APPROPRIATE NON-PARAMETRIC CLASSIFICATION OF INCUBATION PERIOD DATA

A dissertation submitted to the University of Manchester
for the degree of Master of Science
in the Faculty of Science and Engineering

2018

Syed Humza Naqvi Ali
School of Mathematics

Contents

List of Tables

List of Figures

Abstract

Hierarchical clustering is commonly applied with traditional measures of distance between two clusters computed by agglomerative hierarchical algorithms. This dissertation assesses the validity of the classification of non-parametric tests as a measure of dissimilarity for incubation period data. A method of applying p-values obtained from bootstrapped versions of non-parametric tests are discussed and uncertainties regarding bootstrapping, sample size of studies and linkage algorithms are assessed with the Kolmogorov-Smirnov, Anderson-Darling and Cramer-von tests.

All three tests produced similar results for their application with 3 different linkage algorithms, hence the transformation of the distance dissimilarity matrix to a hierarchical structure is not affected by the choice of non-parametric test. The Kolmogorov-Smirnov and Anderson-Darling resulted with the least uncertainties from bootstrapping, where the Kolmogorov-Smirnov test was found to be slightly better and the Cramer-von test was seen to be the least reliable distance measure.

Declaration

No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Intellectual Property Statement

i. The author of this dissertation (including any appendices and/or schedules to this dissertation) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the dissertation, for example graphs and tables ("Reproductions"), which may be described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Dissertation restriction declarations deposited in the University Library,

The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's Guidance on Presentation of Dissertations.

Acknowledgements

I would like to thank my supervisor, Dr. Ian Hall, for his constant support, encouragement and the advice he provided throughout my dissertation. I have been extremely fortunate to have a supervisor who responded to my questions promptly as well as allocating plenty of time each week to provide feedback and instructions.

My greatest appreciation to Adedoyin Awofisayo-Okuyelu for providing the data used within this dissertation, also for her paper, which this dissertations methodology is based on. Without the methodology proposed in the paper, this dissertation would not be possible.

I must express my gratitude to the tex.stackexchange community for their support, suggestions and advice on LaTeXand the Stackoverflow community for introducing me to several functions and providing support for R questions. Specifically I am very grateful to Tal Galili for developing the R package 'dendextend' and his willingness to answer questions related to his package.

Finally I must give my appreciation to my family for their financial support, especially my mother whos love and guidance are always with me and my friends for their continued inspiration and encouragement.

Chapter 1

Introduction

The incubation period is the time between the exposure or infection of a pathogenic organism and onset of clinical symptoms (Brookmeyer 2014). This time period varies largely depending on the dose of infectious agent, rate of replication of infectious agent, the mechanism of disease development and other underlying factors related to the host (Nishiura Hiroshi 2007). Due to these factors, the incubation period distribution is not clearly defined for many diseases as these factors are highly dependent on the type of disease.

Understanding and quantifying the variability of the incubation period acquires many benefits in various fields. For example, in clinical practices the distribution can be used to determine the sources of infection, develop treatments to extend the incubation period and predict disease prognosis. Whereas for public health practice the distribution determines the length of quarantine required for individuals exposed to deadly diseases. Several benefits for other fields are mentioned in (Nishiura Hiroshi 2007). Incorrectly estimating the distribution may result in composing inaccurate case definitions, inaccurate exposure times and excluding outbreaks as sporadic or travel related cases. Therefore, understanding the distribution of the incubation period plays an important role in methods for estimating the direction in which the epidemic or an outbreak is heading. For example, Gail & Brookmeyer (1988) developed three methods for projecting the short-term course of the AIDS epidemic, in particular the

back-calculation technique with the incubation period to estimate HIV prevalence and predict the future incidence of AIDS (Bacchetti & Jewell 1991).

As mentioned, the incubation period is highly dependent on the underlying factors of an epidemic thus epidemics distributions are unknown as the time period varies accordingly to specific cases. There have been several attempts and many methodologies developed in the last century to understand and define the incubation distribution. These developments are summarised by Nishiura Hiroshi (2007) which documents the early efforts of modeling the incubation period and evolution of models proposed.

1.1 Models Proposed for the Incubation Distribution

Although incubation period data tend to follow a continuous distribution, for pragmatic reasons the incubation period reports as a discrete distribution. Miner (1922) documented the first explicit model of the incubation period for typhoid fever by analysing the change in variance and calculating moments of outbreaks, deriving equations to explain the epidemic curve. Miner modeled the Pearson's type I distribution due to its acceptability of right-skewness. The log normal distribution, proposed by Sartwell (1949) who determined the incubation period of acute infectious diseases tends to follow a lognormal distribution and modeled the lognomral distribution with various diseases. Sartwell suggested the use of an estimated median and a dispersion factor as a measure of variability due to distributions often being skewed to the right and from this he developed a method to estimate the time of exposure during a point source outbreak. Since Sartwell's findings, many predominant Japanese epidemiologists proposed improvements for Sartwell's work. For example, Tango (1998) suggested the maximum likelihood estimate as a method to obtain reasonable estimators to the parameters of the log-normal distribution. More recently, the gamma distribution has been proposed for SARS and smallpox by (Eichner & Dietz 2003) and further research by Donnelly *et al.*,(2003) and others have presented the gamma distribution as a suitable model for incubation period. Hence both the gamma and log-normal distributions are commonly assumed as apriori models for incubation period distribution.

Much of the focus in the last century has prioritized on understanding and identifying the distribution of incubation period with a substantial increase in the number of studies for several epidemics. A considerable amount of these studies focuses on AIDS in which accurate estimations of the incubation period are discussed with both prospective and retrospective studies. Before Bacchetti & Jewell (1991) all studies investigated the incubation period of AIDS with only parametric methods which have several drawbacks when compared to non-parametric techniques. Non-parametric estimates provide a useful approach in identifying possible parametric models and assessing the goodness of fit of a parametric analysis. This is particularly useful due to the lack of understanding of the disease mechanism making it difficult to find a priori support for a specific parametric description.

The main goal for epidemiologist's concerns estimating the distribution of the incubation period of any disease resulting in many desirable practical applicabilities. Unfortunately, there is very little literature on classifying and clustering incubation period data parametrically and non-parametrically. Recently, (Awofisayo-Okuyelu et al. 2017) conducted a systematic review and meta-analysis on the incubation period data of campylobacteriosis. In which the authors aimed to estimate the distribution of incubation period data for campylobacteriosis to identify and explain the variation in the distribution between studies. They tested for heterogeneity with traditional measures and developed a method to classify discrete distributions of data non-parametrically with the Kolmogorov-Smirnov test as its basis. By deploying hierarchical clustering they grouped the studies together with a bootstrapped version of the Kolmogorov-Smirnov test as a distance metric with the complete-linkage algorithm to identify relationships between studies. This dissertation aims to evaluate the methodology within this paper, specifically the uncertainties within hierarchical clustering.

