), \rho(c_2,c_3)\} < \rho(c_1,c_3).$$

  This contradicts $d_{SL}^{\mathcal{C}(\mathbf{P})} = \rho$, $\mathbf{P}^2$-a.s. and so $\mathrm{CR}(\mathbf{P}) < 1$.

So CR($\mathbf{P}$) = 1 $\Leftrightarrow$ card(supp($\mathbf{P}$)) ≤ 2.

□

We can now consistently estimate CR($\mathbf{P}$) by CR($\mathbf{P}_n$) where $\mathbf{P}_n$ is the empirical measure:

Theorem 4.3.1 *For any clustered measure* $\mathbf{P}$ *on* $\mathbf{R}$ *:*

$$\lim_n CR(\mathbf{P}_n) = CR(\mathbf{P}), \tag{4.2}$$

almost surely.

Proof: The denominator of CR($\mathbf{P}_n$) is a U-statistic.

The numerator is the U-statistic

$$\int\int d_{SL}^{\mathcal{C}(\mathbf{P})}(x,y)\mathbf{P}_n(dx)\mathbf{P}_n(dy)$$

plus the term

$$\int\int [d_{SL}^{\mathcal{C}(\mathbf{P}_n)} - d_{SL}^{\mathcal{C}(\mathbf{P})}](x,y)\mathbf{P}_n(dx)\mathbf{P}_n(dy). \tag{4.3}$$

But (4.3) is bounded by:

$$\max\{[d_{SL}^{\mathcal{C}(\mathbf{P}_n)} - d_{SL}^{\mathcal{C}(\mathbf{P})}](X_i, X_j), 1 \leq i,j \leq n\}.$$

$$= \max_{1 \leq i,j \leq n} \Delta_n(X_i, X_j)$$

which converges to 0 almost surely by Theorem 3.3.1. So what we have is, essentially, the ratio of two U-statistics and the result follows from Theorem 3.1.1. □

Furthermore, we now get an asymptotic distribution for CR($\mathbf{P}_n$) on the real line.

Theorem 4.3.2 *Under the assumptions:*

- $\mathbf{P}$ *is a clustered measure on* $\mathbf{R}$*,*
- *card*($\mathcal{C}(\mathbf{P})$) > 1 *and*

- **P** *has a density* f *(with respect to Lebesgue measure) that satisfies:*

$$\inf\{f(x),\ x \in supp(\mathbf{P})\} \geq \delta > 0,$$

there exists a $\sigma^2_{CR} \geq 0$ *such that:*

$$\lim_n \mathcal{L}(\sqrt{n}(CR(\mathbf{P}_n) - CR(\mathbf{P}))) = N(0, \sigma^2_{CR}).$$

Proof: We begin by defining the 2-dimensional vectors:

- $D^{(2)}(\mathbf{P}) := (D(\mathbf{P}), R(\mathbf{P}))$,
- $D^{(2)}_n(\mathbf{P}, \mathbf{P}_n) := (\int\int d_{SL}^{\mathcal{C}(\mathbf{P})}(x, y)\mathbf{P}_n(x)\mathbf{P}_n(y), R(\mathbf{P}_n))$ and
- $D^{(2)}_n(\mathbf{P}_n) := (D(\mathbf{P}_n), R(\mathbf{P}_n))$.

Clearly:

$$\begin{aligned}
& \| D^{(2)}_n(\mathbf{P}_n) - D^{(2)}_n(\mathbf{P}, \mathbf{P}_n) \|_2 && (4.4)\\
= & \int\int [d_{SL}^{\mathcal{C}(\mathbf{P}_n)} - d_{SL}^{\mathcal{C}(\mathbf{P})}](x, y)\mathbf{P}_n(dx)\mathbf{P}_n(dy) && (4.5)\\
\leq & \max_{1 \leq i,j \leq n} \Delta_n(X_i, X_j) && (4.6)\\
= & o_p(n^{-1/2}) && (4.7)
\end{aligned}$$

because of Proposition 4.1.1.

Define now the 2×2 matrix V by letting:

$$\begin{aligned}
v_{11} &= \mathrm{Var}_{\mathbf{P}}(\int d_{SL}^{\mathcal{C}(\mathbf{P})}(X_1, x)\mathbf{P}(dx))\\
v_{22} &= \mathrm{Var}_{\mathbf{P}}(\int \rho(X_1, x)\mathbf{P}(dx))\\
v_{12} = v_{21} &= \mathrm{Cov}_{\mathbf{P}}(\int d_{SL}^{\mathcal{C}(\mathbf{P})}(X_1, x)\mathbf{P}(dx), \int \rho(X_1, x)(dx))
\end{aligned}$$

and:

$$V = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}.$$

Then, by Remark 3.1.2:

$$\lim_n \mathcal{L}(\sqrt{n}(D^{(2)}_n(\mathbf{P}, \mathbf{P}_n) - D^{(2)}(\mathbf{P})) = N(0, 4V)$$

and, because of (4.7):

$$\lim_n \mathcal{L}(\sqrt{n}(D_n^{(2)}(\mathbf{P}_n) - D^{(2)}(\mathbf{P})) = N(0, 4V).$$

Consider now the function:

$$f : [0, \infty) \times (0, \infty) \mapsto \mathbf{R} : (s, t) \longrightarrow s/t.$$

Since f is differentiable at $(D(\mathbf{P}), R(\mathbf{P}))$:

$$\begin{array}{cl} & \sqrt{n}\,[f(D(\mathbf{P}_n), R(\mathbf{P}_n)) - f(D(\mathbf{P}), R(\mathbf{P}))] \\ = & \dfrac{\sqrt{n}}{R(\mathbf{P})}\,[D(\mathbf{P}_n) - D(\mathbf{P})] - \dfrac{\sqrt{n}D(\mathbf{P})}{R(\mathbf{P})^2}\,[R(\mathbf{P}_n) - R(\mathbf{P})] \\ + & \sqrt{n} \cdot o\left(\| D_n^{(2)}(\mathbf{P}_n) - D^{(2)}(\mathbf{P}) \|_2\right) \\ = & \sqrt{n} \cdot \left(\dfrac{1}{R(\mathbf{P})}, -\dfrac{D(\mathbf{P})}{R(\mathbf{P})^2}\right) \cdot \left(D_n^{(2)}(\mathbf{P}_n) - D^{(2)}(\mathbf{P})\right) \\ + & \sqrt{n} \cdot o\left(\| D_n^{(2)}(\mathbf{P}_n) - D^{(2)}(\mathbf{P}) \|_2\right) \\ \xrightarrow{\mathcal{L}} & N(0, \sigma_{CR}^2) \end{array}$$

where:

$$\begin{aligned} \sigma_{CR}^2 &= \left(\frac{1}{R(\mathbf{P})}, -\frac{D(\mathbf{P})}{R(\mathbf{P})^2}\right) \cdot 4V \cdot \left(\frac{1}{R(\mathbf{P})}, -\frac{D(\mathbf{P})}{R(\mathbf{P})^2}\right)^t \\ &= \frac{4v_{11}}{R(\mathbf{P})^2} - 2\frac{4v_{12}D(\mathbf{P})}{R(\mathbf{P})^3} + \frac{4v_{22}D(\mathbf{P})^2}{R(\mathbf{P})^4}. \end{aligned}$$

□

Chapter 5

Using the edges of the MST

5.1 Statistics related to the MST

In chapter 3, we discussed the consistency of single-link distances. Therefore we know that when cluster structure is present, the largest single-link distance is going to converge to a positive number. Conversely, when the support of the clustered measure is connected, the single-link distances are going to converge uniformly to 0 (see Theorem 3.3.2). However, we will need more information to decide whether those distances are significantly large, when the dimension is arbitrary.

The discussion of asymptotics on the real line, in chapter 4, began with a study of spacings. It is therefore natural to look for multivariate analogues of spacings hoping to reach similar results. One such possibility is explored by [DEMR88]. It has also been suggested that nearest neighbor distances can be used to assess homogeneity of the sample ([Boc85]) and there has been considerable work on the asymptotics of these distances ([Hen83, DH89, ST88, BB83, Hen88]). However, given the already established relation between single-link distances and minimal spanning trees (Proposition 2.4.1), the edges of the MST seem to be a more natural choice.

In the recent literature, considerable attention has been paid to the *total length* of the MST formed by n iid observations ([Ste88, BvR90, AB92]) mainly because of its use in approximating optimal solutions to much harder problems in combinatorial optimization (e.g. the traveling salesman problem). It should be clear, however, that this statistic has little to do with clustering. In a different direction, J. Friedman and L. C. Rafsky have used the minimal spanning tree to construct multivariate nonparametric 2-sample tests (see [FRf79]). In our case however, the quantity that appears to be of interest is the longest edge of the minimal spanning tree.

In this chapter we will give bounds for that edge of the MST of n iid random variables drawn from a probability measure $\mathbf{P}$ with compact and connected support. Some additional assumptions about the density of P (with respect to Lebesgue measure) and the shape of supp($\mathbf{P}$) will be needed.

5.2 Upper bounds

Definition 5.2.1 *Let $B(a, r)$ denote the open ball $\{x \in \mathbf{R}^d :\| x - a \|< r\}$ for $a \in \mathbf{R}^d$, $r > 0$. Then let $D_{a,b} = B(a, \| a - b \|) \cap B(b, \| a - b \|)$, $a, b \in R^d$.*

Definition 5.2.2 *A class $\mathcal{G}$ of subsets of $\mathbf{R}^d$ is going to be called* **ball-like** *if there is a functional $r : \mathcal{G} \mapsto \mathbf{R}^+$ such that:*

- *$r(G_1) = sr(G_2) \Rightarrow \lambda(G_1) = s^d\lambda(G_2)$ for $G_1, G_2 \in \mathcal{G}$ and $s > 0$,*
- *$\exists G_0 \in \mathcal{G}$ such that $r(G_0) = 1$.*

For such an r, and for $G \in \mathcal{G}$, $r(G)$ will be called the radius of G.

Remark 5.2.1 The class $\mathcal{G} = \{D_{a,b}\}_{a,b \in \mathbf{R}^d}$ is *ball-like* with

$$r(D_{a,b}) =\| a - b \| .$$

The following is a crucial property of the MST:

Lemma 5.2.1 *Let $V = \{x_1, x_2, \ldots, x_n\} \subset \mathbf{R}^d$ and let $T = (V, E)$ be an MST of V. Then, $\forall i, j$ such that $(x_i, x_j) \in E$, $D_{x_i,x_j} \cap \{x_1, x_2, \ldots, x_n\} = \emptyset$.*

Proof: Let $\mathcal{C} = \{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$. Since $(x_i, x_j) \in E$, we have: $d^{\mathcal{C}}_{SL}(x_i, x_j) =\| x_i - x_j \|$, by Proposition 2.4.1. If $x_k \in D_{x_i,x_j}$ then the path $p = (x_i, x_k, x_j)$ from x_i to x_j has size equal to:

$$\max\{\| x_i - x_k \|, \| x_j - x_k \|\} <\| x_i - x_j \|= d^{\mathcal{C}}_{SL}(x_i, x_j)$$

which contradicts the definition of *single-link* distances (Definition 2.3.3). □

Now we can give the following upper bound:

Theorem 5.2.1 *Let* $\mathbf{P}$ *satisfy the following assumptions:*

1. *supp*($\mathbf{P}$) *is compact and connected* $\subset \mathbf{R}^d$,

2. $\mathbf{P}$ *has a density* f *(w.r.t. Lebesgue measure) such that:*

$$\inf\{f(x),\ x \in supp(\mathbf{P})\} = \delta > 0,$$

3. *there is a class* $\mathcal{G}$ *of* **ball-like** *sets with radius* $r(\cdot)$*, a* $G_0 \in \mathcal{G}$ *with* $r(G_0) = 1$ *and a constant* $c > 0$ *such that:*

$$\forall x, y \in supp(\mathbf{P}),\ \exists G \in \mathcal{G},\ G \subset D_{x,y} \cap supp(\mathbf{P}) \text{ and } r(G) \geq c \parallel x - y \parallel .$$

Let $X_1, X_2, \ldots, X_n$ *iid* $\sim \mathbf{P}$*, and let* M_n *be the length of the longest edge of any minimal spanning tree of* $\{X_1, X_2, \ldots, X_n\}$*. Then:*

$$\lim_{n\to\infty} \Pr\left(M_n^d \leq \frac{k}{\delta c^d \lambda(G_0)} \frac{\log n}{n}\right) = 1,$$

for all $k > 2$.

Proof:

Let:

$$r_n := \frac{k}{\delta c^d \lambda(G_0)} \cdot \frac{\log n}{n}$$

and let n_0 be chosen so that:

$$n \geq n_0 \Rightarrow \text{diam}(\text{supp}(\mathbf{P}))^d > r_n.$$

Also, for $1 \leq i < j \leq n$, let $G_{i,j} \in \mathcal{G}$ be such that:

$$r(G_{i,j}) \geq c \parallel X_i - X_j \parallel$$

and

$$G_{i,j} \subset D_{X_i,X_j} \cap \text{supp}(\mathbf{P})$$

as guaranteed by the third assumption of the theorem. If $M_n^d > r_n$ then there must exist an edge (X_i, X_j) such that $\parallel X_i - X_j \parallel^d > r_n$ and (because of Lemma 5.2.1),

$$D_{X_i,X_j} \cap \{X_1, X_2, \ldots, X_n\} = \emptyset.$$

So for $n \geq n_0$:

$$
\begin{aligned}
& \Pr(M_n^d > r_n) \\
\leq & \sum_{1 \leq i < j \leq n} \Pr(\| X_i - X_j \|^d > r_n \text{ and } D_{X_i, X_j} \cap \{X_1, X_2, \ldots, X_n\} = \emptyset) \\
\leq & \sum_{1 \leq i < j \leq n} \Pr(D_{X_i, X_j} \cap \{X_1, X_2, \ldots, X_n\} = \emptyset \,\Big|\, \| X_i - X_j \|^d > r_n) \\
\leq & \sum_{1 \leq i < j \leq n} \Pr(G_{i,j} \cap \{X_1, X_2, \ldots, X_n\} = \emptyset \,\Big|\, \| X_i - X_j \|^d > r_n) \\
\leq & \sum_{1 \leq i < j \leq n} E[1_{(R^d - G_{i,j})^{n-2}} \,\Big|\, \| X_i - X_j \|^d > r_n] \\
\leq & \sum_{1 \leq i < j \leq n} E[(1 - 1_{G_{i,j}})^{n-2} \,\Big|\, \| X_i - X_j \|^d > r_n] \\
\leq & \sum_{1 \leq i < j \leq n} E[(1 - \delta \lambda(G_{i,j}))^{n-2} \,\Big|\, \| X_i - X_j \|^d > r_n] \\
\leq & \sum_{1 \leq i < j \leq n} E[(1 - \delta c^d \| X_i - X_j \|^d \lambda(G_0))^{n-2} \,\Big|\, \| X_i - X_j \|^d > r_n] \\
\leq & \sum_{1 \leq i < j \leq n} (1 - \delta c^d r_n \lambda(G_0))^{n-2} \\
= & \sum_{1 \leq i < j \leq n} (1 - \frac{k \log n}{n})^{n-2} \\
= & \binom{n}{2} (1 - \frac{k \log n}{n})^{n-2} \sim \binom{n}{2} \frac{1}{n^k} \longrightarrow_{n \to \infty} 0
\end{aligned}
$$

since $k > 2$. □

5.3 Lower bounds

To develop a lower bound for the longest edge of the MST, we will be comparing M_n with other statistics for which asymptotic information is available. First we need the following:

Lemma 5.3.1 *Let* **P** *have compact support and density f (with respect to Lebesgue measure) such that:*

$$\inf\{f(x),\ x \in supp(\mathbf{P})\} = \delta > 0.$$

Let $X_1, X_2, \ldots, X_n$ iid $\sim \mathbf{P}$ and:

$$B_n := \min_{1 \leq i \leq n} \rho(X_i, \partial supp(\mathbf{P})).$$

Then:

$$B_n = O_p\left(\frac{1}{n}\right) \text{ as } n \to \infty.$$

Proof: We will need a lower bound for the d-dimensional volume of the set of points in supp($\mathbf{P}$) that are close to ∂supp($\mathbf{P}$). This is provided by the following lemma:

Lemma 5.3.2 *For any compact set $K \subset \mathbf{R}^d$ with $\lambda(K) > 0$ and any $t > 0$, let $K_t := \{x \in K : \rho(x, \partial K) < t\}$. Then, there is a $\gamma > 0$ such that $\lambda(K_t) \geq \gamma t$ for all t with $0 < t \leq \gamma$.*

Proof: Let $\bar{B}(x,t) := \{y \in R^d : \| x - y \| \leq t\}$. Then let:

$$L_t := K \setminus K_t = \{x \in \mathbf{R}^d : \bar{B}(x,t) \subset K\}.$$

If $\forall t > 0$, $L_t = \emptyset$ then $\lambda(K_t) = \lambda(K)$ and the lemma holds with $\gamma = \lambda(K)^{1/2}$. Since $L_t \uparrow$ as $t \downarrow$, if the set $\{t > 0 : L_t \neq \emptyset\}$ is not empty, it will be an interval. In that case, we proceed as follows:

Since $L_t + \bar{B}(0,t) \subset K$ we can apply the Brunn-Minkowski inequality ([Dud89], page 167) to get:

$$\lambda(K)^{1/d} \geq \lambda(L_t + \bar{B}(0,t))^{1/d} \geq \lambda(L_t)^{1/d} + \lambda(\bar{B}(0,t))^{1/d}.$$

Let ω_d be the volume of the unit ball in $\mathbf{R}^d$ and let $c_d := \omega_d^{1/d}$. Then $\lambda(\bar{B}(0,t))^{1/d} = c_d t$. So:

$$\begin{aligned} \lambda(K) &\geq \lambda(L_t) + dc_d t \lambda(L_t)^{\frac{d-1}{d}} \text{ (by the binomial theorem)} \\ &\geq \lambda(L_t)[1 + dc_d t \lambda(K)^{-1/d}] \text{ (because } \lambda(L_t) \leq \lambda(K)) \\ \Rightarrow \lambda(L_t) &\leq \frac{\lambda(K)}{1 + dc_d t \lambda(K)^{-1/d}}. \end{aligned}$$

Now, for any $0 < \epsilon \leq 1$:

$$\frac{1}{1+\epsilon} \leq 1 - \frac{\epsilon}{2}$$

(since this is equivalent to $1 \leq 1 + \epsilon - \frac{\epsilon}{2} - \frac{\epsilon^2}{2} \Leftrightarrow 1 \leq 1 + \frac{\epsilon(1-\epsilon)}{2}$ which is true). So:

$$\lambda(L_t) \leq \lambda(K) \left[1 - \frac{dc_d t \lambda(K)^{-1/d}}{2} \right]$$

provided that $dc_d t \lambda(K)^{-1/d} \leq 1 \Leftrightarrow t \leq \frac{\lambda(K)^{1/d}}{dc_d}$. Then:

$$\lambda(K_t) = \lambda(K) - \lambda(L_t) \geq \left(\frac{\lambda(K)^{\frac{d-1}{d}} dc_d}{2} \right) t.$$

We can now let $\gamma := \min\{\frac{\lambda(K)^{1/d}}{dc_d}, \frac{dc_d \lambda(K)^{\frac{d-1}{d}}}{2}, \sup\{t \geq 0 : \ L_t \neq \emptyset\}\}$ and the lemma follows. □

Proof: (of Lemma 5.3.1). Let $M > 0$. If $B_n > M/n$ then the set $K_{M/n}$ of Lemma 5.3.2 does not contain any of the observations $X_1, X_2, \ldots, X_n$. By Lemma 5.3.2 and because supp($\mathbf{P}$) is compact, there exists a $\gamma > 0$ such that $\lambda(K_{M/n}) \geq \gamma M/n$ for all n such that $M/n \leq \gamma$. Therefore:

$$\begin{aligned} \Pr\left(B_n > \frac{M}{n}\right) &\leq \left(1 - \mathbf{P}(K_{M/n})\right)^n \\ &\leq \left(1 - \delta\lambda(K_{M/n})\right)^n \\ &\leq \left(1 - \delta\gamma\frac{M}{n}\right)^n \\ &\leq e^{-\delta\gamma M}, \ \forall n. \end{aligned}$$

Since $e^{-\delta\gamma M} \to 0$ as $M \to \infty$, this proves Lemma 5.3.1. □

Now recall that minimal spanning trees were defined in Chapter 2 (Definition 2.4.2) on arbitrary graphs (not necessarily graphs of points in $\mathbf{R}^d$). In fact, Proposition 2.4.1 proved the equivalence of single-link distances and MSTs on any *clustering*. So the following lemma involves no new concepts:

Lemma 5.3.3 *Let* $\mathbf{P}$ *have compact support and* $x_1, x_2, \ldots, x_n$ *iid* $\sim \mathbf{P}$. *Let* M_n *be the length of the longest edge of the MST on* $\{X_1, X_2, \ldots, X_n\}$ *and* M_n^* *the length of the longest edge of the MST on:*

$$\{\partial supp(\mathbf{P})\} \cup \{\{X_i\}, \ X_i \notin \partial supp(\mathbf{P})\}.$$

Then:

$$\max\{M_n, B_n\} \geq M_n^*.$$

Proof: Let $\mathcal{C} = \{\{X_i\},\ 1 \leq i \leq n\}$ and:

$$\mathcal{C}^* = \{\partial\text{supp}(\mathbf{P})\} \cup \{\{X_i\},\ X_i \notin \partial\text{supp}(\mathbf{P})\}.$$

By the definition of single-link distances with respect to a clustering, we have:

$$\begin{aligned} d_{SL}^{\mathcal{C}^*}(X_i, X_j) &\leq d_{SL}^{\mathcal{C}}(X_i, X_j),\ 1 \leq i, j \leq n \\ &\leq M_n\ (= \max_{i,j} d_{SL}^{\mathcal{C}}(X_i, X_j)). \end{aligned}$$

Also:

$$d_{SL}^{\mathcal{C}^*}(X_i, \partial\text{supp}(\mathbf{P})) \leq \max\{d_{SL}^{\mathcal{C}^*}(X_i, X_j), d_{SL}^{\mathcal{C}^*}(X_j, \partial\text{supp}(\mathbf{P}))\}.$$

Taking X_j to be an observation for which:

$$B_n = \rho(X_j, \partial\text{supp}(\mathbf{P}))$$

we have:

$$\begin{aligned} d_{SL}^{\mathcal{C}^*}(X_i, \partial\text{supp}(\mathbf{P})) &\leq \max\{d_{SL}^{\mathcal{C}^*}(X_i, X_j), B_n\} \\ &\leq \max\{d_{SL}^{\mathcal{C}}(X_i, X_j), B_n\} \\ &\leq \max\{M_n, B_n\}. \end{aligned}$$

So $M_n^* \leq \max\{M_n, B_n\}$. □

We will now compare M_n^* to statistics based on nearest neighbor distances.

Definition 5.3.1

1. *Let* $\mathbf{P}$ *have compact support and* $X_1, X_2, \ldots, X_n$ *iid* $\sim \mathbf{P}$. *Then define:*

$$Z_n := \max_{1 \leq i \leq n} \min\{\min\{\rho(X_i, X_j),\ j \neq i\}, \rho(X_i, \partial supp(\mathbf{P}))\}.$$

2. Let **P** *have a density f with respect to Lebesgue measure and define:*

$$D_n := \max_{1\leq i\leq n} f(X_i)^{1/d}\min\{\min\{\rho(X_i,X_j), j\neq i\}, \rho(X_i, \partial supp(\mathbf{P}))\}.$$

Then we have the following:

Lemma 5.3.4 *Let* **P** *have compact support and $X_1, X_2, \ldots, X_n$ iid $\sim$* **P**. *Then $M_n^* \geq Z_n$.*

Proof: If $Z_n = \rho(X_{i_0}, X_{j_0})$ for $1 \leq i_0, j_0 \leq n$ then:

$$d_{SL}^{\mathcal{C}^*}(X_{i_0}, X_{j_0}) = \rho(X_{i_0}, X_{j_0}) = Z_n$$

because the best path from X_{i_0} to X_{j_0} is simply the edge $\{X_{i_0}, X_{j_0}\}$. So $M_n^* \geq Z_n$. Similarly we can handle the case:

$$Z_n = \rho(X_{i_0}, \partial\text{supp}(\mathbf{P}))$$

for some $1 \leq i_0 \leq n$. □

The following is known about D_n:

Proposition 5.3.1 *Let* **P** *be concentrated on an open bounded subset U of $\mathbf{R}^d$ whose closure is supp*(**P**) *and assume it has a continuous density f satisfying:*

- $\inf\{f(x),\ x \in supp(\mathbf{P})\} > 0$ *and*
- $\sup\{|f(x) - f(y)|\ :\ |x-y| < s\} = o((-\log s)^{-1})$ *as $s \to 0$.*

Let ω_d denote the volume of the unit ball in $\mathbf{R}^d$.
Then, for $X_1, X_2, \ldots, X_n$ iid $\sim$ **P**:

$$\lim_{n\to\infty} \Pr\left(n\omega_d D_n^d - \log n < \xi\right) = \exp(-e^{-\xi}),\ \forall \xi \in \mathbf{R}.$$

Proof: This is the main result of [Hen83]. □

A lower bound for M_n can be derived under the same assumptions on f.

Theorem 5.3.1 *Let* $\mathbf{P}$ *satisfy the following assumptions:*

- $\mathbf{P}$ *is concentrated on an open bounded subset* U *of* $\mathbf{R}^d$ *whose closure is supp*$(\mathbf{P})$ *and*

- $\mathbf{P}$ *has a continuous density* f *for which:* $\sup\{|f(x)-f(y)| \mid |x-y|<s\} = o((-\log s)^{-1})$ *as* $s \to 0$.

Let $X_1, X_2, \ldots, X_n$ *be iid* $\sim \mathbf{P}$, $\Delta := \max\{f(x),\ x \in supp(\mathbf{P})\}$ *and* ω_d *the volume of the unit ball in* $\mathbf{R}^d$. *Then, for every* $k<1$*:*

$$\lim_{n\to\infty} \Pr\left(M_n^d \geq \frac{k}{\Delta\omega_d}\cdot\frac{\log n}{n}\right) = 1.$$

Proof: By Lemma 5.3.3:

$$M_n^* \leq \max\{M_n, B_n\} \Rightarrow M_n^{*d} - B_n^d \leq M_n^d.$$

Therefore:

$$\begin{aligned}\Delta M_n^d &\geq \Delta M_n^{*d} - \Delta B_n^d \\ &\geq \Delta Z_n^d - \Delta B_n^d \text{ (Lemma 5.3.4)} \\ &\geq D_n^d - \Delta B_n^d \text{ (Definition 5.3.1).}\end{aligned}$$

So:

$$\begin{aligned}\Pr\left(M_n^d < \frac{k}{\Delta\omega_d}\cdot\frac{\log n}{n}\right) &= \Pr\left(\Delta\omega_d n M_n^d < k\log n\right) \\ &\leq \Pr\left(n\omega_d D_n^d - \Delta n\omega_d B_n^d < k\log n\right) \\ &= \Pr\left(n\omega_d D_n^d - \log n < \Delta n\omega_d B_n^d - (1-k)\log n\right) \\ &= \Pr\left(\frac{n\omega_d D_n^d - \log n}{(1-k)\log n - \Delta n\omega_d B_n^d} < -1\right)\end{aligned}$$

for large n, since Lemma 5.3.1 guarantees that:

$$(1-k)\log n - \Delta\omega_d n B_n^d \xrightarrow{\mathrm{P}} +\infty,$$

as $n \to \infty$ (and thus is, eventually, positive). From Proposition 5.3.1, we know that:

$$L_n := n\omega_d D_n^d - \log n \xrightarrow{\mathcal{L}} L$$

where L is a random variable for which:

$$\Pr(L < \xi) = \exp(-e^{-\xi}), \ \forall \xi \in \mathbf{R}.$$

Also:

$$k_n := \frac{1}{(1-k)\log n - \Delta\omega_d n B_n^d} \xrightarrow{\mathrm{P}} 0.$$

Therefore:

$$k_n L_n \xrightarrow{\mathcal{L}} 0 \Rightarrow \lim_{n\to\infty} \Pr(k_n L_n < -1) = 0.$$

□

Chapter 6

Clustering under the Density Model

6.1 Chaining and breakdown points

The main drawback of the single-link method is the *chaining effect.* A few observations between clusters can create a *chain*, i.e. a path of small size joining the clusters and thus making the single link distances small (Figure 6.1). (By Definition 2.3.2, the size of a path is the length of the longest edge of the path).

We can describe this effect more precisely using the terminology of robust statistics. Let $\mathcal{P}_C$ denote the family of clustered probability measures on $\mathbf{R}^d$ that have at least two clusters. By Definition 3.3.1, $M(\mathbf{P})$ denotes the length of the longest edge of any minimal spanning tree on $\mathcal{C}(\mathbf{P})$ (the clustering of $\mathbf{P}$). Therefore, $M(\mathbf{P}) > 0$ for any $\mathbf{P} \in \mathcal{P}_C$. By Theorem 3.3.2, we have:

$$M(\mathbf{P}_n) \xrightarrow{\text{a.s.}} M(\mathbf{P})$$

as $n \longrightarrow \infty$.

Recall now the definition of the gross-error breakdown point of a sequence of estimators (see also [Hub81], page 13 or [HRRS86], page 97).

Definition 6.1.1 *Let* $\mathbf{P}$ *be a Borel probability measure on* $\mathbf{R}^d$*. Suppose that* $\theta \in \Theta \subset \mathbf{R}^k$ *is estimated by a sequence of estimators of the form* $T_n(X_1, X_2, \ldots, X_n)$ *where:*

$$T_n : \mathbf{R}^{d \times n} \mapsto \mathbf{R}^k.$$

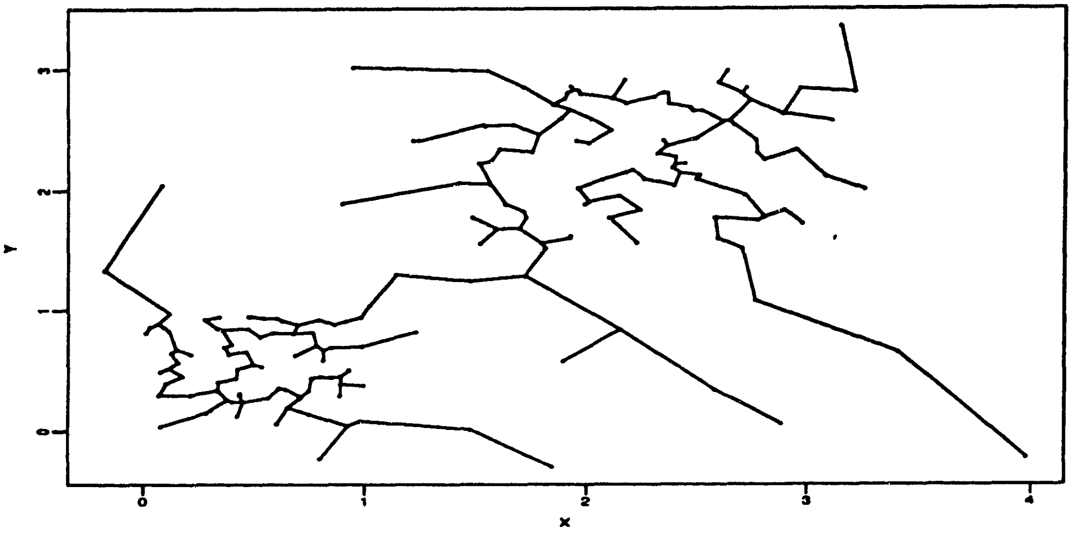


Figure 6.1: Chaining effects

The gross-error breakdown point of T_n *at* $\mathbf{P}$ *is defined to be:*

$$\begin{aligned} b(\mathbf{P},\Theta,T_n) \;:=\; & \sup\{\epsilon \le 1 \,:\, \exists \textit{ a compact set } K_\epsilon \subset \Theta \textit{ such that} \\ & \lim_n \Pr(T_n(X_1,X_2,\ldots,X_n) \in K_\epsilon) = 1\} \end{aligned}$$

where $X_1,X_2,\ldots,X_n$ *iid* $\sim \mathbf{Q}$*,* $\mathbf{Q} = (1-\epsilon)\mathbf{P} + \epsilon\mathbf{R}$ *and* $\mathbf{R}$ *is an arbitrary probability measure.*

When $\mathbf{P} \in \mathcal{P}_C$ then $M(\mathbf{P}) > 0$. Therefore, the natural parameter space Θ for $\theta = M(\mathbf{P})$ is $\Theta = (0,\infty)$. We now have the following:

Theorem 6.1.1 *For every* $\mathbf{P} \in \mathcal{P}_C$*,* $b(\mathbf{P},(0,\infty),M(\mathbf{P}_n)) = 0$.

Proof: For every $0 < \epsilon \le 1$, we can find a probability measure $\mathbf{R}$ such that the measure:

$$\mathbf{Q}_\epsilon = (1-\epsilon)\mathbf{P} + \epsilon\mathbf{R}$$

has compact and connected support. For example, we can choose $\mathbf{R}$ as follows: Since $\mathbf{P}$ is a clustered measure, its support is compact. So, there exists a ball $B_\mathbf{P}$ large enough so that supp($\mathbf{P}$) $\subset B_\mathbf{P}$. Let $\mathbf{R}$ be the uniform distribution on $B_\mathbf{P}$. Then $\mathbf{Q}_\epsilon$ has compact and connected support. If $X_1,X_2,\ldots,X_n$ *iid* $\sim \mathbf{Q}_\epsilon$ and

$$\mathbf{Q}_{\epsilon,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

is the empirical measure, then $\lim_n M(\mathbf{Q}_{\epsilon,n}) = 0$ a.s. by Theorem 3.3.2. Any compact set K_ϵ in Θ must have a positive distance from 0 and so:

$$\Pr(M(Q_{\epsilon,n} \in K_\epsilon) \longrightarrow 0$$

as $n \to \infty$. So, the breakdown point is 0. □

Remark 6.1.1 When dealing with real data, one is more likely to come across a *contaminated* version of a $\mathbf{P} \in \mathcal{P}_C$ than $\mathbf{P}$ itself. Since the statistic $M(\mathbf{P}_n)$ is so easily affected by such contamination we will have to make adjustments. Ideally, one would like to 'strip' the contamination $\epsilon\mathbf{R}_\epsilon$ from the measure $\mathbf{Q}_\epsilon$ and *then* estimate the functional $M(\mathbf{P})$. The idea is similar to the use of *trimmed means* as location estimators (instead of simply using the non-robust sample mean). In the case of location estimators the type of contamination we mostly want to avoid is that from long-tailed distributions. Therefore, in the sample version we choose to remove the extreme observations and compute the mean of the rest. In our case the contamination to be avoided is that of *low density* measures. (The picture of islands (clusters) in a sea of low-density contamination has often been invoked to describe this case). It would then seem that the appropriate action would be to remove low-density observations.

The rest of this chapter is devoted to the implementation of this idea. We will need to adjust the class of measures we will be dealing with and also discuss the use of density estimation.