1.2 Clustering

A widely used approach in machine learning and statistics is clustering. Cluster analyses is a method for creating groups of objects or clusters together in such a way that objects within the same cluster share some similarity compared to objects in different clusters that are distinct. Bock (1989) discussed the criteria required for all objects in a cluster to satisfy. Clusters must share the same or closely related properties, show small mutual distances/dissimilarities, have 'relations' with at least one other object in the group and be clearly distinguishable from the rest of the objects in the data set. Carmichael et al. (1968) suggested the set contains clusters of points if the distribution of the points are continuous for relatively dense populated regions of space and if they are surrounded by continuous and empty regions of space.

In the literature review of data clustering, similarity measures/coefficients and dissimilarity measures/distances describe quantitatively the similarity or dissimilarity of two objects within a cluster or two clusters. Generally, all clustering algorithms are based on an index of similarity or dissimilarity between two objects (Jain 1988). Several traditional measures of dissimilarity distances such as the Euclidean, Manhattan, Minkowski and others are thoroughly discussed by Sneath (1973), Anderberg (1973), Gordon (1999) and Everitt (1980). When conducting a cluster analysis, the choice of dissimilarity is subjective, hence there is no realistic method to insure the optimal dissimilarity measure has been chosen. This relies heavily on the type of multivariate data, but certain clustering techniques are applicable for specific data sets. For example, for large multivariate data sets, K-means is especially useful whereas hierarchical clustering becomes impractical unless other techniques are incorporated (Zaït & Messatfa 1997). Once a dissimilarity measure is finalised a linkage method for hierarchical clustering is required. This too is subjective although in previous literature it has been suggested that the complete algorithm is favored over single and average linkage due to its ability to retain information and form compact clusters that are as internally homogeneous as possible (Baker 1972).

Clustering techniques can be split into Hard and fuzzy clustering. Hard clustering can further be broken down into partitional and hierarchical. Hierarchical clustering divides a data set into a sequence of nested partitions and contains agglomerative and divisive methods. Agglomerative can be seen as a bottom-up method where every cluster starts in its own cluster and is repeatedly paired with its closest neighbour via some similarity criteria resulting in one cluster. Divisive Hierarchical clustering is a top-down method where all objects initialise in one cluster and repeatedly divide into smaller clusters. Both methods of hierarchical clustering suffer from the same drawbacks where objects are incorrectly assigned to a group at an early stage and therefore unable to be grouped later if a pair has higher similarity. Secondly, different similarity measures abd algorithms within each method lead to diverse results requiring subjective analysis.

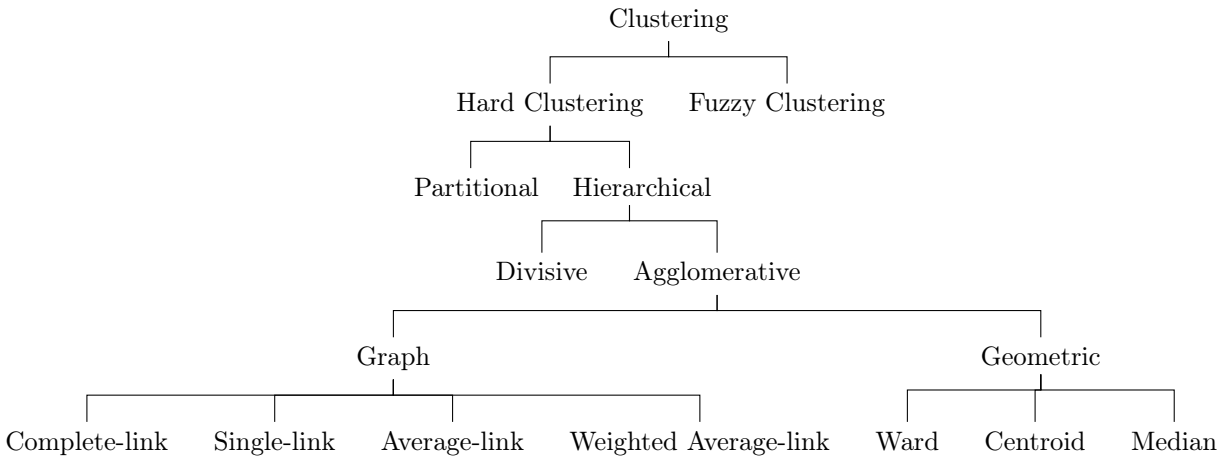


Figure 1.1: *Diagram of clustering algorithms with common methods for hierarchical clustering.*

Once a dissimilarity measure and linkage algorithm are selected and applied to a data set, the hierarchical clustering structure can be represented graphically and symbolically. The most common are n-tree, banner, pointer representation, packed representation, icicle plot and dendrogram. A dendrogram is the traditional method of visualizing hierarchical clustering, consisting of an n-tree where each internal node is associated with a height that satisfies some conditions such as the ultrametric (Gordon 1987).

1.3 Validation of Hierarchical Clustering

Generally, cluster validity is split into statistical testing and non-statistical testing. Statistical testing includes external criteria and internal criteria whereas non-statistical testing consists of relative criteria. External criteria evaluates clustering results based on a respecified structure reflecting the intuitive structure of a data set by employing the Monte Carlo technique for two approaches (Halkidi et al. 2002). Internal criteria evaluates a clustering structure with quantities and features acquired from the data set and relative criteria chooses the best clustering result from a set of predefined criteria.

1.3.1 Comparing two Hierarchical Clusters

Several methods have been proposed to compare two hierarchical clustering structures in literature. Johnson (1968) suggested Lorenz curves with the curve furthest away from the diagonal indicates the better hierarchical structure. Others derived indices such as Rand (1971) proposed the R_k index, Anderberg (1973) the Jaccard index and Hubert & Levin (1976) the δ index. These indices are functions of the matrix $[m_{ij}]$ therefore their approach to distinguish two clusters are similar but differ in terms of definition of the index. Arabie & Boorman (1973) studied the similarity between the measures with multidimensional scaling from $[m_{ij}]$ and information theory. Baker (1974) focused on the goodness of fit of dendrograms with the Goodman-Kruskal gamma coefficient in which he compared the complete-linkage algorithm to single-linkage. Fowlkes & Mallows (1983) proposed the method of B_k with a sequence of measures as the basis for a plotting procedure. They compared this method to both Rand's and Baker's indices by Monte-Carlo simulation for perturbation, multiple cluster and other experiments in which they concluded, comparing two hierarchical clusters is not a one-dimensional concept.