6.2 Density clustered measures

From the discussion in the previous section, it would appear that we need to handle distributions that exhibit cluster structure, although they may have connected support such as the contaminated measure $\mathbf{Q}_\epsilon$ in the proof of Theorem 6.1.1. To do that, we will have to adopt a broader definition of cluster structure and appropriate functionals to assess this structure.

Definition 6.2.1 *A probability measure* $\mathbf{P}$ *on* $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$ *will be called* $[\delta_1, \delta_2]$*-clustered iff:*

- $\mathbf{P}$ *has a density* f *with respect to Lebesgue measure which is continuous on supp*$(\mathbf{P})$ *and bounded,*

- $0 < \delta_1 < \delta_2 < \sup_{x \in supp(\mathbf{P})} f(x)$,
- $\exists k \in \mathbf{N}$ *such that for every* $\delta_1 \leq \delta \leq \delta_2$*:*
  - $\mathbf{P}([f \geq \delta]) > 0$ *and*
  - *the probability measure* $\mathbf{P}_\delta$ *with density:*

$$f_\delta(x) = \frac{f(x) \cdot 1_{[f \geq \delta]}(x)}{\mathbf{P}([f \geq \delta])},\ x \in \mathbf{R}^d$$

    is a clustered measure (Definition 3.2.1) with the additional properties:
    - $card(\mathcal{C}(\mathbf{P}_\delta)) \leq k$,
    - $\forall C_1 \neq C_2,\ C_3 \neq C_4 \in \mathcal{C}(\mathbf{P}_\delta)$*:*

$$\rho(C_1, C_2) = \rho(C_3, C_4) \Rightarrow \begin{cases} C_1 = C_3 \quad and \quad C_2 = C_4 \\ \qquad\qquad or \\ C_1 = C_4 \quad and \quad C_2 = C_3. \end{cases}$$

Remark 6.2.1 Notice that measures like the standard multivariate normal are not clustered (their support is not compact). However, the standard normal is $[\delta_1, \delta_2]$-clustered for any:

$$0 < \delta_1 < \delta_2 < \frac{1}{(2\pi)^{d/2}}.$$

Since every $\mathbf{P}_\delta$ is a clustered measure with clustering $\mathcal{C}(\mathbf{P}_\delta)$, we can define single-link distances on supp$(\mathbf{P}_\delta)$ (Definition 2.3.3), single-link hierarchies (Definition 2.3.4) and minimal spanning trees (Definition 2.4.2). Unlike Chapters 3-5 where we studied functionals defined on $\mathbf{P}$ such as $M(\mathbf{P})$ and $D(\mathbf{P})$ (Definition 3.3.1) and estimated them by statistics of the form $M(\mathbf{P}_n)$ and $D(\mathbf{P}_n)$, we will now define families of functionals of the form $T(\mathbf{P}_\delta)$, $\delta \in [\delta_1, \delta_2]$ and estimate them by stochastic processes of the form $T_{\delta,n}(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ $iid \sim \mathbf{P}$.

Definition 6.2.2 *Let* $\mathbf{P}$ *be a* $[\delta_1, \delta_2]$*-clustered measure on* $\mathbf{R}^d$ *with density* f*. Let* $\delta_1 \leq \delta \leq \delta_2$*.*

- *Let TREE$(\delta) = (\mathcal{C}(\mathbf{P}_\delta), E_\delta)$ be the unique minimal spanning tree on the clustering $\mathcal{C}(\mathbf{P}_\delta)$. Then, let:*

$$M(\mathbf{P}, \delta) := \max\{\| e \|, e \in E_\delta\}.$$

- *Let $r_1, r_2, r_3 \in \mathbf{N}$. If card$(\mathcal{C}(\mathbf{P}_\delta)) = 1$, let:*

$$T^{r_1,r_2,r_3}(\mathbf{P}, \delta) = 0.$$

  If card$(\mathcal{C}(\mathbf{P}_\delta)) > 1$, let $\mathcal{P}_\delta = (A_\delta, B_\delta)$ be the unique partition of supp$(\mathbf{P}_\delta)$ which we obtain after removing the longest edge of TREE(δ). Then define:

$$T^{r_1,r_2,r_3}(\mathbf{P}, \delta) = \mathbf{P}(A_\delta)^{r_1} \cdot M(\mathbf{P}, \delta)^{r_2} \cdot \mathbf{P}(B_\delta)^{r_3}.$$

Definition 6.2.3 *A $[\delta_1, \delta_2]$-clustered measure $\mathbf{P}$ is going to be called $[\delta_1, \delta_2]$-unimodal iff card$(\mathcal{C}(\mathbf{P}_\delta)) = 1$ for all $\delta \in [\delta_1, \delta_2]$.*

6.3 Estimating $T^{r_1,r_2,r_3}(\mathbf{P}, \delta)$

Let $X_1, X_2, \ldots, X_n$ be *iid* $\sim \mathbf{P}$, where $\mathbf{P}$ is a $[\delta_1, \delta_2]$-clustered measure. We would like to construct estimators of both $M(\mathbf{P}, \delta)$ and $T^{r_1,r_2,r_3}(\mathbf{P}, \delta)$, for $\delta \in [\delta_1, \delta_2]$. Since these functionals involve the unknown density f of $\mathbf{P}$, we will need to use a density estimator f_n of f. There is of course, a wide range of choices as well as an enormous literature on density estimation (see e.g. [Rao83] and [Sil86]).

Combining density estimation with other clustering methods is not new. In [Kit76, SDJ79, Kit79], there is a discussion of a mode seeking method that appears to resemble single-link clustering. Another hybrid clustering method using density estimation is described in [Won82] and [WL83].

In the rest of this chapter, we will make use of one particular class of density estimators, the kernel class. Such estimators were first suggested by [Ros56] for univariate densities and were extended to the multivariate case by [Cac66]. We make no claim, however, that this class performs better in conjunction with single-link than any other. For completeness, we give the following definition.

Definition 6.3.1 *Suppose $K : \mathbf{R}^d \mapsto \mathbf{R}$ is Borel measurable and satisfies:*

1. $\int_{\mathbf{R}^d} K(x)dx = 1$,

2. $\sup\{|K(x)|, x \in \mathbf{R}^d\} < \infty$,

3. $\int_{\mathbf{R}^d} |K(x)|dx < \infty$ *and*

4. $\lim_{\|y\|\to\infty} \| y \|^d K(y) = 0$.

Let h_n *be a sequence of positive numbers. Then, if* $X_1, X_2, \ldots, X_n$ *iid* $\sim f$*, the function* $f_n : \mathbf{R}^d \mapsto \mathbf{R}$ *defined by:*

$$f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K(\frac{x - X_i}{h_n})$$

will be called the density estimator with kernel K *and window* h_n.

For our purposes, we are interested in the uniform consistency of such estimators. The following theorem was proved by Devroye and Wagner in 1976 and a proof is included in [Rao83], pages 185-188. For a different set of conditions, leading to the same conclusion, see [BR78].

Theorem 6.3.1 *Suppose* K *is a density on* $\mathbf{R}^d$ *satisfying:*

1. $\sup\{K(x), x \in \mathbf{R}^d\} < \infty$,

2. $\sup\{\| x \|^d K(x)\} < \infty$ *and*

3. $\sup_x\{|K(x + y) - K(x)|\} \le C \| y \|$, *for* $y \in R^d$, *for some* $C > 0$.

Also, assume that:

- $h_n \to 0$ *as* $n \to \infty$ *and*

- $nh_n^d / \log n \to \infty$ *as* $n \to \infty$.

Then, for every uniformly continuous density f*, such that, for some* $\gamma > 0 :$ $\int_{\mathbf{R}^d} \| x \|^d f(x)dx < \infty$*, we have:*

$$\sup_{x \in R^d} |f_n(x) - f(x)| \to 0$$

a.s. as $n \to \infty$.

Having now a density estimator allows us to construct estimators for $M(\mathbf{P},\delta)$ and $T^{r_1,r_2,r_3}(\mathbf{P},\delta)$.

Definition 6.3.2 *Let* $\mathbf{P}$ *be a Borel measure in* $\mathbf{R}^d$ *with density* f. *Let* $X_1, X_2, \ldots, X_n$ *be iid* $\sim \mathbf{P}$. *Let* f_n *be a density estimator of* f. *For* $\delta \geq 0$ *let:*

- $S_n(f_n,\delta) := \{X_i : f_n(X_i) \geq \delta\}$,
- $TREE_{\delta,n} = (S_n(f_n,\delta), E_{\delta,n})$ *be a minimal spanning tree on* $S_n(f_n,\delta)$,
- $M_n(f_n,\delta) := \max\{\| e \|, e \in E_{\delta,n}\}$ *(set to 0 if* $E_{\delta,n} = \emptyset$*),*
- $\mathcal{P}_{\delta,n} = (A_{\delta,n}, B_{\delta,n})$ *be the partition of* $S_n(f_n,\delta)$ *obtained when we remove the (a.s. unique) longest edge of the tree* $TREE_{\delta,n}$ *and*
- $T_n^{r_1,r_2,r_3}(f_n,\delta) := \mathbf{P}_n(A_{\delta,n})^{r_1} \cdot M_n(f_n,\delta)^{r_2} \cdot \mathbf{P}_n(B_{\delta,n})^{r_3}$.

We turn now to the question of computing the stochastic processes $M_n(f_n,\delta)$ and $T_n^{r_1,r_2,r_3}(f_n,\delta)$ given a sample $X_1, X_2, \ldots, X_n$. Let $\delta_i := f_n(X_i)$ for $1 \leq i \leq n$ and let $\delta_{(1)} \leq \delta_{(2)} \leq \ldots \leq \delta_{(n)}$ be the order statistics of $\delta_1, \delta_2, \ldots, \delta_n$. Then:

$$M_n(f_n,\delta) = \begin{cases} 0 & \text{for } \delta > \delta_{(n-1)} \\ M_n(f_n,\delta_{(i)}) & \text{for } \delta_{(i)} \geq \delta > \delta_{(i-1)},\ 2 \leq i < n \\ M_n(f_n,\delta_{(n)}) & \text{for } \delta \leq \delta_{(1)} \end{cases}$$

and

$$T_n^{r_1,r_2,r_3}(f_n,\delta) = \begin{cases} 0 & \text{for } \delta > \delta_{(n-1)} \\ T_n^{r_1,r_2,r_3}(f_n,\delta_{(i)}) & \text{for } \delta_{(i)} \geq \delta > \delta_{(i-1)},\ 2 \leq i < n \\ T_n^{r_1,r_2,r_3}(f_n,\delta_{(n)}) & \text{for } \delta \leq \delta_{(1)}. \end{cases}$$

So, we only need to compute $M_n(f_n,\delta_{(i)})$ and $T_n^{r_1,r_2,r_3}(f_n,\delta_{(i)})$ for $i = 1, 2, \ldots, n$. As a first step, we reduce the computation of these statistics to the computation of $E_{\delta_{(i)},n}$, the set of edges of the MST on $S_n(f_n,\delta_{(i)})$. First we need the following:

Lemma 6.3.1 *There is a constant* N_d, *depending only on* d, *such that for any MST in* $\mathbf{R}^d$ *and for any vertex of that MST, the number of edges adjacent to the vertex is bounded by* N_d. *Furthermore,* $N_d = O(2.77^d)$ *as* $d \to \infty$.

Proof: The existence of such a constant is a special case of Lemma 2.4 in [Ste88]. It is based on the fact that, by an argument similar to that in Lemma 5.2.1, no two edges of the MST can form an angle of less than 60°. Therefore, the number of edges adjacent to any vertex must be bounded by the maximum possible number N_d of edges adjacent to a point in $\mathbf{R}^d$ without any of the edges forming an angle smaller than 60°. Clearly, N_d only depends on d.

Furthermore, N_d is bounded by the maximum number of 30° caps than can be packed on the surface of the unit sphere S^{d-1}. This, in turn, is bounded by the ratio of the surface area A_d^{180} of S^{d-1} to the surface area A_d^{30} of a 30° cap. Let $f(x) = \sqrt{1-x^2}$. Then:

$$\begin{aligned} A_d^{30} &= \int_{\sqrt{3}/2}^{1} f(x)^{d-2} \cdot A_{d-1}^{180} \cdot \sqrt{1+[f'(x)]^2}dx \\ &= A_{d-1}^{180} \int_{\sqrt{3}/2}^{1} f(x)^{d-3}dx \end{aligned}$$

and similarly:

$$A_d^{180} = 2A_{d-1}^{180} \int_0^1 f(x)^{d-3}dx.$$

Since $f(x) \le f(0)$ for $x \in [0,1]$:

$$A_d^{180} \le 2A_{d-1}^{180} f(0)^{d-3}(1-0) = 2A_{d-1}^{180}.$$

Note that the function f is decreasing in $[\sqrt{3}/2, 1]$. So, for $0 < \epsilon < 1-\sqrt{3}/2$:

$$\begin{aligned} A_d^{30} &> A_{d-1}^{180} \int_{\sqrt{3}/2}^{1-\epsilon} f(x)^{d-3}dx \\ &\ge A_{d-1}^{180} \left(1-\epsilon-\frac{\sqrt{3}}{2}\right) [f(1-\epsilon)]^{d-3} \\ &= A_{d-1}^{180} \left(1-\epsilon-\frac{\sqrt{3}}{2}\right) (2\epsilon-\epsilon^2)^{\frac{d-3}{2}} \\ &> A_{d-1}^{180} \left(1-\epsilon-\frac{\sqrt{3}}{2}\right) \epsilon^{\frac{d-3}{2}}. \end{aligned}$$

So:

$$N_d < \frac{2}{\left(1-\epsilon-\frac{\sqrt{3}}{2}\right)\epsilon^{\frac{d-3}{2}}} = \frac{2\epsilon^{3/2}}{\left(1-\epsilon-\frac{\sqrt{3}}{2}\right)}(\epsilon^{-1/2})^d = O((\epsilon^{-1/2})^d)$$

as $d \to \infty$. So, letting $\epsilon = 0.13$, we get $N_d = O(2.77^d)$. □

An MST (S_n, E_n) on $S_n = \{X_1, X_2, \ldots, X_n\}$ can be represented in a variety of ways. We choose the following form: Define an array t of dimension nN_d. Fill the block $t((i-1)N_d + 1)$ to $t(iN_d)$ by the integers j such that $(i,j) \in E_n$ in ascending order. Any remaining values are filled with zeros. By Lemma 6.3.1, an array of dimension nN_d will be sufficient. As is generally done in the combinatorial optimization literature (see e.g. [PS82], page 161), we assume that all integers to be treated by the algorithm can be transcribed in unit time. Under this assumption, we have:

Lemma 6.3.2 *Given an MST $TREE_n = (S_n, E_n)$ on $S_n = \{X_1, X_2, \ldots, X_n\}$ with the above described representation and any edge $e \in E_n$, we can obtain the two trees $TREE_n^A(e) = (V_n^A(e), E_n^A(e))$ and $TREE_n^B(e) = (V_n^B(e), E_n^B(e))$ resulting from removing the edge e in $O(n)$ time and using $O(2.77^d n)$ space as $n, d \to \infty$.*

Proof: If $e = (X_i, X_j)$, we search within the blocks $t((i-1)N_d + 1)$ to $t(iN_d)$ and $t((j-1)N_d + 1)$ to $t(jN_d)$ to locate vertices adjacent to i and j. Then we search within the blocks corresponding to these vertices and continue in this fashion until we have exhausted all the vertices in S_n. Every entry in every block corresponds to an edge of the MST. Conversely, every edge appears in exactly two blocks (the ones corresponding to its endpoints). There are exactly $n-1$ edges in the MST and, therefore, there are exactly $2(n-1)$ non-zero entries in the array t. So, it will only take $O(n)$ time to complete the search and transcription (regardless of how large d is).

The space needed to store the array t is $O(nN_d)$. (We know that only $2(n-1)$ entries will be non-zero but we don't know where these entries are going to be.) By Lemma 6.3.1, this makes the space needed $O(2.77^d n)$. □

Remark 6.3.1 It is now obvious from Definition 6.3.2 that, once we have the MST on $S_n(f_n, \delta_{(i)})$, we will need $O(i)$ time to find $M_n(f_n, \delta_{(i)})$ and (because of Lemma 6.3.2) $O(i)$ time to find the sets $A_{\delta_{(i)},n}$ and $B_{\delta_{(i)},n}$ and, therefore, compute $T_n^{r_1,r_2,r_3}(f_n, \delta_{(i)})$.

We now need an algorithm that computes the MST on the set $S_n(f_n, \delta_{(i)})$, for $1 \leq i \leq n$. We will present an algorithm that introduces a new vertex w to an MST TREE$_n = (V_n, E_n)$ of n vertices. Then, the n MSTs on $S_n(f_n, \delta_{(i)})$ can be computed by successive use of this algorithm. The existence of a density f assures us that every MST that we will come across will have a unique longest edge. (When working in floating-point arithmetic, this will happen with probability very close to 1). We also agree that an MST with only one vertex has longest edge equal to 0.

ALGORITHM

Input:

1. The tree $\mathrm{TREE}_n = (V_n, E_n)$.
2. The vertex $w \notin V_n$.

Output:

The tree $\mathrm{TREE}_{n+1} = (V_n \cup \{w\}, E_{n+1})$.

```
1. LET $\rho_w := \rho(w, u_w) := \min_{u \in V_n} \rho(w, u)$.
   LET $e_n$ be the longest edge of $(V_n, E_n)$.

2.   • IF $\rho_w \geq \| e_n \|$ THEN:
         - add $(w, u_w)$ to $E_n$ and $w$ to $V_n$
         - STOP
     • ELSE compute the trees:

           $\mathrm{TREE}_n^A(e_n) = (V_n^A(e_n), E_n^A(e_n))$
           $\mathrm{TREE}_n^B(e_n) = (V_n^B(e_n), E_n^B(e_n))$

       resulting from the removal of $e_n$.
       LET $\rho_w^A := \rho(w, u_w^A) := \min_{u \in V_n^A} \rho(w, u)$.
       LET $\rho_w^B := \rho(w, u_w^B) := \min_{u \in V_n^B} \rho(w, u)$.
       IF $\rho_w^A > \rho_w^B$ SWAP:

           $\mathrm{TREE}_n^A(e_n) \leftrightarrow \mathrm{TREE}_n^B(e_n)$
           $\rho_w^A \leftrightarrow \rho_w^B$

       IF $\rho_w^A < \| e_n \| \leq \rho_w^B$ THEN
```

```
  - include $E_n^B$ and $e_n$ in the new MST,
  - to complete the MST, run the algorithm on $\mathrm{TREE}_n^A(e_n)$ and
    $w$.
  - STOP
IF $\rho_w^A \leq \rho_w^B <\| e_n \|$ THEN
  - include $(w, u_w^A)$ and $(w, u_w^B)$ in the new MST,
  - to complete the MST, run the algorithm on:
    (a) $\mathrm{TREE}_n^A(e_n)$ and $w$,
    (b) $\mathrm{TREE}_n^B(e_n)$ and $w$.
  - STOP
```

Remark 6.3.2 The algorithm terminates, since each recursive call on it reduces the number of the remaining edges by at least one.

Remark 6.3.3 Since the computation of $\min_{u \in V_n} \rho(w, u)$ alone will take $O(n)$ time, the algorithm needs at least that much time. In the worst case, however, it will need $O(n^2)$ as in the univariate example shown in Figure 6.2.

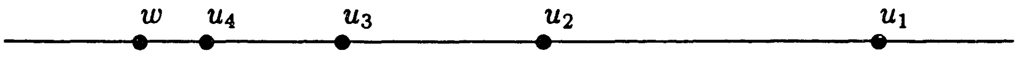


Figure 6.2: An example of worst case performance.

Clearly, to add the new vertex w to the MST on $\{u_1, u_2, u_3, u_4\}$, we will need to recursively call the algorithm on $\{u_4, u_3, u_2\}$, then on $\{u_4, u_3\}$ and finally on $\{u_4\}$. Each recursion will compute trees TREE_n^A and TREE_n^B and this takes $O(n)$ time (Lemma 6.3.2). Therefore, we will need a total of $O(n^2)$ time.

This means that in the worst case, we will need $O(n^3)$ time to build the MSTs on the sets $S_n(f_n, \delta_{(i)})$ for $1 \leq i \leq n$. In practice, however, we expect to do better than that. In our use of the algorithm, the new vertex w must satisfy:

$$f_n(w) \leq \min\{f_n(u_1), f_n(u_2), f_n(u_3), f_n(u_4)\}.$$

Therefore, cases such as the one in Figure 6.2 have a small probability. In fact, when dealing with unimodal distributions such as the multivariate normal, new observations will be added to the MST beginning with the mode and moving towards the tails. So, most of the time we will have $\rho_w \geq \| e_n \|$ and recursive calls will be avoided.

6.4 Simulation results

In this section we will attempt to assess the behavior of the stochastic process $T_n^{r_1,r_2,r_3}(f_n, \delta)$ as an estimator of $T^{r_1,r_2,r_3}(\mathbf{P}, \delta)$ by a series of simulations. The density estimator f_n will have the (d-dimensional) standard normal density as kernel and will have window $h_n = O(n^{-\frac{1}{d+4}})$. This window size is known to minimize the mean integrated square error of a kernel density estimator (see also [Sil86]). Our main goal is to compare the ability of the above mentioned stochastic process (and statistics based on them) to distinguish between unimodal and multimodal densities. The statistics used will be:

1. $\mathrm{SUM}_n := \mathbf{P}_n(A_{\delta_n^{max},n}) + \mathbf{P}_n(B_{\delta_n^{max},n})$
2. $\mathrm{MIN}_n := \min\{\mathbf{P}_n(A_{\delta_n^{max},n}), \mathbf{P}_n(B_{\delta_n^{max},n})\}$

where:

- $\delta_n^{max} := \min\{\delta_i \,:\, T_n^{r_1,r_2,r_3}(f_n, \delta_i) = T_{max,n}^{r_1,r_2,r_3}\}$,
- $\delta_i := f_n(X_i)$, for $1 \leq i \leq n$ and
- $T_{max,n}^{r_1,r_2,r_3} := \max_{1\leq i\leq n}\{T_n^{r_1,r_2,r_3}(f_n, \delta_i)\}$.

We begin by simulating unimodal distributions. As long as such distributions satisfy $\Pr(f = \delta) = 0$ for every δ, we expect that the maximum of $T_n^{r_1,r_2,r_3}(f_n, \delta_i)$ will be achieved close to $\delta_{(1)} = \min_{1\leq i\leq n} \delta_i$. Therefore we expect that:

- $\mathrm{SUM}_n \approx 1$, since $\mathbf{P}_n(f_n \geq \delta_{(1)}) = 1$ and

- $\text{MIN}_n \approx 0$, since all the mass is concentrated in either $A_{\delta_n^{max},n}$ or $B_{\delta_n^{max},n}$.

Our choices of r_1, r_2, r_3 are:

1. $r_1 = r_2 = r_3 = 1$, i.e. $\mathbf{P}_n(A_{\delta,n}) \cdot M(f_n, \delta) \cdot \mathbf{P}_n(B_{\delta,n})$,
2. $r_1 = r_3 = 1$, $r_2 = d$, i.e. $\mathbf{P}_n(A_{\delta,n}) \cdot M(f_n, \delta)^d \cdot \mathbf{P}_n(B_{\delta,n})$ and
3. $r_1 = r_3 = 0$, $r_2 = 1$, i.e. $M(f_n, \delta)$.

We also choose the following distributions: For dimensions $d = 2, 3, 4$, we use:

1. The standard normal (N),
2. the standard normal truncated by 5% in radius (N5) and
3. the standard normal truncated by 10% in radius (N10).

For each distribution in each dimension and for each choice of r_1, r_2, r_3 we use 100 samples of sample size $n = 100$. For each sample i, $1 \leq i \leq 100$, we compute the statistics SUM_i and MIN_i and report the intervals:

- $(\text{SUM}_{(5)}, \text{SUM}_{(96)})$ and
- $(\text{MIN}_{(5)}, \text{MIN}_{(96)})$.

The results are summarized in Table 6.1.

In Figures 6.3,6.4,6.5 at the end of the chapter, you can examine the results of these simulations in the form of boxplots for the empirical distribution of the statistics SUM_i and MIN_i in each of the cases mentioned in Table 6.1. The boxplots are drawn using the default parameters of the S-function *boxplot* ([BCW88], page 402). So, the boxes extend to the upper and lower quartile (with the median highlighted by a horizontal line within the box), while the whiskers are drawn from each quartile to the farthest observation that lies within 1.5 interquartile distances from that quartile on the side away from the median. Observations still farther out are shown individually.

Before moving to multimodal distributions, it would be interesting to examine the behavior of the uniform distribution on the unit cube. Since

d	P	(1,1,1)		(1,d,1)		(0,1,0)	
		SUM$_n$	MIN$_n$	SUM$_n$	MIN$_n$	SUM$_n$	MIN$_n$
2	N	(0.54,1.00)	(0.02,0.19)	(0.79,1.00)	(0.01,0.10)	(0.93,1.00)	(0.01,0.01)
	N5	(0.55,1.00)	(0.02,0.23)	(0.53,1.00)	(0.01,0.20)	(0.06,1.00)	(0.01,0.01)
	N10	(0.58,1.00)	(0.02,0.20)	(0.65,1.00)	(0.02,0.20)	(0.45,1.00)	(0.01,0.01)
3	N	(0.52,1.00)	(0.01,0.16)	(0.83,1.00)	(0.01,0.04)	(0.67,1.00)	(0.01,0.01)
	N5	(0.46,1.00)	(0.02,0.15)	(0.74,1.00)	(0.01,0.06)	(0.05,1.00)	(0.01,0.01)
	N10	(0.40,1.00)	(0.02,0.17)	(0.63,1.00)	(0.01,0.09)	(0.07,1.00)	(0.01,0.01)
4	N	(0.43,1.00)	(0.02,0.09)	(0.87,1.00)	(0.01,0.03)	(0.75,1.00)	(0.01,0.01)
	N5	(0.42,0.99)	(0.02,0.13)	(0.79,1.00)	(0.01,0.05)	(0.03,1.00)	(0.01,0.01)
	N10	(0.51,1.00)	(0.02,0.10)	(0.65,1.00)	(0.01,0.05)	(0.02,1.00)	(0.01,0.01)

Table 6.1: Intervals for SUM$_n$ and MIN$_n$: Unimodal distributions.

the uniform is a unimodal distribution but can also be approximated very well by multimodal distributions, it appears to lie near the boundary between unimodal and multimodal distributions. Table 6.2 summarizes the simulation results for dimensions 2, 3 and 4. The result can also be viewed in the form of boxplots in Figure 6.6.

d	(1,1,1)		(1,d,1)		(0,1,0)	
	SUM$_n$	MIN$_n$	SUM$_n$	MIN$_n$	SUM$_n$	MIN$_n$
2	(0.43,1.00)	(0.03,0.30)	(0.46,1.00)	(0.02,0.30)	(0.03,0.99)	(0.01,0.01)
3	(0.42,1.00)	(0.01,0.20)	(0.45,1.00)	(0.01,0.20)	(0.02,1.00)	(0.01,0.01)
4	(0.36,1.00)	(0.01,0.14)	(0.46,1.00)	(0.01,0.12)	(0.02,1.00)	(0.01,0.01)

Table 6.2: Intervals for SUM$_n$ and MIN$_n$: Uniform distributions.

We now proceed to simulate bimodal distributions. As test distributions we will use mixtures of two d-dimensional normals. The simulated samples will be drawn from the distribution with density:

$$\frac{1}{2} f_{d,0} + \frac{1}{2} f_{d,m} \tag{6.1}$$

where:

- $f_{d,0}$ is the standard d-dimensional normal density and
- $f_{d,m}$ is the d-dimensional normal density with mean $(0, 0, \ldots, 0, m)$ and covariance matrix the identity matrix I_d.

Remark 6.4.1 The distribution with density given by (6.1) is bimodal iff $m > 2$.

Therefore, we will use $d = 2, 3, 4$ and $m = 3, 4, 5$.

As with the unimodal distributions of Table 6.1, we use 100 samples of sample size $n = 100$ for each choice of m and each choice of r_1, r_2, r_3. In each case we report the intervals:

- $(\mathrm{SUM}_{(5)}, \mathrm{SUM}_{(96)})$ and
- $(\mathrm{MIN}_{(5)}, \mathrm{MIN}_{(96)})$.

The results are summarized in Table 6.3.

d	m	$(1,1,1)$		$(1,d,1)$		$(0,1,0)$	
		SUM_n	MIN_n	SUM_n	MIN_n	SUM_n	MIN_n
2	3	(0.35,0.92)	(0.05,0.37)	(0.26,0.99)	(0.02,0.33)	(0.02,1.00)	(0.01,0.01)
	4	(0.53,0.98)	(0.22,0.47)	(0.39,0.86)	(0.16,0.42)	(0.03,0.21)	(0.01,0.01)
	5	(0.71,1.00)	(0.33,0.50)	(0.54,0.95)	(0.24,0.47)	(0.03,0.18)	(0.01,0.01)
3	3	(0.34,0.99)	(0.02,0.38)	(0.19,1.00)	(0.01,0.34)	(0.02,1.00)	(0.01,0.01)
	4	(0.40,0.96)	(0.16,0.46)	(0.19,0.90)	(0.05,0.40)	(0.02,0.22)	(0.01,0.01)
	5	(0.66,1.00)	(0.30,0.50)	(0.40,0.95)	(0.15,0.47)	(0.02,0.16)	(0.01,0.01)
4	3	(0.36,0.99)	(0.02,0.30)	(0.14,1.00)	(0.01,0.25)	(0.02,1.00)	(0.01,0.01)
	4	(0.36,0.97)	(0.05,0.41)	(0.16,1.00)	(0.01,0.38)	(0.02,0.17)	(0.01,0.01)
	5	(0.57,0.99)	(0.26,0.49)	(0.19,0.96)	(0.07,0.48)	(0.02,0.12)	(0.01,0.01)

Table 6.3: Intervals for SUM_n and MIN_n: Bimodal distributions.

Again, the same results can be examined in the following pages in the form of boxplots for the empirical distributions of the statistics SUM_i and MIN_i in each of the cases mentioned in Table 6.3.

We can now summarize the conclusions of these simulations in the following:

- The SUM_n statistic is only helpful in distinguishing between the unimodal and the bimodal distributions of the simulations when used with the process $T_n^{0,1,0}(f_n, \delta)$. Then, the boxplots suggest that SUM_n is close to 1 most of the time in the unimodal cases (except for the uniform case) and close to 0 most of the time in the multimodal cases. Notice also that, as the dimension increases, it becomes more difficult to detect bimodality unless m is large. (Compare e.g. $d = 2$, $m = 3$ with $d = 3$ or 4, $m = 3$.)

There are, however, certain problems with its use that make it unreliable:

1. It completely fails to recognize the unimodality of distributions close to the uniform. In all fairness, this appears to be a problem with the uniform distribution itself, rather than with SUM_n, but the effect on SUM_n is more evident.
2. Furthermore, it seems to do reliably well with unimodal distributions only when these have tails (e.g. the normal). In such a case, $\mathrm{SUM}_n \geq 0.67$ at least 95% of the time (Table 6.1). (In fact, as seen in Figures 6.3,6.4,6.5, $\mathrm{SUM}_n = 1$ at least 75% of the time.) When the distribution has compact support (e.g. N5, N10), $\mathrm{SUM}_{(5)}$ can be as low as 0.02. The fact that a unimodal density can easily be mistaken for a multimodal has been pointed out by Donoho ([Don88]), at least in the univariate case, and it becomes here the main concern in using SUM_n and $T_n^{0,1,0}(f_n, \delta)$.
3. Finally, the last remark leads us to another observation. Although this case has not come up in the simulations, the importance that SUM_n and $T_n^{0,1,0}(f_n, \delta)$ attributes to the tails of the distribution also indicates a sensitivity to outliers. Even significant cluster structure can be ignored when an outlier is present. Instead of $T_n^{0,1,0}(f_n, \delta)$ taking its maximum value at a partition corresponding to the true clusters, it would favor a partition of the sample into the outlier and the rest of the sample. Then, of course, $\mathrm{SUM}_n \approx 1$ which would indicate unimodality.

- When dealing with the MIN_n statistic, we expect $\mathrm{MIN}_n \approx 0$ for a unimodal distribution and MIN_n to be bounded away from 0 for a multimodal distribution. In this sense, the process $T_n^{1,1,1}(f_n, \delta)$ appears to discriminate better than $T_n^{1,d,1}(f_n, \delta)$ between the unimodal and bimodal cases considered in the simulations, especially in higher dimensions.

- Overall, the simulations seem to suggest the use of the $T_n^{1,1,1}(f_n, \delta)$ process (although $T_n^{1,d,1}(f_n, \delta)$ is a viable alternative) and the computation of the MIN_n statistic to measure the degree of multimodality.

  How this information is going to be used may vary in each particular case. A classical approach would be to set up a test where the null hypothesis H_0 would be that of unimodality, a hypothesis rejected when MIN_n is large. Judging from Table 6.1 and letting $\alpha = 0.1$, a value of 0.2 or higher for MIN_n would indicate non-unimodality. This, however, would result in poor power against the bimodal alternatives of Table 6.3 (especially when $m = 3$).

  Notice that finding the maximum of $T_n^{1,1,1}(f_n, \delta)$ also provides a clustering of at least part of the sample and produces a new clustering method. Although based on single-link clustering, this method should be less sensitive to chaining. Therefore, a more realistic decision process might have two stages:

  1. Decide whether there is *some* evidence of multimodality (based on the value of MIN_n), enough to justify the computation of the corresponding partition.
  2. After the partition is computed, decide whether it is an appropriate one, based on data-dependent considerations.

Since the final decision on whether cluster structure is indeed present is defered to the second stage, we might be willing to allow for a large *size* in the first stage (e.g. as large as 0.5) in order to allow for the inspection of the partition. This *size* would then correspond to a value of MIN_n much lower than 0.2. A value of 0.05 will allow us to proceed to the second stage in 95% of the time in all the bimodal cases of Table 6.3 except for the cases $m = 3$, $d = 3$ or 4. Notice, however, that the case $m = 3$ is close enough to being unimodal that high power against it is difficult to achieve in high dimensions with a sample size of 100.

We will see this decision process in practice in the next chapter where we are going to deal with real data sets.