Generally, there are three aspects to consider when comparing tree's or dendrograms produced from hierarchical clustering: constructing a consensus of a set of trees, measuring the degree of consensus among trees in a given set and measuring dissimilarity between trees. The methods above specifically focus on comparing two partitions by

cutting the dendrograms at certain stages during the clustering process. Others have dedicated more attention to comparing two dendrograms obtained from the same set of objects. Sokal & Rohlf (1962) devised a method to compare two dendrograms directly by calculating an ordinary product-moment correlation coefficient between corresponding elements of two matrices of cophenetic values. Waterman & Smith (1978) defined a metric on binary trees, which counts the minimum number of nearest neighbor interchanges required to change one tree to another. Day (1985) introduced the ideas of algorithms for comparing trees with labeled leaves. Reilly et al. (n.d.) developed Cohen's k statistic by understanding the null distribution of the maximum number of k clusters, and derived Cohen's kappa to measure the correlation between two clustering methods. The most recent work published by Morlini & Zani (2012) proposes a new index Z for measuring the global dissimilarity between two hierarchical clusterings for dendrograms.

Therefore, in the literature review there have been many attempts to understand how two hierarchical clusterings differ by either comparing the clustering procedure or the dendrograms. These indices and metrics provide a method to compare the clusterings and although many use different methods of assessing the difference, the results are generally conclusive and consistent.

1.4 Aims and Objectives

Following on from Awofisayo-Okuyelu et al. (2017) systematic review of incubation period for campbylobacterosis, the main objective of this dissertation is to evaluate the methodology proposed by Dr. Hall. This dissertation uses similar appropriate non-parametric tests to the KS test as a distance metric. Firstly, the subgroup analysis for the same data collected by Awofisayo-Okuyelu will be reproduced and then the same will be tested for different tests. This leads to the first aim of this project which is to validate this method to understand whether the p-values of non-parametric tests are reliable in classifying incubation period data. The second aim is to assess the measures behind the uncertainty with this methodology and the magnitude of each uncertainty.

Factors which may affect the clustering are:

1. Bootstrapping of non-parametric tests.
2. The effect of sample size of studies. i.e. whether removing a big study has a big impact to the clustering.
3. The effect of agglomerative hierarchical algorithms.

Therefore the main goal of this dissertation is to, firstly, assess if transforming non-parametric tests to distance is a reliable method to cluster incubation period data and, secondly, determine which test has the least uncertainty under certain regulatory conditions such as bootstrap sample size.

Chapter 2

Methodology

The methodology chapter begins with a brief introduction to the data collected by Awofisayo-Okuyelu et al. (2017) and proceeds to general applications of non-parametric tests, by defining test statistics and methods of calculating p-values for several tests. Bootstrapping is introduced with the statistical computational methods of bootstrapped non-parametric tests. The next sections focus on the process of clustering which is mathematically defined and then demonstrating the importance of the dissimilarity measure within clustering and the interpretation of non-parametric tests defined in section 2.2 as measures of dissimilarity between objects. The concept of hierarchical clustering is introduced as a subgroup of hard clustering and the agglomerative algorithms are mathematically defined in terms of their relationship with the Lance-Williams formula (2.12). A graphical representation of hierarchical clustering is described and measures of comparing two hierarchical clusterings are discussed.

2.1 Introduction to Data

The methods of collection and quality control of the incubation period data are briefly summarized below as stated by Awofisayo-Okuyelu et al. (2017) in the systematic review and meta-analysis of campylobacteriosis. A systematic literature search for peer review publications was carried out for campylobacterosis, where each article went through an initial assessment stage to effectively evaluate the quality of the

incubation period data reported based on a set of criteria and not the quality of the overall study. With a predetermined format, raw and summary data were extracted along with summary statistics such as the mean,median, mode and range. The data then went through a final cleaning procedure before analysis.

The final result consists of 30 studies containing general information, study characteristics, organism characteristics, outcome measures and raw data. The incubation time for each study is reported which is the data for the clustering analysis in this dissertation. For further information on the criteria in the initial assessment stage and a detailed explanation of the data, refer to the search strategy and selection process in the meta-analysis.

2.2 Non-Parametric Tests

Non-parametric procedures are classified as distribution-free or inferencing with specified distributions and unspecified parameters. The mathematical procedure is a hypothesis test for the form of the population or for some characteristics of the probability distribution of the sample data. A common use for such tests is to obtain the characteristics of a population from which a sample is drawn, known as goodness-of-fit test for one-sample which can be generalized for two-samples. In the one sample case, characteristics of the population are obtained by testing for statistical evidence of how well the sample distribution represents the population. In the two-sample case, two samples are compared to identify whether both samples originate from the same population. Common non-parametric tests include the Kolmogorov-Smirnov, Anderson-Darling and Cramer-Von-Mises. These three non-parametric tests measure how close two empirical cumulative distribution functions have to be in order for two independent samples to originate from the same population. i.e they are not significantly different. If the null hypotheses (two samples come from the sample population) is not rejected, the two samples originate from the same population thus the empirical distribution functions of the two samples are reliable estimates of their population cumulative distribution function. Therefore if H_0 is accepted, the two empirical distributions should be similar Gibbons (2003).

The Kolmogorov-Smirnov test is the general standard for non-parametric tests when comparing two samples due to easy interpretation of the test statistic. The test exhibits increased power against deviations in the middle whereas the Anderson-Darling exhibits higher sensitivity to the tails and increased power to fatter tails than specified. The Cramer-von-mises is a combination of the Anderson-Darling and Kolmogorov-Smirnov test, attempting to balance the deviations in the middle and the sensitivity to tails but it tends to hold the properties the Kolmogorov-Smirnov over the Anderson-Darling.

2.2.1 Kolmogorov-Smirnov Test

As a Goodness-of-fit criterion, the KS test quantifies the distance between the empirical distribution function of a random sample with a hypothesized cumulative distribution. Adaptation to the two-sample case, compares two empirical distribution functions of two samples Gibbons (2003). Defining the empirical distribution function of the sample as $S_m(x)$, the proportion of sample observations that are less than or equal of x for all real numbers then $S_m(x)$ is seen as a consistent point estimator for the true distribution $F_x(x)$. Where $F_x(x)$ designates the unknown distribution functions of sample data.