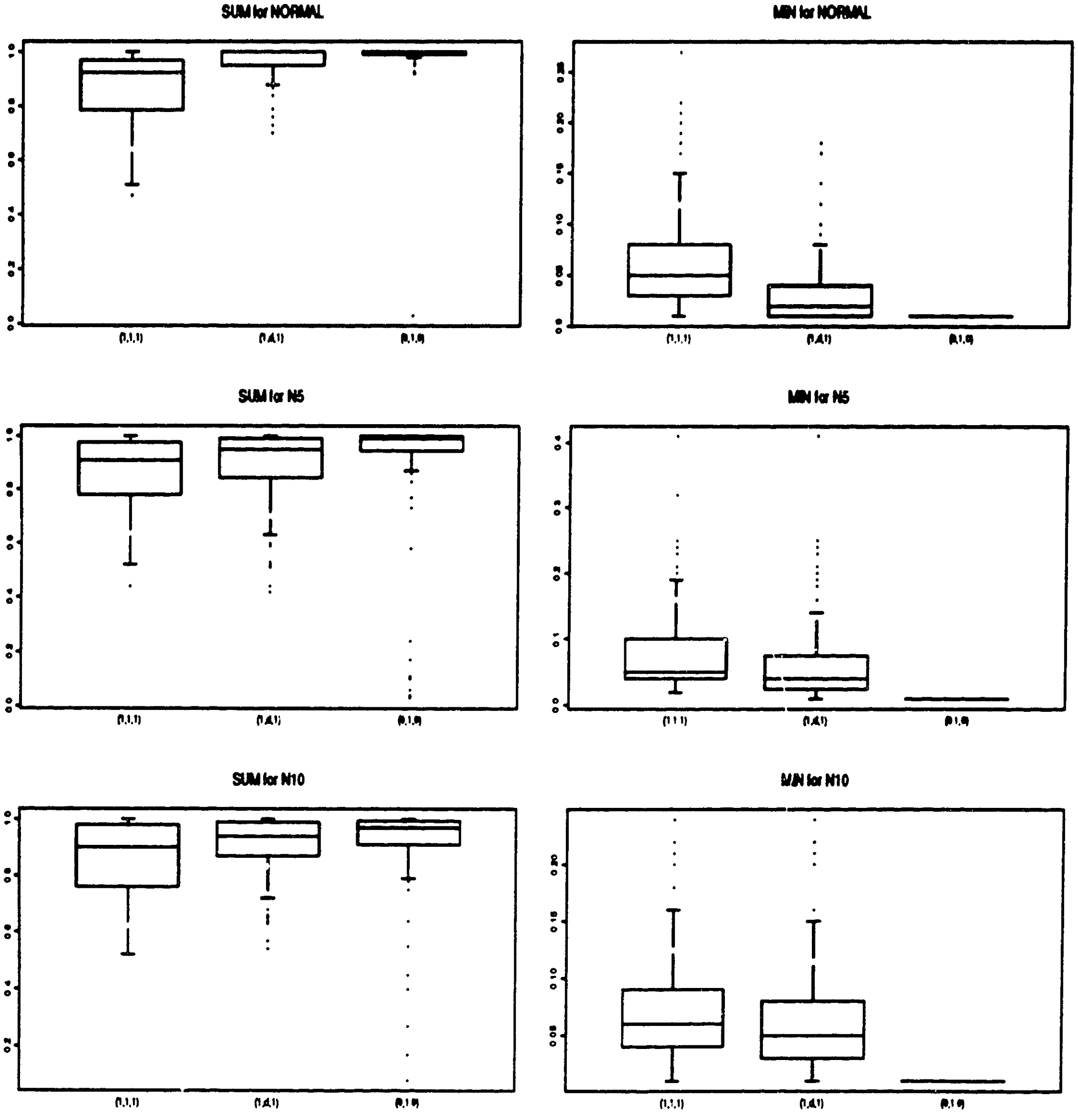


Figure 6.3: 2-dimensional unimodal distributions

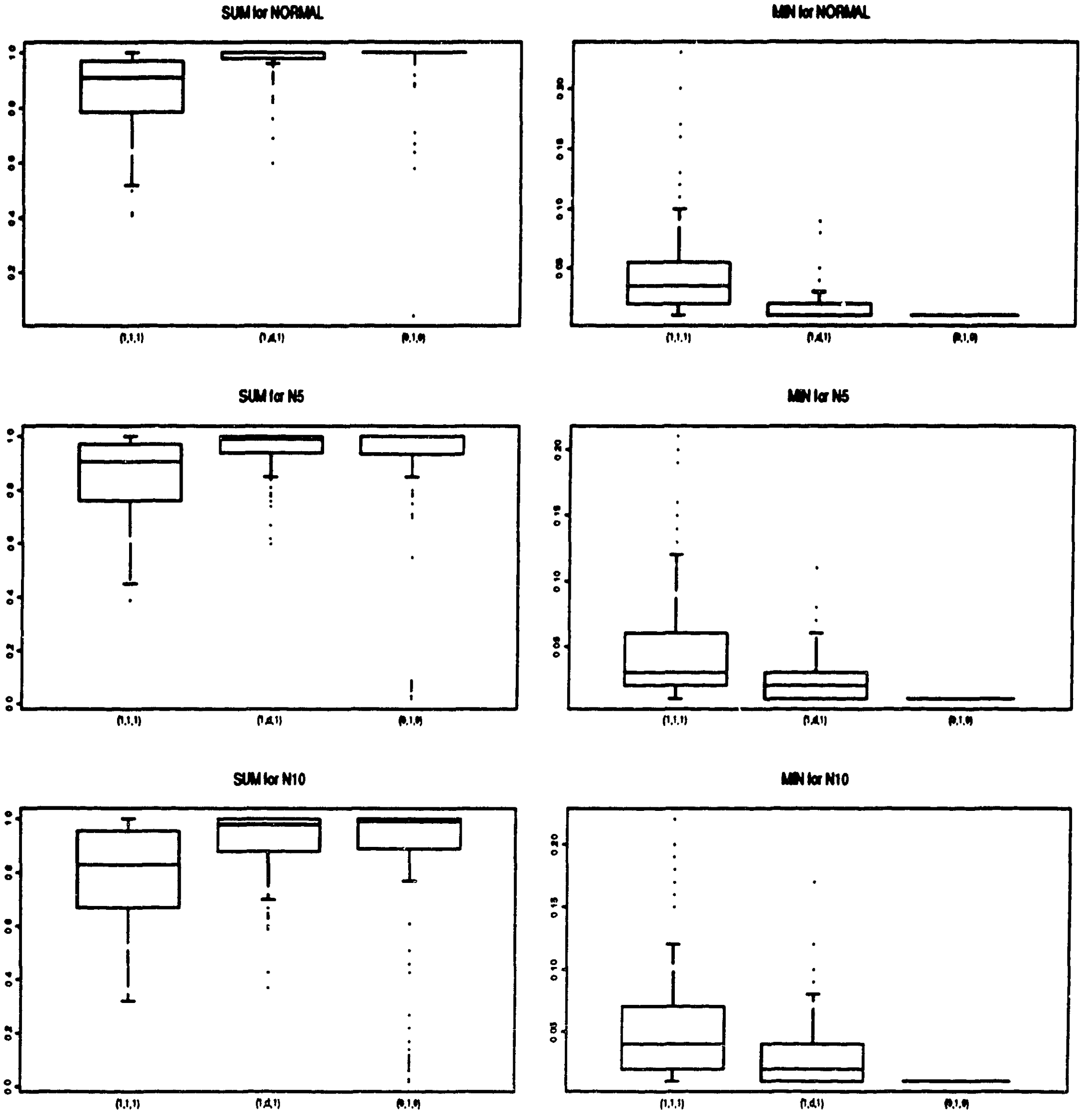


Figure 6.4: 3-dimensional unimodal distributions

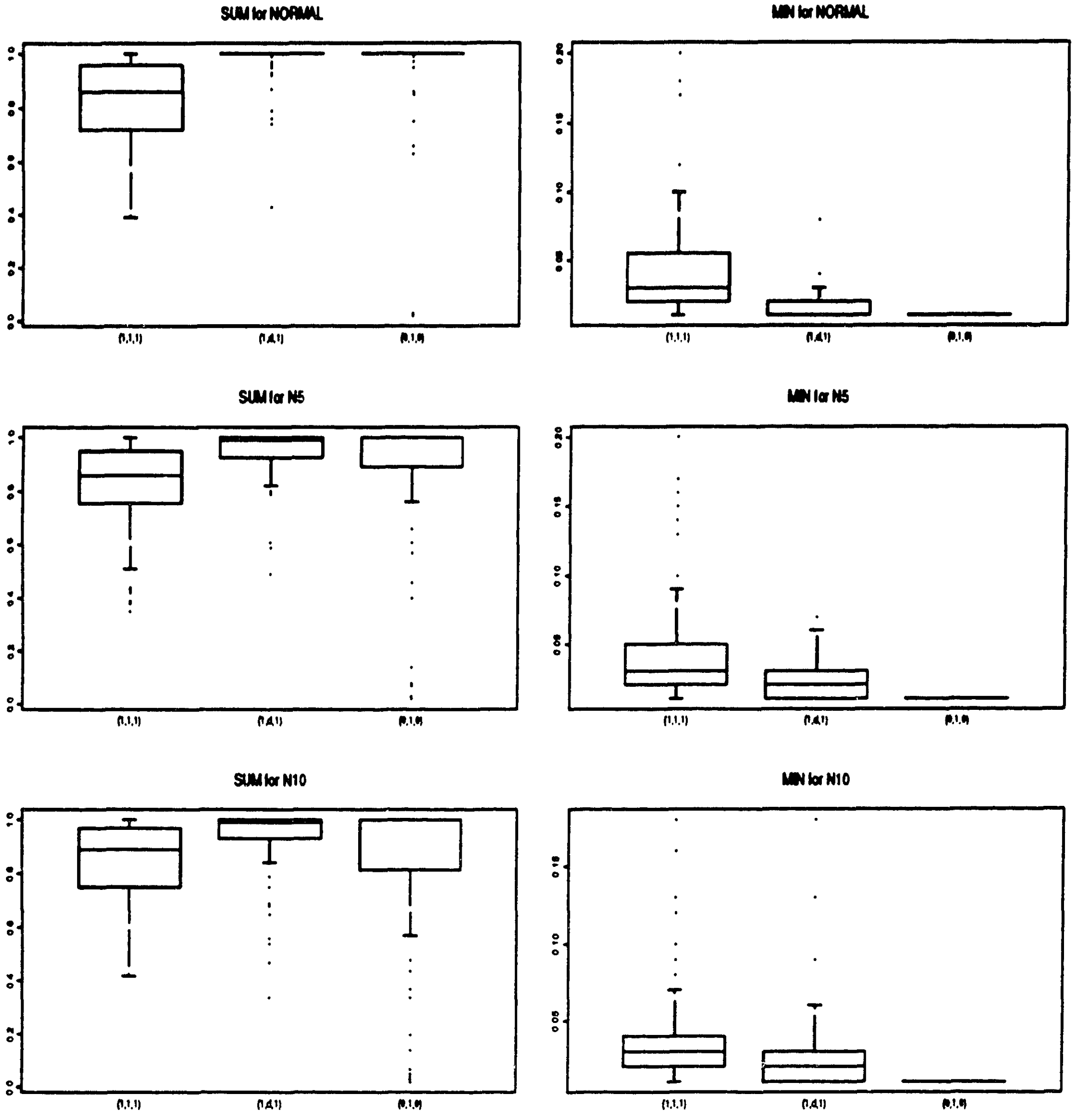


Figure 6.5: 4-dimensional unimodal distributions

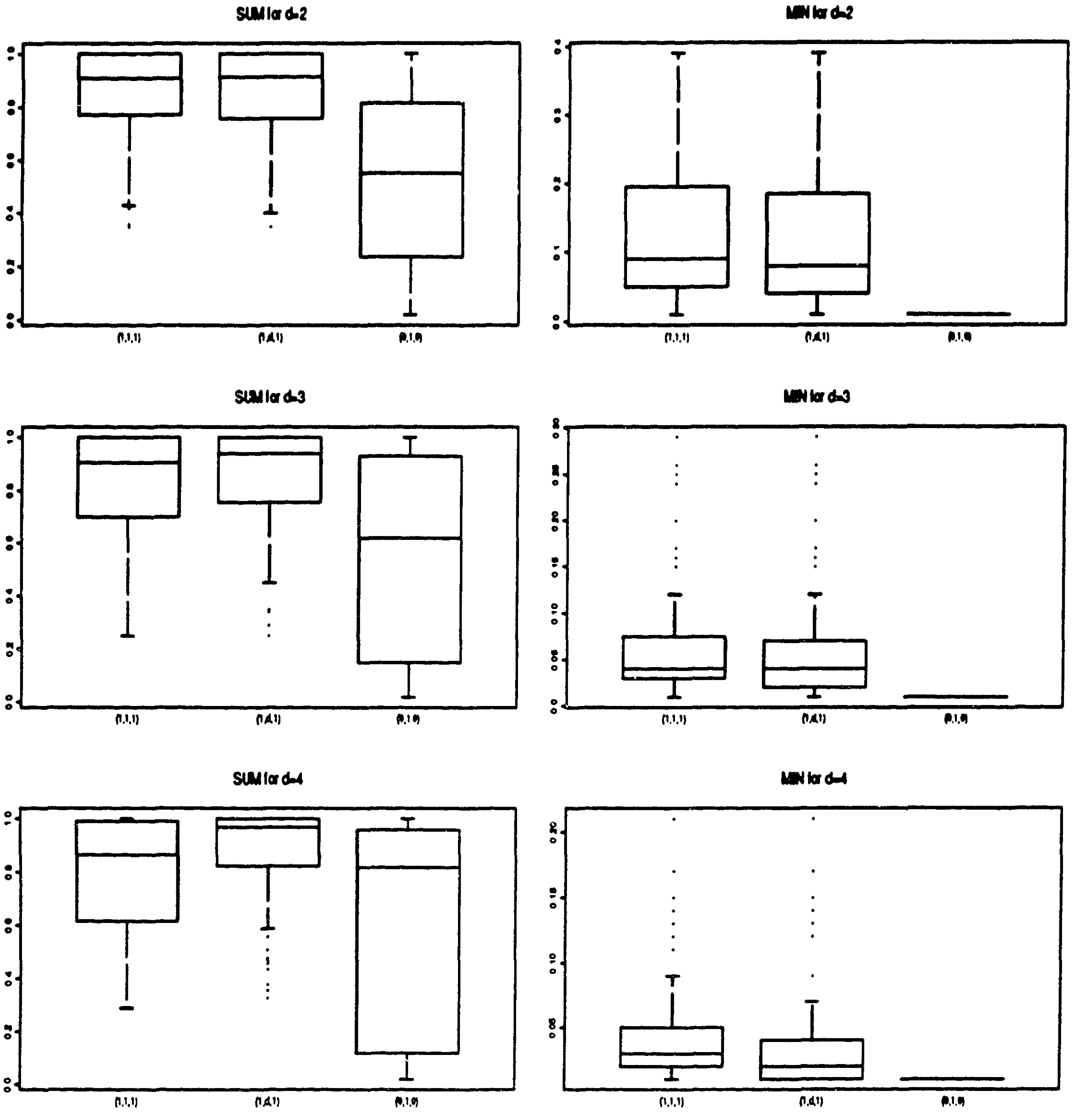


Figure 6.6: Uniform distributions on the unit cube

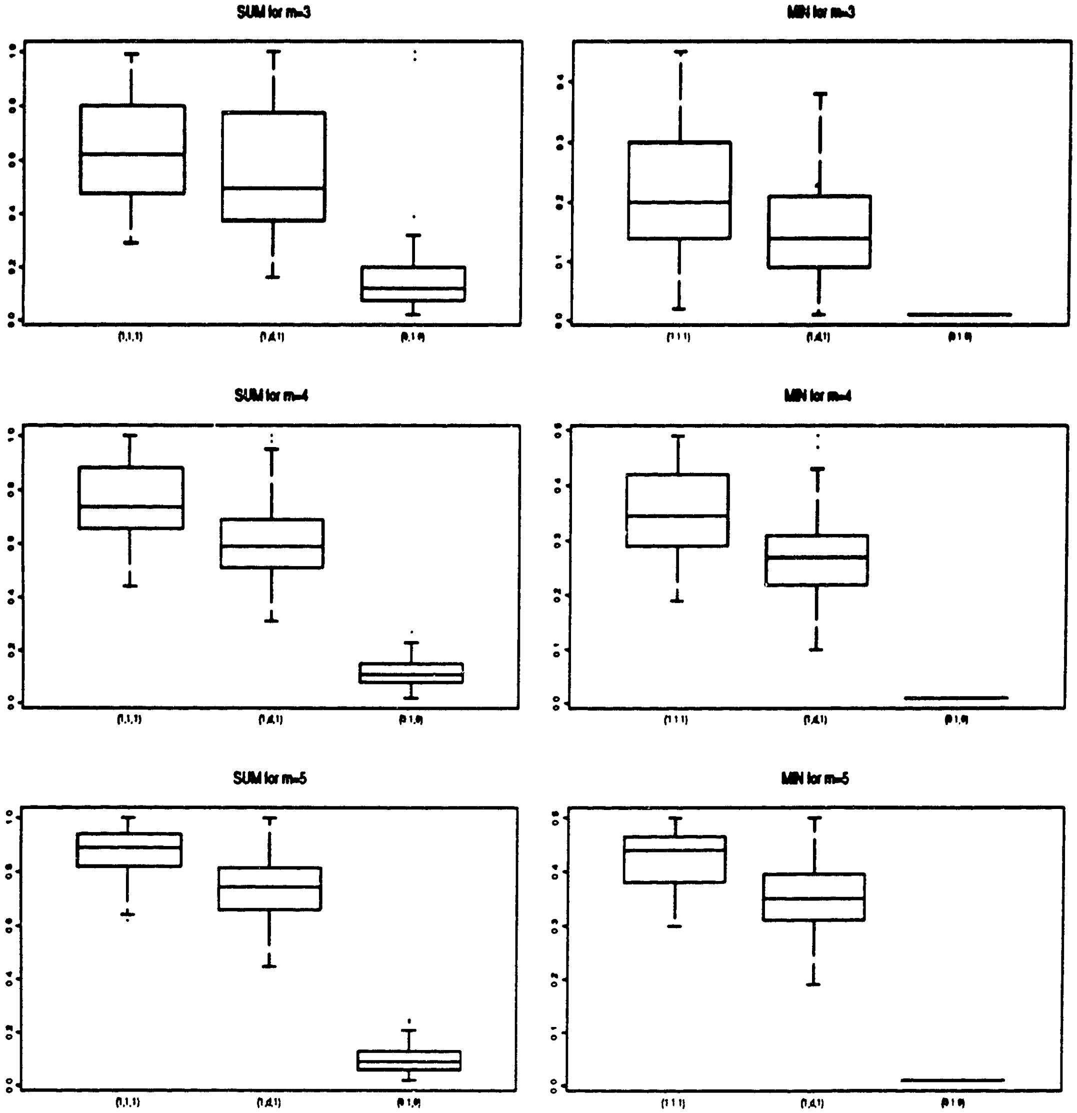


Figure 6.7: 2-dimensional bimodal distributions

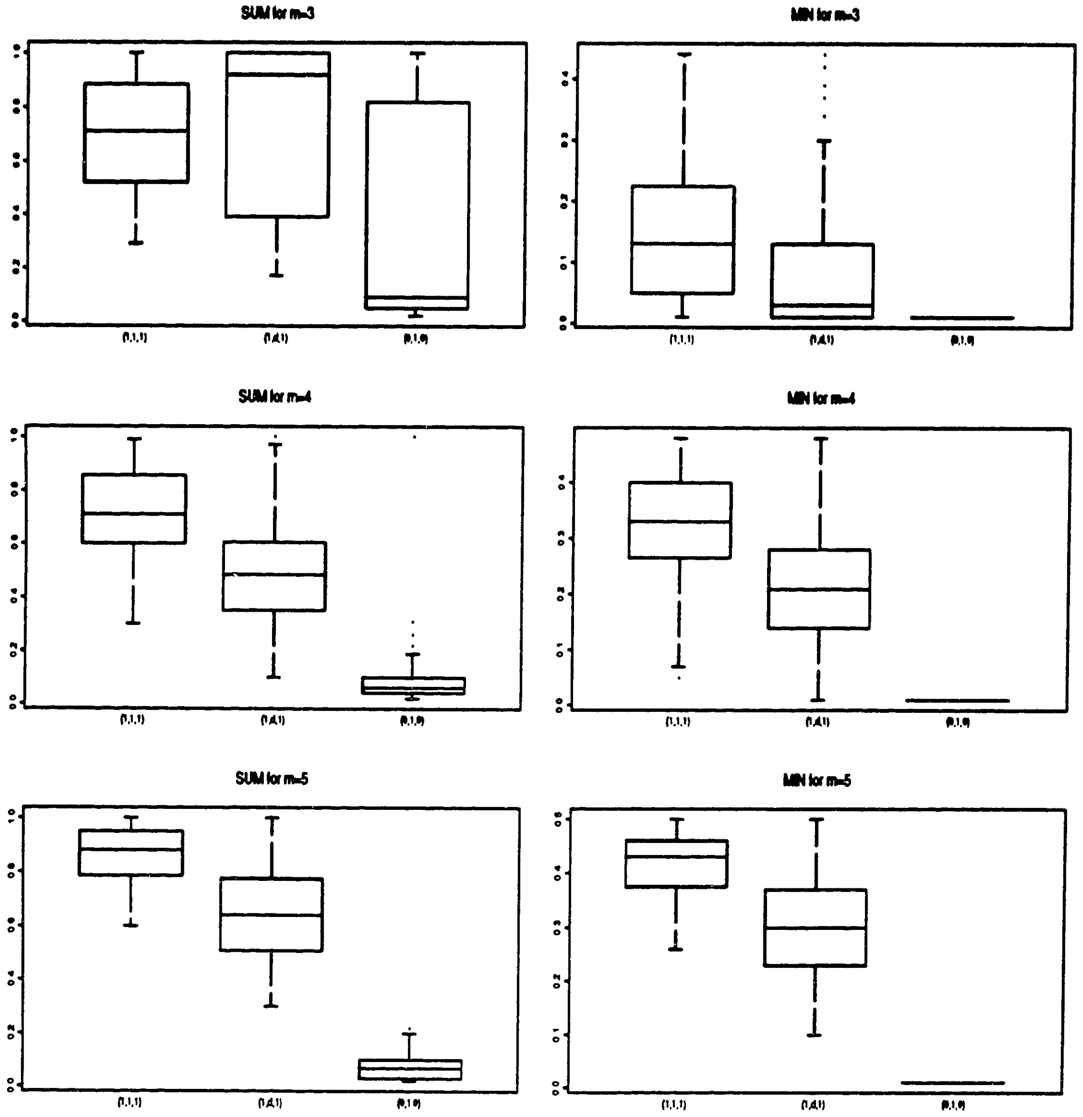


Figure 6.8: 3-dimensional bimodal distributions

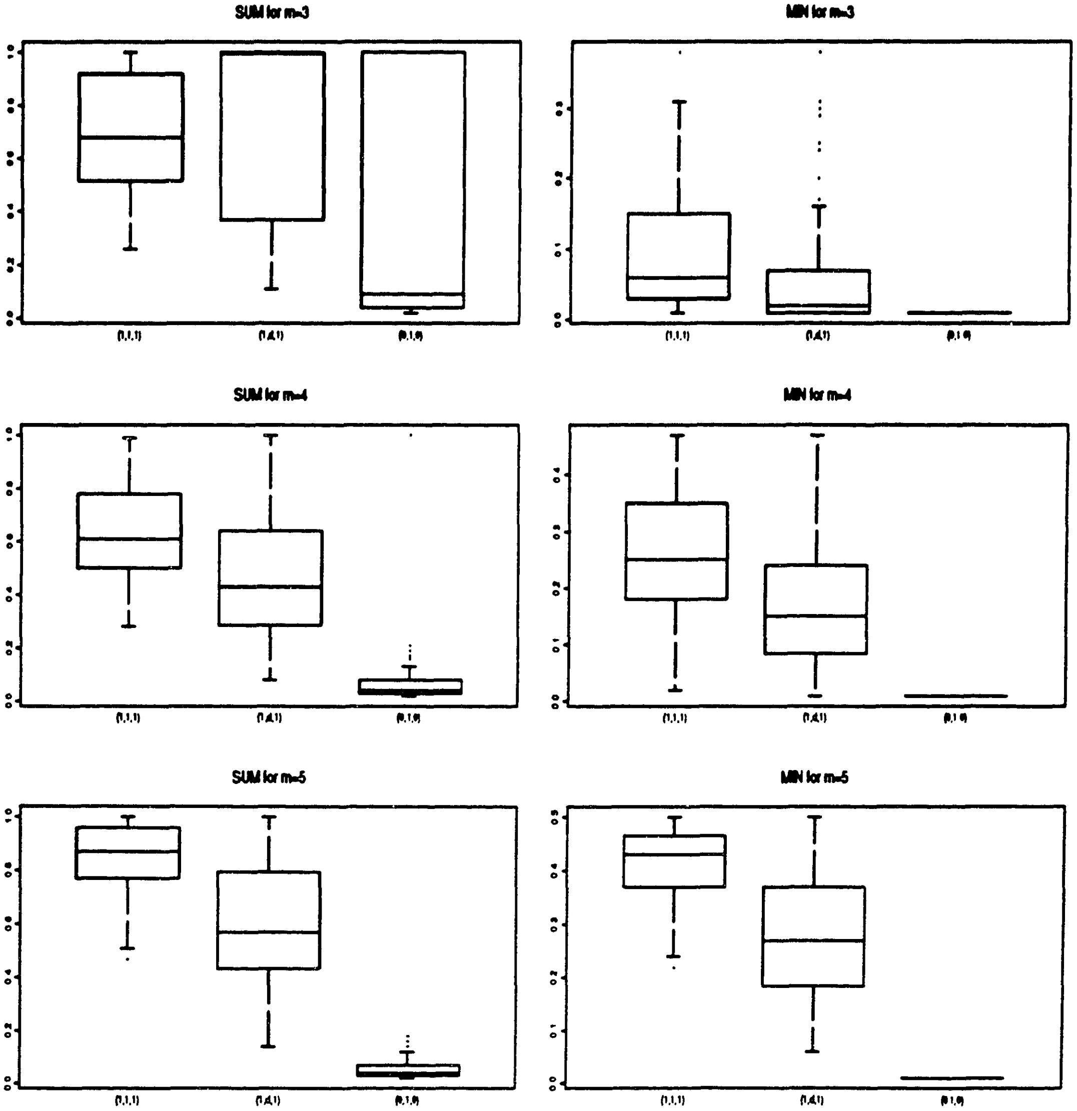


Figure 6.9: 4-dimensional bimodal distributions

Chapter 7

Finding Groups in Data

7.1 Improving on single-link clustering

It is now time to try the techniques developed mainly in the last chapter on some real data. Our point of view in this chapter is slightly different. We are not just interested in detecting multimodality but in actually finding groups in data. For this purpose, we will make use of the process $T_n^{1,1,1}(f_n, \delta)$ and the partitions of the sample it is associated with.

Our hope is that we can divise a procedure that retains the good properties of single-link clustering but at the same time, avoids the chaining problems. Our procedure will include three steps:

Step 1: Plot the process $T_n^{1,1,1}(f_n, \delta)$ and find its maximum $T_{\mathbf{max},n}^{1,1,1}$.

Step 2: Find the optimal truncation level:

$$\delta_n^{\max} := \min\{f_n(X_i)) : T_n^{1,1,1}(f_n, f_n(X_i)) = T_{\mathbf{max},n}^{1,1,1}\}.$$

Find the partition of:

$$\{X_j : f_n(X_j) \geq \delta_n^{\max}\}$$

that corresponds to $T_n^{1,1,1}(f_n, \delta_n^{\max})$.

Step 3: Decide what to do with the observations in:

$$\{X_j : f_n(X_j) < \delta_n^{\max}\}.$$

Remark 7.1.1 Notice that the procedure provides us with a partition into two groups. If further clustering is desired, the procedure can be repeated on each of the groups. Alternatively, once the low-density observations in:

$$\{X_j : f_n(X_j) < \delta_n^{\max}\} \tag{7.1}$$

have been removed (hopefully including the troublesome observations lying between clusters and causing chaining), we can rely on single-link to reveal the cluster structure of the data.

Remark 7.1.2 What to do with the observations in (7.1) will depend on what we want to do with the data. If a partition of the whole sample is desired, then we may attempt to attach each of these observations to one of the groups formed by the rest of the data. Alternatively, we may treat them as special cases forming their own group.

Thera are cases, however, where a complete partitioning of the sample is not important. Instead, we want to identify *dense* clusters only. Such an example will be given in Section 7.3.

7.2 Food Data

We will begin with a simple example involving only 27 observations in 5 dimensions. In such a case, it is not impossible to recognize the clustering structure (if such exists) and therefore we are able to check on our results.

The data in Table 7.1 are a list of the nutrients contained in 27 different kinds of meat, fish and fowl. The nutrients listed are: food energy (caloric intake), protein, fat, calcium and iron. Except for fat which is given in grams, the rest are given in percentage of the recommended daily allowance. In this case this appears to be the most informative way of scaling the variables.

The data are included on page 87 of [Har75], where they are used to demonstrate the k-means clustering method. For that purpose, k is taken equal to 3. Before applying the method of Chapter 6, we can try to identify groups by simply looking at the data. Graphical techniques such as Chernoff faces can be of assistance (see Figure 7.1)

From a preliminary examination of the data, it seems that we are dealing with the following groups:

No	Food	Energy	Protein	Fat	Calcium	Iron
1	Beef, braised	11	29	28	1	26
2	Hamburger	8	30	17	1	27
3	Beef, roast	13	21	39	1	20
4	Beef, steak	12	27	32	1	26
5	Beef, canned	6	31	10	2	37
6	Chicken, broiled	4	29	3	1	14
7	Chicken, canned	5	36	7	2	15
8	Beef, heart	5	37	5	2	59
9	Lamb leg, roast	8	29	20	1	26
10	Lamb shoulder, roast	9	26	25	1	25
11	Ham, smoked	11	29	28	1	25
12	Pork, roast	11	27	29	1	25
13	Pork, simmered	11	27	30	1	25
14	Beef tongue	6	26	14	1	25
15	Veal cutlet	6	33	9	1	27
16	Bluefish, baked	4	31	4	3	6
17	Clams, raw	2	16	1	10	60
18	Clams, canned	1	10	1	9	54
19	Crabmeat, canned	3	20	2	5	8
20	Haddock, fried	4	23	5	2	5
21	Mackerel, boiled	6	27	13	1	10
22	Mackerel, canned	5	23	9	20	18
23	Perch, fried	6	23	11	2	13
24	Salmon, canned	4	24	5	20	7
25	Sardines, canned	6	31	9	46	25
26	Tuna, canned	5	36	7	1	12
27	Shrimp, canned	3	33	1	12	26

Table 7.1: Nutrients in Meat, Fish and Fowl

- A high protein, high energy, high fat group (e.g. beef, ham and pork),
- a high protein, low energy, low fat group (e.g. chicken and fish) and
- special cases like clams (high in iron) or sardines (high in calcium).

Let us now use the methods developed in Chapter 6 to obtain a more precise clustering of the data. The first step would be to compute the process $T_n^{r_1,r_2,r_3}(f_n, \delta)$ for some choice of r_1, r_2, r_3. In view of the results of the simulations in Chapter 6, we start by letting:

$$r_1 = r_2 = r_3 = 1.$$