For two independent random samples $X_1, X_2, \ldots, X_m$, $Y_1, Y_2, \ldots, Y_n$ from continuous populations $F_x(x)$ and $F_y(x)$ respectively and assuming the data are measured on at least an ordinal scale, the empirical distribution functions, denoted as S_m and S_n, are defined as

$$
\begin{aligned}
S_m(x) &= \frac{\text{number of observed } X\text{'s} \leq x}{m} \\
&= \begin{cases} 0 & \text{if } x < X_{(1)} \\ k/m & \text{if } X_{(k)} \leq x < X_{(k+1)} \quad \text{for } k = 1, 2, \ldots, m-1 \\ 1 & \text{if } x \geq X_{(m)} \end{cases}
\end{aligned}
$$

and

$$S_n(x) = \frac{\text{number of observed } Y\text{'s} \leq x}{n}$$
$$= \begin{cases} 0 & \text{if } x < Y_{(1)} \\ k/n & \text{if } Y_{(k)} \leq x < Y_{(k+1)} \quad \text{for } k = 1, 2, \ldots, n-1 \\ 1 & \text{if } x \geq Y_{(n)} \end{cases}$$

then the two-sided hypotheses for the Kolmogorov-Smirnov test are:

$H_0 : F_x(x) = F_y(x)$ for all x

$H_1 : F_x(x) \neq F_y(x)$ for at least one value of x

The test statistic of the Kolmogorov-Smirnov test, denoted by $D_{m,n}$ is the maximum absolute difference between two empirical distributions

$$D_{m,n} = \max_x |S_m(x) - S_n(x)|$$

Graphically, this is the maximum vertical distance between $S_m(x)$ and $S_n(x)$. Small values of $D_{m,n}$ suggest little difference between $S_m(x)$ and $S_n(x)$ implying there is not enough evidence to reject the null hypothesis. If the difference between $S_m(x)$ and $S_n(x)$ is large, then $D_{m,n}$ is big implying there is enough evidence to reject the null hypothesis, therefore the two samples do not originate from the same population.

2.2.2 Anderson-Darling Test

Among the empirical distribution function (EDF) statistics, the Anderson Darling statistic is known to be the most powerful (Arshad et al. 2003). It can be seen as a modification of the Kolmogorov-Sminov test by providing more weight to the tails (Farrell & Rogers-Stewart 2006). Therefore the procedure and methodology behind the Anderson-Darling test is very similar to that of the KS test but a difference in calculating the test statistic and p-values.

The two sample test statistic proposed by Darling (1957) for a random independent sample $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$, denoted by A^2_{mn}, is

$$A^2_{mn} = \frac{mn}{N} \int_{-\infty}^{\infty} \frac{\{F_m(x) - G_n(x)\}^2}{H_N(x)\{1 - H_N(x)\}} dH_N(x) \tag{2.1}$$

where $F_m(x)$ is defined as the proportion of the sample $X_1, X_2, \ldots, X_m$ that is not greater than x, $G_n(x)$ is the empirical distribution function of the second independent sample $Y_1, Y_2, \ldots, Y_n$ obtained from a continuous population with distribution function $G(x)$, and $H_N(x) = \{mF_m(x) + nG_n(x)\}1/N$, with $N = m + n$, is the empirical distribution function of the pooled sample. Then the hypotheses for the Anderson-Darling test are

$H_0 : F_m(x) = G_n(x)$ for all x

$H_1 : F_m(x) \neq G_n(x)$ for at least one value of x

Defining a computational formula is necessary as (2.1) is not practically applicable. By applying Kiefer (1959) treatment of the k-sample analogue of the Cramer-von Mises test, Scholz & Stephens (1987) defined the Anderson-Darling test statistic for the k-sample case to be

$$A^2_{kN} \sum_{i=1}^{k} n_i \int_{B_N} \frac{\{F_{in_i(x)} - H_N(x)\}^2}{H_N(x)\{1 - H_N(x)\}} dH_N(x) \tag{2.2}$$

where $B_N = \{x \in R : H_N(x) < 1\}$. For the case of $k = 2$, (2.2) reduces to (2.1) and under the continuity assumption on the F_i, the probability of ties is zero. The pooled ordered sample is then $Z_1 < Z_2 < \ldots Z_N$ which results in the computational formula

$$AD = A^2_{kN} = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{(NM_{ij} - jn_i)^2}{j(N - j)} \tag{2.3}$$

where M_{ij} is the number of observations in the ith sample that are not greater than Z_j, which is the jth observation in the pooled ordered sample Scholz & Stephens (1987).

2.2.3 Cramer-von Mises Criterion

The Cramer-von Mises criterion is another measure of difference between two empirical distribution functions and is similar to the Kolmogorov-Smirnov test and Anderson-Darling in this nature. The hypotheses are

$H_0 : F_x(x) = G_x(x)$ for all x from $-\infty$ to ∞

$H_1 : F_x(x) \neq G_x(x)$ for at least one value of x

for two independent samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$, with unknown empirical distribution functions $F(x)$ and $G(x)$, respectively. The Cramer-von Mises statistic can be defined as

$$T_2 = \frac{mn}{(m+n)^2} \sum_{x=X_i} \sum_{X=Y_j} [S_1(x) - S_2(x)]^2 \tag{2.4}$$

$$= \frac{mn}{(m+n)^2} \left\{ \sum_{i=1}^{n} [S_1(X_i) - S_2(X_i)]^2 + \sum_{j=1}^{m} [S_1(Y_j) - S_2(Y_j)]^2 \right\} \tag{2.5}$$

where $S_1(x)$ and $S_2(x)$ are the empirical distribution functions of the two samples and the squared difference in the summation is computed at each X_i and each Y_j (Conover 1999). Another version has been proposed by Baringhaus & Franz (2004) called the Cramer test which is seen to be an improvement of the original Cramer-von mises criterion and is defined as

$$T_{m,n} = \frac{mn}{m+n} \int_{-\infty}^{\infty} [F_m(t) - G_n(t)]^2 dt, \tag{2.6}$$

where $F_m(t)$ and $G_m(t)$ are the empirical distribution functions of $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ respectively.