We plot this in Figure 7.2. Notice that each point in the plot corresponds to a partition of a number of observations. Next to each such point, we mark the number of the last food item of Table 7.1 included in the partition. As we can see, the maximum is obtained when 25 of the 27 foods are used. The other 2 foods which correspond to the points in the data set where the density estimator f_n attains its 2 lowest values are:

1. sardines, canned (No 25) and
2. shrimp, canned (No 27).

We now use the single-link method on the rest. The results are shown in the dendrogram of Figure 7.3. The first 3 divisive steps of the single-link method discover only small groups (1 – 3 foods) and separate them from the rest of the data. The fourth step, however, separates the remaining 20 foods into two groups of 8 and 12 foods. None of the subsequent steps discovers any other significantly large group. Therefore, we have discovered 2 major food groups:

Group I: Beef braised, Hamburger, Beef roast, Beef steak, Beef canned, Lamb leg roast, Lamb shoulder roast, Ham cooked, Pork roast, Pork simmered, Beef tongue, Veal cutlet.

Group II: Chicken broiled, Chicken canned, Bluefish baked, Crabmeat canned, Haddock fried, Mackerel boiled, Perch fried, Tuna canned.

In addition, we discovered 5 foods that do not seem to belong to either of the above mentioned groups and which the dendrogram links in the following way:

1. (a) Beef heart.

   (b) Clams raw, Clams canned[1].

2. Mackerel canned, Salmon canned.

To these we could add the two foods (canned sardines and canned shrimp) that were originally separated from the rest of the data.

It seems that:

- Group I includes the foods that are high in protein and high in energy and fat but low in other nutrients.

- Group II includes the foods that are high in protein but low in energy, fat and other nutrients.

As for the other 7 foods, they all seem to differ from the ones in the two groups by containing unusually high doses of iron or calcium. They do not really form a group because there are considerable differences among them but they may very well play a special role in, e.g. any balanced diet.

The results seem to confirm our initial guess based on Table 7.1 and Figure 7.1.

7.3 Market Segmentation Data

Some of the most interesting applications of cluster analysis come from marketing research (see [Chu91] and [PS83]). One particular problem is the use of a questionnaire to obtain information about the existence of segments in a particular market. So, for example, a manufacturer may be interested in identifying homogeneous groups (segments) among clients in order to target them with products particularly suited for those groups. Since the process of developing such *custom-made* products is always costly, the manufacturer must be convinced that such segments do exist. The questions asked in such

[1] These may very well be considered two forms of the same food.

a research will almost certainly include demographic characteristics (age, income, number of children etc) but other questions may also be included. For an informative introduction to the subject, see [Win78].

In our example, we will use data collected on behalf of Fabhus, an Atlanta manufacturer of prefabricated homes who saw their business decline in the late 80's, after a booming start in the late 70's. The researchers mailed questionnaires to old customers in an effort to reveal the *customer profile* as well as collect information about preferences, previous housing choices and degree of satisfaction. 293 questionnaires were returned. We will concentrate on the demographic questions, namely:

Question 1: Number of children living at home:

0: 0,

1: 1,

2: 2,

3: 3,

4: 4,

5: 5 or more,

9: blank.

Question 2: Age of household head:

0: blank,

1: under 20,

2: 20-24,

3: 25-34,

4: 35-44,

5: 45-54,

6: 55-64,

7: 65 or more.

Question 3: Occupation of head of household:

0: blank,

1: professional or official,

2: technical or manager,

3: proprietor,

4: farmer,

5: craftsman,

6: clerical or sales,

7: labor or machines operator,

8: foreman,

9: service worker,

10: retired,

11: other,

12: if more than one checked.

Question 4: Family income bracket:

0: blank,

1: Under $6,000,

2: $6,000-$11,999,

3: $12,000-$17,999,

4: $18,000-$23,999,

5: $24,000-$29,999,

6: $30,000-$35,999,

7: $36,000-$41,999,

8: $42,000 or over.

Question 5: Spouse's employment status:

0: blank,

1: spouse employed full time,

2: spouse employed part time,

3: spoude not employed,

4: not married.

Notice that the variables are not continuous and, therefore, there is no true density f estimated by f_n. In this case, f_n simply measures how isolated an observation is. This is, however, exactly what we need to know in order to avoid chaining effects. Since the scale of the variables is different, we began by scaling them (dividing by their sample standard deviation). Then, to obtain a first look at the data we plotted the first two principal components (Figure 7.4). Because two or more persons may have given identical answers to the 5 questions, each observation is represented by a circle with radius proportional to the number of observations with the same coordinates. There seems to be indication of the existence of two clusters, a small one on the left and a much larger one on the right. However, because of the large number of observations that lie in between, it would be impossible to recover this structure by the single-link method.