2.2.4 Bootstrapping of non-parametric tests

Non-parametric tests compute the exact probability of obtaining observed results under H_0, which is suitable for small sample size studies. Another method calculates p-values based on asymptotic properties which is suitable for large sample size, referred

to as asymptotic nonparametric tests. However standard non-parametric tests do not perform well with small sample size studies in many occasions, therefore permutation tests or bootstrap tests can be combined with non-parametric tests for accommodation of accurate estimation for small sample size studies. In the bootstrap procedure, the test statistic values for a non-parametric test are obtained by resampling, with replacement, the bootstrap test under the null hypothesis.

In non-parametric bootstrap tests, it is assumed the sample represents the empirical distribution of the population function and a large number of bootstrap samples are drawn from this distribution for constructing a sampling distribution of a test statistic, known as a bootstrap sample distribution. The p-values are obtained by locating the observed statistic from the observed sample on the bootstrap distribution. This procedure is heavily dependent on the non-parametric test statistic (Dwivedi et al. 2017). The following algorithm proposed by (Efron 1993) is described as:

1. Let $x = x_1, x_2, \ldots, x_m$ be the observed sample 1 of size, and $y = y_1, y_2, \ldots, y_m$.
2. Evaluate the non-parametric test statistic T_{Obs}. For the Kolmogorov-Smirnov test, the r function *ks.boot()*; Anderson darling, the r function (kSamples::ad.test()) and for the Cramer test, the r function CvM.test() calculates each tests bootstrapped statistic.
3. Create two transformed error data sets, $x^* = x_1 - \bar{x} + \bar{z}$, $x_2 - \bar{(x)} - \bar{(z)}, \ldots, x_m - \bar{x} - \bar{z}$ and $y^* = y_1 - \bar{y} + \bar{z}$, $y_2 - \bar{(y)} - \bar{(z)}, \ldots, y_n - \bar{y} - \bar{z}$, where $\bar{z}$ is the mean of the combined sample.
4. Draw a bootstrap sample of size m observations with replacement ($x^{*\prime}$) and of size n observations with replace ($y^{*\prime}$) and compute the necessary statistics for each test statistic.
5. Repeat steps 3 and 4 N times to obtain N test statistics
6. Approximate p-value = $\frac{\text{number of times } t^{*\prime} \geq T_{Obs}}{N}$

2.3 Clustering

Clustering, also referred to as classification, is a method of unsupervised learning described as dividing a set of objects into a smaller number of clusters in such a way that objects in the same cluster are similar to one another and dissimilar to objects in other clusters (Gordon 1987).

Figure 1.1 shows that clustering can be divided into Hard and Fuzzy clustering. In hard clustering each object is assumed to belong to one and only one cluster. Mathematically a class label $l_i \in \{1, 2, \ldots, k\}$ is assigned to each object x_i to identify the cluster's class, where k is the number of clusters. The result is expressed as a hard k-partition of a data set D in a $k \times n$ matrix

$$\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k1} & u_{k2} & \cdots & u_{kn} \end{pmatrix} \tag{2.7}$$

where n denotes the number of records in the data set, k denotes the number of clusters, and u_{ij} satisfies the following constraints

$$u_{ji} \in \{0, 1\}, \quad 1 \leq j \leq k, \quad 1 \leq i \leq n, \tag{2.8}$$

$$\sum_{j=1}^{k} u_{ji} = 1, \quad 1 \leq i \leq n, \tag{2.9}$$

$$\sum_{i=1}^{n} u_{ji} > 0, \quad 1 \leq j \leq k. \tag{2.10}$$

Constraint (2.8) implies an object can either belong to one cluster or not, constraint (2.9) indicates each object belongs to one and only one cluster and constraint (2.10) suggests a cluster must not be empty. Therefore a cluster is valid if and only if there is exactly one object allocated which has not been allocated to any other cluster. Fuzzy clustering is similar but with relaxed assumptions, for example it is not necessary for a cluster to only allocate one object inside. Mathematically the clustering process of a data set D can be represented by an assignment function $f : D \to [0, 1]^k, x \to f(x)$,

defined as

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_k(x) \end{pmatrix} \tag{2.11}$$

where $f_i(x) \in [0, 1]$ for $i = 1, 2, \ldots, k$ and $x \in D$, and

$$\sum_{i=1}^{k} f_i(x)1 \quad \forall x \in D$$

Then for every $x \in D$, $f_i(x) \in \{0, 1\}$ the clustering represented by f is hard clustering (Gan 2007).

2.3.1 Non-Parametric Tests as a Measure of Dissimilarity

Distances and similarity measures play an important role in cluster analysis (Jain 1988, Anderberg 1973) where they quantitatively describe the similarity or dissimilarity of two objects or clusters. Data to be analysed with clustering is commonly presented as an $n \times p$ raw data matrix, $X \equiv (x_{ik})$, where x_{ik} denotes the value of the kth variable observed for ith object or a $n \times n$ matrix of pairwise dissimilarities $D \equiv (d_{ij})$, where d_{ij} denotes the dissimilarity between the ith and jth objects (Gordon 1987). In the majority of clustering methods, the raw data matrix is transformed into a pairwise dissimilarity matrix and occasionally the data occurs naturally in dissimilarity format for which many different measures of dissimilarity have been proposed to construct a relevant measure of pairwise dissimilarity mentioned in the literature review of clustering. In order for a dissimilarity matrix or raw data to be considered a dissimilarity measure, the matrix must satisfy the following minimum conditions

1. $d_{ij} \leq 0, d_{ii} = 0$
2. $d_{ij} = d_{ji}$ for all i, j belonging to the set of objects, Ω

Data in an asymmetric difference matrix, d^*_{ij}, can be transformed into a pairwise distance matrix by $d_{ij} = \frac{1}{2}(d^*_{ij} + d^*_{ji})$ or other generalizations of clustering strategies (Gordon 1987).

Awofisayo-Okuyelu et al. (2017) developed a new methodology, characterizing the Kolmogorov-Smirnov test as a measure of dissimilarity between the studies. The methodology behind this relies on the definition of p-values and the transformation of a distance matrix into a pairwise dissimilarity matrix. P-values are transformations of a test statistic into a standard form hence they are random variables that are uniformly distributed when the null hypothesis is true and all other assumptions are met (Murdoch et al. 2008). This is a direct result of the definition of α as the probability of a type I error. It is desirable to result in a probability of rejecting a true null hypothesis of α which the distribution of the p-value is more weighted towards zero.