Instead, we proceed in the way outlined in Section 7.1. We choose again to compute the process $T_n^{1,1,1}(f_n, \delta)$ which is plotted in Figure 7.5. The maximum is attained when 133 observations are included. The partition that this maximum corresponds to can be seen in the dendrogram of Figure 7.6.

What is most interesting perhaps is the interpretation of these two groups.

Group I: The smaller group on the left includes 32 individuals, all of them over 55. 29 of the 32 are retired. Their spouses are not employed. None of them has children currently living in their house and their income belongs to the three lower income brackets ($ 17,999 or less).

Group II: The larger group on the right includes 101 individuals belonging to various professions. Their ages range between 22 and 44 and their income belongs to the middle brackets ($ 12,000 to 29,999).

When we now project these 133 observations onto the plane defined by the first two principal components of the whole data set, we can clearly see the two groups (Figure 7.7). It is important to note that we have succeeded in visualizing cluster structure in 5 dimensions where a more traditional visualization method (principal components) failed. This is partly because the removal of low density observations besides removing the observations that cause chaining, also removes outliers to which principal components are sensitive (see [Hub85]).

Of course, we only obtained a partition of 45% of the data. However, this is completely satisfactory in this case. It is unlikely that finding a group for *each* individual that returned the questionnaire is going to be possible or even useful. What Fabhus would have liked to know is the *strongest* groups of customers in their market. For such groups, they can develop new marketing strategies in the future. Besides, a well-known rule in marketing states that 80% of the business comes from 20% of the customers. It is this 20% that deserves the attention of Fabhus.

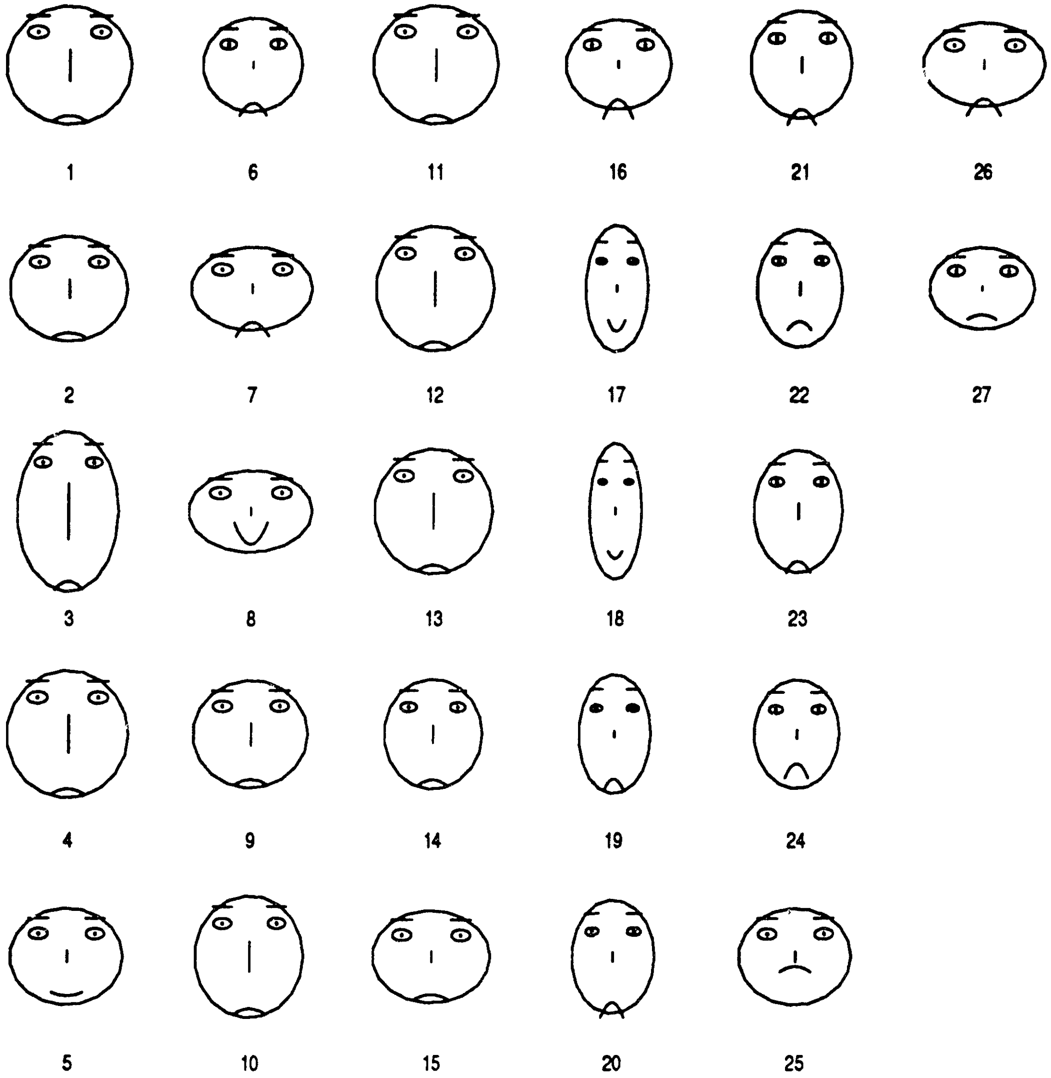


Figure 7.1: Chernoff faces for the food nutrient data

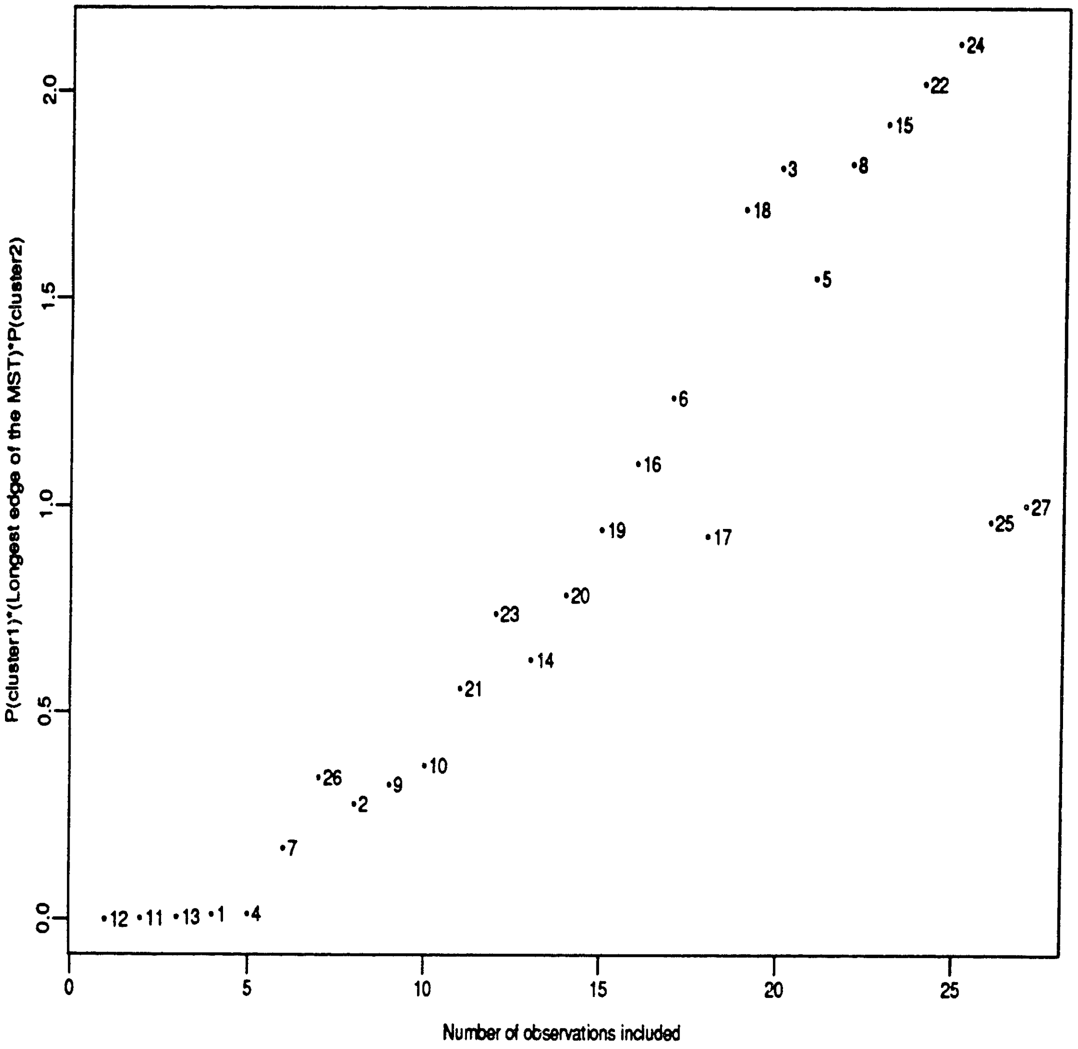


Figure 7.2: Choosing the best truncation level using $T_n^{1,1,1}(f_n, \delta)$.

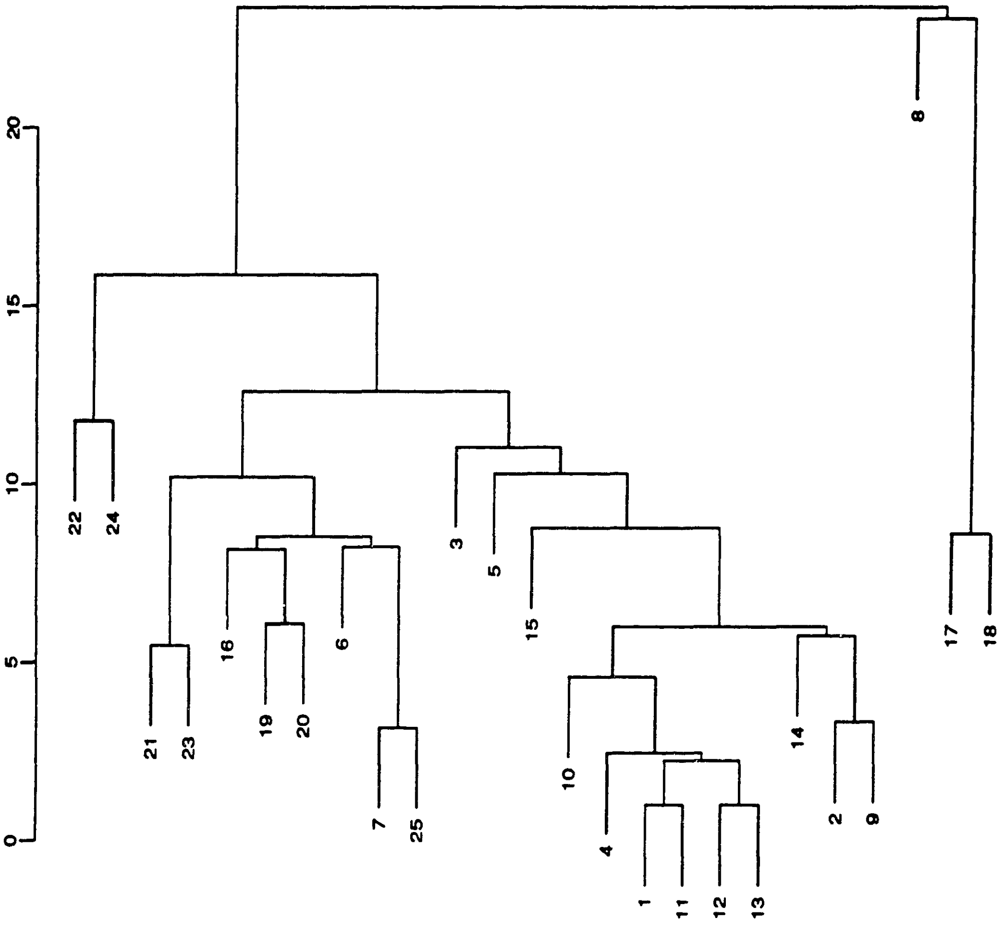


Figure 7.3: Single-link dendrogram for the food data.

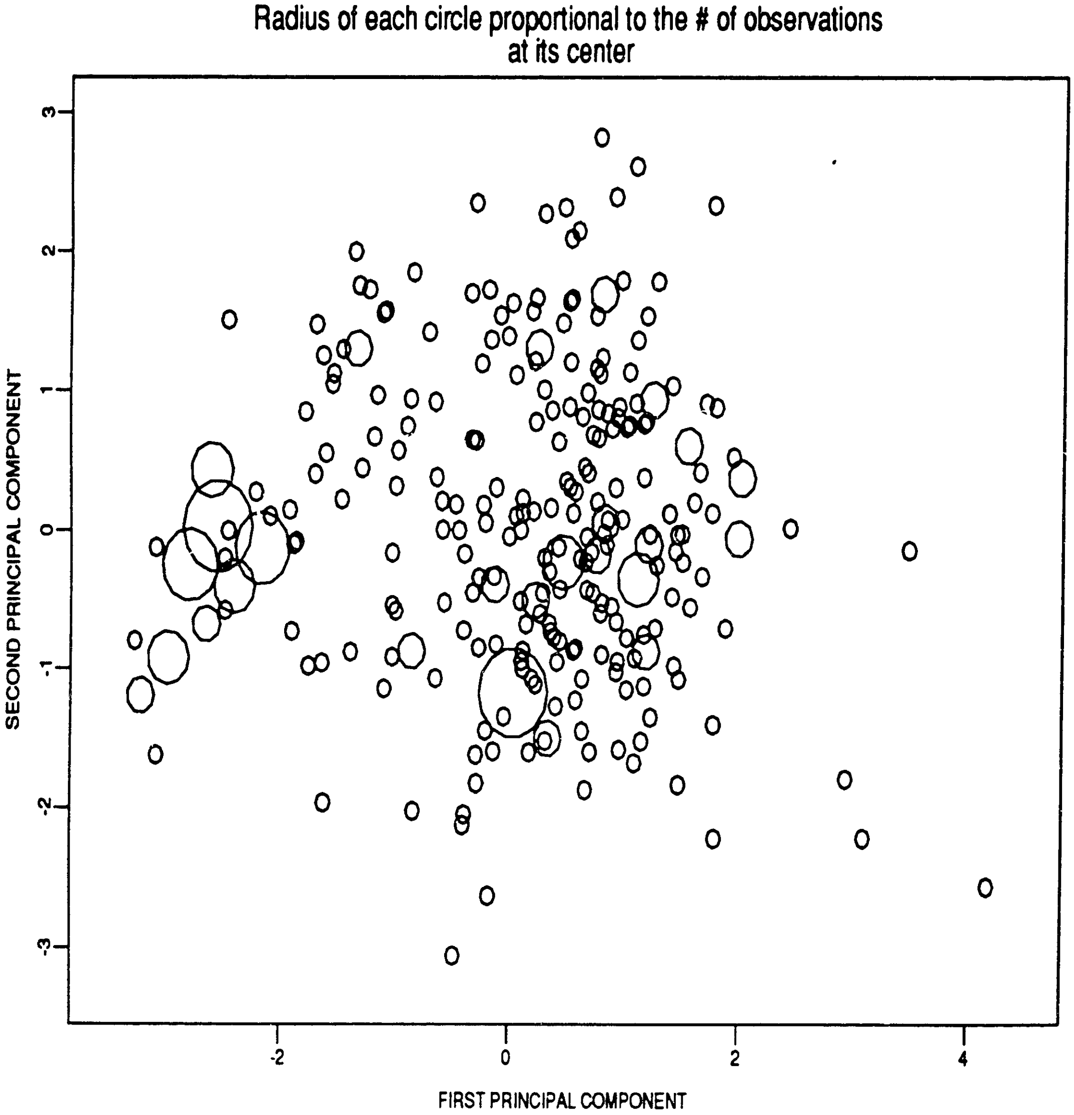


Figure 7.4: The first two principal components for the Fabhus data.