Due to P-values representing a standard metric, the test statistic is not interpreted as the measure of distance as different non-parametric tests have different levels of measures for their test statistics. Therefore the transformation of a test statistic into a p-value allows for the comparison of non-parametric tests. This will be the basis for the measure of dissimilarity. Firstly, the bootstrapped versions of each test mentioned in sections above will be applied to individual studies against each other and recorded. Then the dissimilarity matrix is calculated by subtracting 1 from the p-values and transforming the resulting matrix by comparing the distances between p-values. Then the hierarchical clustering algorithms, defined in section (2.3.2) are applied to the generated dissimilarity matrices.

2.3.2 Hierarchical Clustering and Agglomerative Algorithms

Hard clustering can be further split into hierarchical and partitional clustering. The main difference between partitional and hierarchical clustering is how the data sets are divided. Partitional clustering divides a data set into a single partition which creates a one-level non-overlapping partitioning of the data point whereas hierarchical clustering divides a data set into a sequence of nested partitions. Hierarchical clustering is of particular interest when investigating how clusters of similar objects are related to one another.

The Central Agglomerative procedure can be seen as the main algorithm for hierarchical clustering and all other linkage methods described later are variations of this. The algorithm is stated in Chapter 6 of Anderberg (1973) as the following:

Let s_{ij} be the similarity between i and j by a measure of association. Assuming the similarities are symmetric i.e. $(s_{ij} = s_{ji})$ and non-negative then all $\binom{n}{2} = \frac{1}{2}n(n-1)$ possible pairwise combinations may be arrayed in a lower triangular similarity matrix. Then the general procedure for agglomerative clustering is

1. Initialise with n clusters with exactly one object allocated to each cluster and label the clusters $1, 2, \ldots, n$.

2. Label the most similar pair of clusters p and q and their associated similarity S_{pq}, for $p > q$. The general method for calculating this distance if p and q are represented as C_i and C_j then for some cluster C_k:

$$\begin{aligned} d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) \\ +\gamma |d(C_i, C_k) - d(C_j, C_k)| + \delta_i h(C_i) + \delta_j h(C_j) \\ +\epsilon h(C_k) \end{aligned} \tag{2.12}$$

   is known as the Lance-williams formula, where $h(C_i)$ is the height of cluster C_i and $\theta \equiv (\alpha_i, \alpha_j, \beta, \gamma, \delta_i, \delta_j, \epsilon)$ is a set of parameters whose values specify the clustering strategy in Table 2.1 (G. N. Lance & W. T. Williams 1966, N. Lance & T. A. Williams 1967).

3. Reduce the number of clusters by 1 and label the product of the merger q and update the similarity matrix objects to reflect the updated similarities between cluster q and all other existing clusters. Delete the row and column of S which represents cluster p.

4. Repeat steps 2 and 3, $n - 1$ times and at each stage record the identity of the clusters which are merged and the value of the similarity between.

For all agglomerative algorithms, step 1 is the same. In step 2, algorithms measure the similarity between two pairs differently, which depends on their definition and method of calculating the distance between the two pairs of objects and for step 3, algorithms propose varying methods to update the revised similarity matrix.

The three linkage algorithms which are most commonly used in agglomerative hierarchical clustering are the single-link (Sneath 1957), complete-link (Mcquitty 1960) and average-link (Sokal & Michener 1958, Mcquitty 1967). Each provides a different method of weighting the similarity between two clusters, shown in Table 2.1 for the weight w_i associated with each cluster C_i defined to be the number of objects the cluster contains.

Table 2.1: *Clustering strategies from the general agglomerative algorithm*

Name	α_i	β	γ	δ_i	ϵ
Single-link	$\frac{1}{2}$	0	- $\frac{1}{2}$	0	0
Complete-link	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
Average-link	$\frac{w_i}{w_i+w_j}$	0	0	0	0

Single-Linkage Algorithm

The single linkage algorithms, commonly referred to as nearest neighbor method and minimum method, clusters objects together at each stage by joining the single shortest or strongest link between them to measure the dissimilarity between two groups. The distance for three objects, C_i, C_j and C_k, between C_k and $C_i \cup C_j$ can be obtained by substituting the parameters in Table 2.1 into the Lance-Williams formula (2.12) to obtain the distance-like measure

$$\begin{aligned} & D(C_k, C_i \cup C_j) \\ & = \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) - \frac{1}{2}|D(C_k, C_i) - D(C_k, C_j)| \\ & = \min\{D(C_k, C_i), D(C_k, C_j)\} \end{aligned} \tag{2.13}$$

where $D(\cdot, \cdot)$ is a distance between two clusters. From (2.13) it can be seen that

$$D(C, C') = \min_{x \in C, y \in C'} d(x, y) \tag{2.14}$$

where C and C' are two non-empty, non-overlapping clusters and $d(\cdot, \cdot)$ is the distance function to compute the dissimilarity matrix.

The single-linkage algorithm is known to be incapable of distinguishing poorly separated clusters. For example if the two objects in two separate clusters are closer to each other than objects within their respective cluster then the single-link algorithm clusters the objects together. However for distant clusters, the single-link algorithm will cluster the data appropriately. Therefore, for any cluster of two or more objects, every object is more similar to some other object of the same cluster than any other object not in the same cluster. The main advantage single-link clustering has over complete and average-link is that it's invariant to any transformation which leaves the ordering of the similarities unchanged. Johnson (1967) and Jardein *et al.* (1967) comment on the theoretical observations of this. Often criticised, is the tendency of single-link clustering to chain clusters together. Objects at opposite ends of a cluster may be marked dissimilar but in fact they are not. This is due to the method of using the minimum distance between two objects as a measure of dissimilarity between clusters.

Complete-Linkage Algorithm

The complete-linkage algorithm, commonly referred to as farthest neighbor method, links objects in a cluster to each other at some maximum distance or minimum similarity. In comparison to the single-link method, the interpretation of clusters can only be analysed within individual clusters as there is no useful interpretation for comparing different clusters. Using the same notation as before and substituting the parameter

values in Table 2.1 into the Lance-Williams formula to obtain the distance-like measure

$$\begin{aligned} & D(C_k, C_i \cup C_j) \\ & = D(C_k, C_i) + \frac{1}{2} D(C_k, C_j) + \frac{1}{2} |D(C_k, C_i) - D(C_k, C_j)| \\ & = \max\{D(C_k, C_i), D(C_k, C_j)\} \end{aligned} \tag{2.15}$$

From 2.15 it can be seen that

$$D(C, C') = \max_{x \in C, y \in C'} d(x, y) \tag{2.16}$$

where $D(C, C')$ is the distance between the most distant members of the groups C_k and $C_i \cup C_j$. $D(C, C')$ can be interpreted graphically as the diameter of the smallest sphere which encloses the cluster resulting from $C_i \cup C_j$. Hence the complete-link algorithm searches for the furthest distance between pairs of objects contrary to the single-link which uses the minimum distance between pairs. Although this solves the issue of chaining, another issue arises. Outlying objects prevent close clusters to merge together because the furthest distance measure amplifies the effects of outlying data (Odilia Yim & Kylee T. Ramdeen 2015). This results in clustering of objects to be more conservative which has the opposite effect of single-linkage.