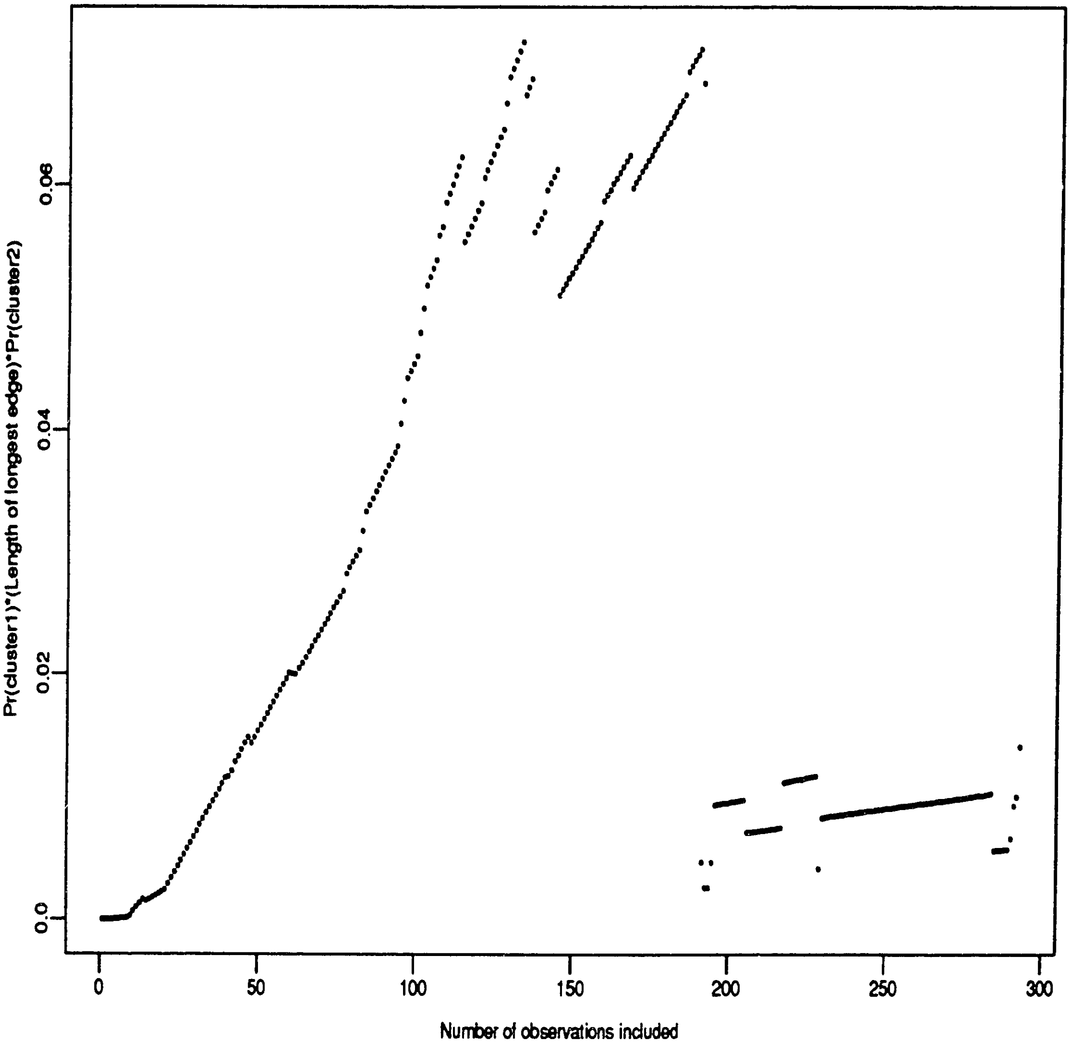


Figure 7.5: The process $T_n^{1,1,1}(f_n, \delta)$ for the Fabhus data.

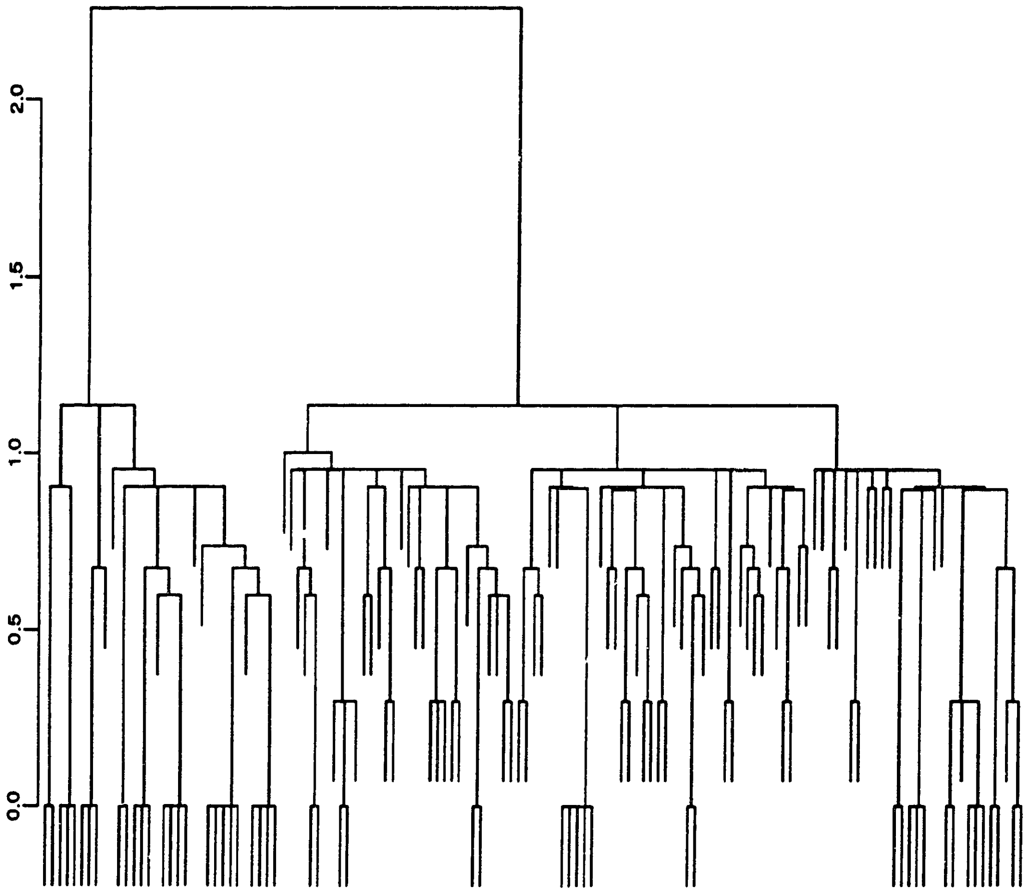


Figure 7.6: Single-link dendrogram for the 133 observations of the Fabhus data.

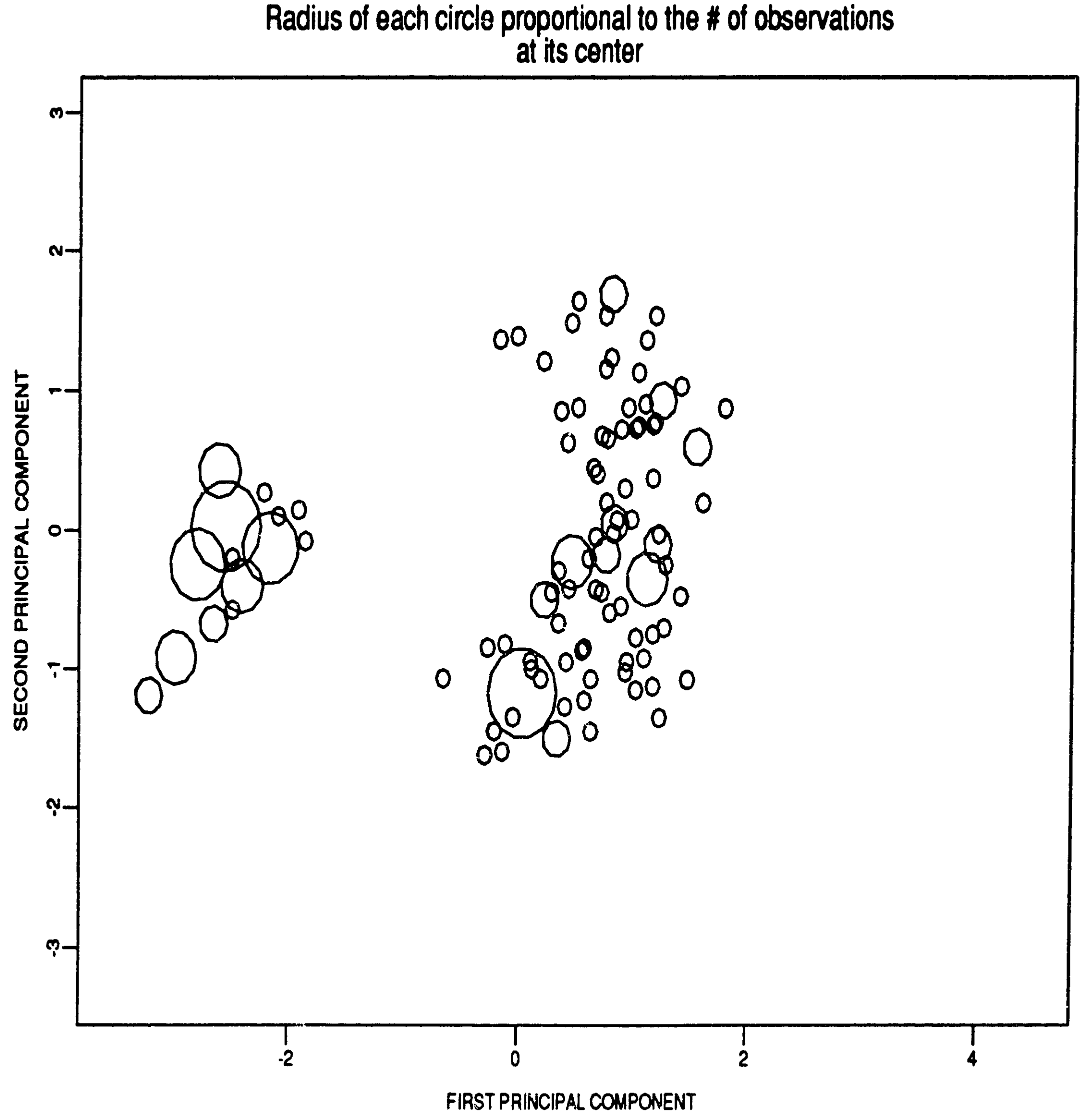


Figure 7.7: The truncated Fabhus data projected on the same plane as before.

Bibliography

[AB92] F. Avram and D. Bertsimas. The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach. *Annals of Applied Probability*, 2:113–130, 1992.

[BB83] Peter Bickel and Leo Breiman. Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Annals of Statistics*, 11:185–214, 1983.

[BCW88] Richard A. Becker, John M. Chambers, and Alan R. Wilks. *The New S Language: A programming environment for data analysis and graphics.* Wadsworth & Brooks/Cole, Pacific Grove, CA, 1988.

[Ber66] Robert Berk. Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, 37:51–58, 1966.

[BFOS84] Leo Breiman, Jerome Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees.* Wadsworth and Brooks/Cole, Pacific Grove, CA, 1984.

[Boc85] H. H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2:77–108, 1985.

[BR78] Monique Bertrand-Retali. Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Rev. Roum. Math. Pures et Appl.*, 30(3):361–385, 1978.

[But86] R. W. Butler. Optimal stratification and clustering on the line using the L^1-norm. *Journal of Multivariate Analysis*, 19:142–155, 1986.

[But88] R. W. Butler. Optimal clustering on the real line. *Journal of Multivariate Analysis*, 24:88–108, 1988.

[BvR90] Dimitris J. Bertsimas and Garett van Ryzin. An asymptotic determination of the minimum spanning tree and minimum matching constants in geometric probability. *Operations Research Letters*, 9:223–231, 1990.

[Cac66] T. Cacoullos. Estimation of a multivariate density. *Ann. Inst. Statist. Math.*, 18:179–189, 1966.

[Chu91] Gilbert A. Churchill, Jr. *Marketing Research, Methodological Foundations.* The Dryden Press, Chicago, fifth edition, 1991.

[CM88] J. A. Cuesta and C. Matran. The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory and Related Fields*, 78:523–534, 1988.

[Dar53] D. A. Darling. On a class of problems related to the random division of an interval. *Annals of Mathematical Statistics*, 24:239–253, 1953.

[Deh84] Paul Deheuvels. Strong limit theorems for maximal spacings from a univariate distribution. *Annals of Probability*, 12:1181–1193, 1984.

[DEMR88] Paul Deheuvels, John Einmahl, David Mason, and Frits H. Ruymgaart. The almost sure behavior of maximal and minimal multivariate k_n spacings. *Journal of Multivariate Analysis*, 24:155–176, 1988.

[DH89] Holger Dette and Norbert Henze. The limit distribution of the largest nearest-neighbour link in the unit d-cube. *Journal of Applied Probability*, 26:67–80, 1989.

[Don88] David L. Donoho. One-sided functional inference about functionals of a density. *Annals of Statistics*, 16:1390–1420, 1988.

[Dud89] R. M. Dudley. *Real Analysis and Probability.* Wadsworth & Brooks/Cole, Pacific Grove, CA, 1989.

[Eub88] R. L. Eubank. Optimal grouping, spacing, stratification and piecewise constant approximation. *SIAM Review*, 30:404–420, 1988.

[Eve74] B. Everitt. *Cluster Analysis.* Halsted Press, New York, 1974.

[FRb67] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62:1159–1178, 1967.

[FRf79] Jerome H. Friedman and Laurence C. Rafsky. Multivariate genaralizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 17:697–717, 1979.

[FRf81] Jerome H. Friedman and Laurence C. Rafsky. Graphics for the multivariate two-sample problem (with comments). *Journal of the American Statistical Association*, 76:277–293, 1981.

[Gil80] Laurence S. Gillick. *Iterative Ellipsoidal Trimming.* PhD thesis, Massachusetts Institute of Technology, 1980.

[Gor81] A. D. Gordon. *Classification.* Chapman and Hall, London, 1981.

[Gor87] A. D. Gordon. A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A*, 150, Part 2:119–137, 1987.

[GR65] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series and Products.* Academic Press, New York and London, fourth edition, 1965.

[GR69] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Appl. Stat.*, 18:54–64, 1969.

[Hal46] Paul R. Halmos. The theory of unbiased estimation. *Annals of Mathematical Statistics*, 17:34–43, 1946.

[Har75] J. A. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

[Har78] J. A. Hartigan. Asymptotic distributions for clustering criteria. *Annals of Statistics*, 6:117-131, 1978.

[Har81] J. A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76:388–394, 1981.

[Har85] J. A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63–76, 1985.

[Har88] J. A. Hartigan. The span test for unimodality. In H. H. Bock, editor, *Classification and Related Methods of Data Analysis*, pages 229–236. North-Holland, Amsterdam, 1988.

[Hen83] Norbert Henze. Ein asymptotischer Satz über den maximalen Minimalabstand von unabhängigen Zufallsvektoren mit Anwendung auf einen Anpassungstest im R^d und auf der Kugel. *Metrika*, 30:245–259, 1983.

[Hen88] Norbert Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Annals of Statistics*, 16:772–783, 1988.

[HH85] J. A. Hartigan and P. M. Hartigan. The dip test for unimodality. *Annals of Statistics*, 13:70–84, 1985.

[Hoe48] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325, 1948.

[HRRS86] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The approach based on Influence Functions.* Wiley, New York, 1986.

[Hu82] T. C. Hu. *Combinatorial Algorithms.* Addison-Wesley, Reading, MA, 1982.

[Hub81] Peter J. Huber. *Robust Statistics.* Wiley, New York, 1981.

[Hub85] Peter J. Huber. Projection pursuit. *Annals of Statistics*, 13:435–475, 1985.

[Hub91] Peter J. Huber. Goals and algorithms. Technical Report PJH-91-2, Massachusetts Institute of Technology, Department of Mathematics, Cambridge, MA 02139, September 27, 1991.

[JD88] A. K. Jain and R. C. Dubes. *Algorithms for clustering data.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[Joh67] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

[Kit76] Josef Kittler. A locally sensitive method for cluster analysis. *Pattern Recognition*, 8:23–33, 1976.

[Kit79] Josef Kittler. Comments on 'Single-link characteristics of a mode-seeking clustering algorithm'. *Pattern Recognition*, 11:71–73, 1979.

[KR90] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data.* Wiley, New York, 1990.

[KS89] Hyune-Ju Kim and David Siegmund. The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, 76:409–423, 1989.

[Lévy39] Paul Lévy. Sur la division d'un segment par des points choisis au hasard. *C.R.Acad.Sci. Paris*, 208:147–149, 1939.

[LMW84] Ludovic Lebart, Alain Morineau, and Kenneth M. Warwick. *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices.* Wiley, New York, 1984.

[MKB79] K .V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, London, 1979.

[NW88] George L. Nemhauser and Laurence A. Wolsey. *Integer and Combinatorial Optimization.* Wiley, New York, 1988.

[PFvN89] Roger Peck, Lloyd Fisher, and John van Ness. Approximate intervals for the number of clusters. *Journal of the American Statistical Association*, 84:184–191, 1989.

[Pol81] David Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9:135–140, 1981.

[Pol82] David Pollard. A central limit theorem for k-means clustering. *Annals of Probability*, 10:919–926, 1982.

[PS82] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization. Algorithms and Complexity.* Prentice-Hall, Englewood Cliff, NJ, 1982.

[PS83] Girish Punj and David W. Stewart. Cluster analysis in marketing research: Review and suggestions for applications. *Journal of Marketing Research*, 20:134–148, 1983.

[Pyk65] R. Pyke. Spacings (with discussion). *Journal of the Royal Statistical Society, Series B*, 27:395–449, 1965.

[Rao83] B. L. S. Prakasa Rao. *Nonparametric Functional Estimation.* Academic Press, Orlando, FL, 1983.

[Rip88] B. D. Ripley. *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge, 1988.

[Ros56] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

[SDJ79] Edward Schaffer, Richard Dubes, and Anil Jain. Single-link characteristics of a mode-seeking clustering algorithm. *Pattern Recognition*, 11:65–70, 1979.

[Ser80] Robert J. Serfling. *Approximation theorems of mathematical statistics.* Wiley, New York, 1980.

[Sil81] B. W. Silverman. Using kernel density estimators to investigate multimodality. *Journal of the Royal Statistical Society, Series B*, 43:97–99, 1981.

[Sil86] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[ST88] J. Michael Steele and Luke Tierney. Boundary domination and the distribution of the largest nearest neighbor link in higher dimensions. *Journal of Applied Probability*, 23:524–528, 1988.

[Sta86] Richard P. Stanley. *Enumerative Combinatorics*, volume I. Wadsworth and Brooks/Cole, Monterey, California, 1986.

[Ste88] J. Michael Steele. Growth rates of euclidean minimal spanning trees with power weighted edges. *Annals of Probability*, 16:1767–1787, 1988.

[Str75] David J. Strauss. A model for clustering. *Biometrika*, 62:467–475, 1975.

[Win78] Yoram Wind. Issues and advances in segmentation research. *Journal of Marketing Research*, 15:317–337, 1978.

[WL83] M. Antony Wong and Tom Lane. A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society, Series B*, 45:362–368, 1983.

[Won82] M. Antony Wong. A hybrid clustering method for identifying high-density clusters. *Journal of the American Statistical Association*, 77:841–847, 1982.