Average-link Algorithm

The Group Average method, commonly referred to as UPGMA which stands for "unweighted pair group method using arithmetic averages" Jain (1988). The distance between two groups is defined as the average of the distances between all possible points that are made up of one data point from each group (Gan 2007). Using the same notation as before and substituting the parameter values in Table 2.1 into the Lance-Williams formula to obtain the distance-like measure

$$\begin{aligned} & D(C_k, C_i \cup C_j) \\ & = \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j) \end{aligned} \tag{2.17}$$

From 2.17 it can be seen that

$$D(C, C') = \frac{1}{|C||C'|} \sum_{x \in C, y \in C'} d(x, y), \tag{2.18}$$

where the proof is shown in the appendix. This method is seemed to be an in-between of two extremes as it's definition is not dependent on extreme values therefore there is no interpretation of minimum or maximum similarity within a cluster.

Summary of linkage-Algorithms

Currently, there is no universal rule or criteria which suggests a linkage algorithm is the best in general. Each algorithm has been developed for a specific reason and works best in certain circumstances. For the incubation period data analysed in this project, all three algorithms will be implemented to compare which algorithm represents the data the best and to test for their uncertainties to see whether they achieve similar or different results.

The Complete-algorithm will be used as a basis for when testing for other uncertainties as the algorithm tends to retain all the information of the pairwise distances. Comparing both the drawbacks of single and complete-linkage it can be seen that complete-linkage is generally seen as the most reliable method as it merges outliers later on the process due to outliers increasing maximum distances hence it is more robust than single-link. The general consensus behind the average-link method is when the single or complete-link methods are not applicable or cannot be distinguished in terms of both their disadvantages applying to the data. Since the average-link can be seen as the in-between measure of single and complete-linkage, it is the safer option.

2.3.3 Representations of Hierarchical Clustering

Hierarchical clustering can be represented either graphically or as a list of abstract symbols. The former is the popular choice due to its easy interpretation of which objects are clustered together. The most common graphical visualisation of hierarchical

clustering is n-tree, for other representations refer to Sneath (1973). An n-tree on a set of objects $\Omega = \{1, 2, \ldots, n\}$ is a set of T of subsets Ω satisfying the conditions

1. $\Omega \in T, \varnothing \notin T, \{i\} \in T$ for all $i \in \Omega$,
2. $A \cap B \in \{\varnothing, A, B\}$ for all $A, B \in T$.

(Bobisud & Bobisud 1972)

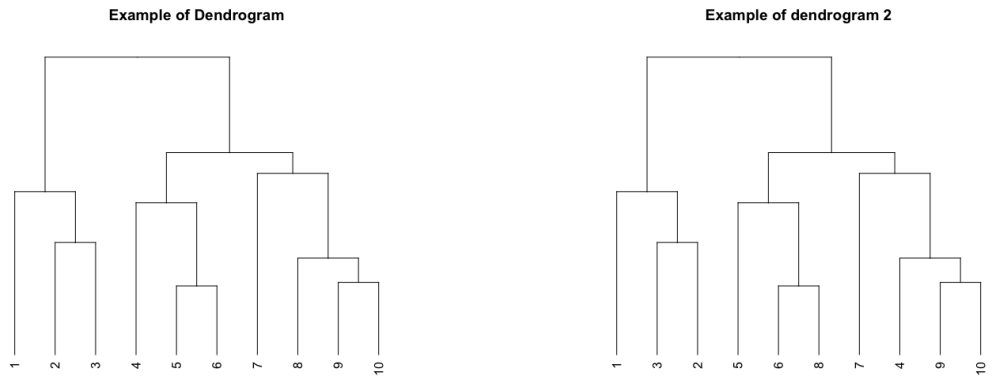
Therefore an n-tree is defined as an unordered rooted tree with labeled leaves, in which no internal node has degree two except the root of the tree which is the node at the top. A dendrogram can then be described as an n-tree diagram where each internal node is associated with a height h satisfying the condition

$$h(a) \leq h(B) \leftrightarrow A \subseteq B$$

for all subset of data points A and B if $A \cap B \neq \varnothing$, where $h(A)$ and $h(B)$ denote the heights of A and B respectively. Let h_{ij} be the height of the internal note specifying the smallest class to which objects x_i and x_j belong to then h_{ij} must satisfy the ultrametric condition

$$h_{ij} \leq \max\{h_i k, h_j k\} \quad \forall i, j, k \in \Omega$$

which is a necessary and sufficient condition for a dendrogram Gordon (1987). An example of two simple dendrogram is shown in Figure 2.1 where 10 objects are clustered together with the complete-linkage algorithm.



(a) *Example 1* **(b)** *Example 2*

Figure 2.1: *Example of two dendrograms*

There is no defined method of referring to an object within a cluster or any notation other than stating the objects within clusters. Therefore when referring to future dendrograms the notation is as follows: The lowest level of clustering, i.e. pairing of two objects is denoted as $\cap$, $|$ represents two clusters joining together within a big cluster and $||$ represents a cluster which contains two clusters paired with another cluster, and $|||$ represents the highest level of cutting the dendrogram which is the top of the dendrogram. For example, the dendrograms in Figure 2.1 **a** and **b** are labeled as

$$1|2 \cap 3|||4|5 \cap 6||7|8|9 \cap 10$$

and

$$1|3 \cap 2|||5|6 \cap 8||7|4|9 \cap 10$$

respectively

This implies 2 and 3, 5 and 6, 9 and 10 are paired together in the smallest level of cluster. 1 is paired together with cluster 2 and 3, 3 is paired together with cluster 5 and 6, 8 is paired together with cluster 9 and 10 and 7 is paired together with cluster 8,9,10. This notation is particularly useful when it is not possible to present all the dendrograms in a study but to just refer to each dendrogram in terms of the structure.

2.4 Comparison of two Hierarchical Clusters

Mentioned in Section 1.3.1, several methods are proposed to compare two hierarchical clusters both in terms of dendrograms and the clustering process. The package 'dendextend' (Galili 2015) implements the correlation methods and provides several graphical techniques to visually compare two dendrograms. Thus comparing two sets of hierarchical clustering or dendrograms can be split into correlation measures and graphical methods.

2.4.1 Correlation Measures

Correlation measures are measures of similarity between two dendrograms of hierarchical clustering. When an agglomerative hierarchical algorithm is applied to a dissimilarity matrix, the ranks orders corresponding to the rank of ordering clusters together is a numerical value of the iteration at which two objects are merged. Comparison of these rank order's is not possible with standard measures of association such as the product moment correlation coefficient due to its inapplicability with ordinal data. Therefore the three measures that were proposed to find the association between ordinal data are the cophenetic correlation and the Baker's Gamma index. These values range from -1 to 1, where near 0 values indicating two dendrograms are not statistically similar and a value of 1 would indicate strong statistical evidence of similarity between two dendrograms.

There are many other methods in which dendrograms can be compared with statistical software. Global comparisons are possible between the merging of clusters, height of clusters and ordering of clusters between dendrograms. The merging displays the ordering of pairing of objects within a cluster which is able to provide information to where exactly the dendrograms changed in terms of pairing of clusters. The height provides information to how different the cophenetic distances are between clusters and the ordering suggests how similar two dendrograms are in terms of the orders of pairings, small and big clusters where a value of 0 indicates identical ordered dendrograms.

Cophenetic Correlation

The cophenetic correlation is derived from the cophenetic distance which is defined to be the distance between the largest two clusters that contain only one object within when they are merged into a single cluster that contains both. The cophenetic correlation is then seen as a measure of how well a dendrogram preserves the pairwise distance dissimilarity matrix. This interpretation implies it is used to detect whether a dendrograms represent the original data but it can be further applied to see whether two dendrograms are similar by the correlation between two cophenetic distance matrices of two trees and is calculated by the following:

For a dendrogram $\{T_i\}$ and data $\{X_i\}$, where $x(i,j)$ is the distance between two data points and $t(i,j)$ is the dendrogrammatic distance between T_i and T_j and let x be the average of $x(i,j)$ and t the average of $t(i,j)$, the cophenetic correlation, c is defined as

$$c = \frac{\sum_{i<j}(x(i,j)-x)(t(i,j)-t)}{\sqrt{\left[\sum_{i<j}(x(i,j)-x)^2\right]\left[\sum_{i<j}(t(i,j)-t)^2\right]}}$$

Baker's Gamma Index

The Baker's gamma index is characterised as the rank correlation between the stages at which pairs of objects combine in each of the two clusters. Calculation of the index involves taking two objects and determining the highest level of k (number of cluster groups created when cutting the cluster) for which the two objects remain in the same cluster. The value of k is recorded and is repeated for the same cluster. This is then repeated for $\frac{n}{2}$ combinations and both values are calculated from each of the two clusters. These two sets of numbers are paired according to the pairing of objects compared and the Spearman correlation is calculated (Galili 2015).

2.4.2 Tanglegram

A tanglegram is a graphical representation of comparing two dendrograms side by side by assessing the position of each object within a cluster. A straight line indicates the object is in the same position and if two objects within a small cluster paired alongside another two objects within a small cluster, results in straight lines it indicates the big cluster, small cluster and pairings within the small clusters have not changed position.

Visually inspecting a tanglegram is useful compared to other representations of hierarchical structures due to their easiness of assessing what objects have changed position between two dendrograms. For example the global comparisons of merging between two pairs of objects will provide an ordered list of the merging procedure of the dendrogram. One would inspect the list for both dendrograms and look for discrepancies. With a tanglegram, one can easily detect if clusters or pairings have changed procedures by simply assessing for lines which are not parallel. An example of a tanglegram between the dendrogram examples in Figure 2.1 is shown in Figure 2.2.

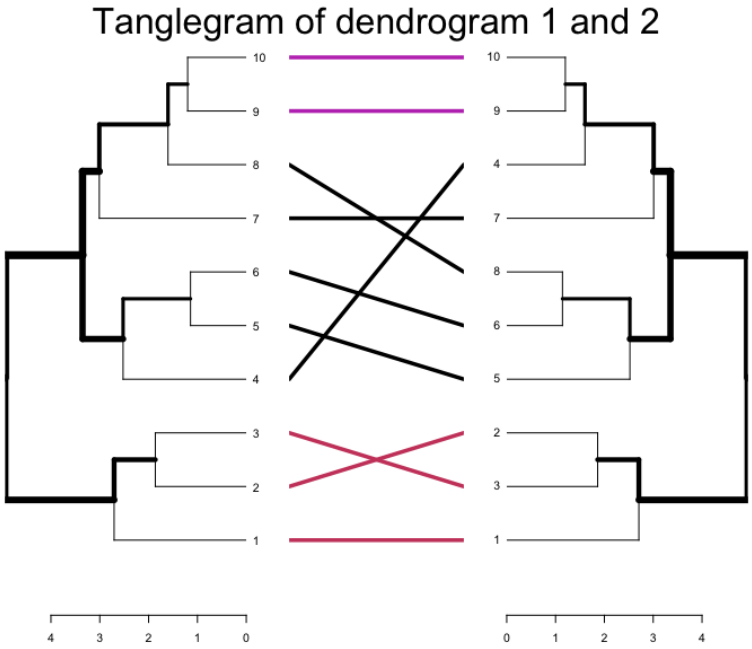


Figure 2.2: *Tanglegram between dendrogram 1 and dendrogram 2*

2.5 Analysis

The methodologies proposed in this chapter will be carried out on R with the incubation period data. The specific packages and procedure of assessing for validity and uncertainty of hierarchical structures is explained below.

2.5.1 Hierarchical Clustering in R

The hierarchical procedure for agglomerative algorithms is implemented in the function hclust() by providing a predefined dissimilarity distance object with traditionally agglomerative linkage methods. An advantage of hclust, compared to other agglomerative hierarchical functions is hclust provides an option for any dissimilarity distance whereas a function such as agnes only allows for traditional distance methods.

2.5.2 Validation of Hierarchical clustering

For each non-parametric test, an initial assessment is carried out for a base dendrogram. i.e 10,000 bootstrap samples with the complete linkage algorithm, in terms of how well the clusters represent the original distance matrix. This is measured by computing the correlation between the cophenetic distances and the original dissimilarity distance object generated by the bootstrapped version of each non-parametric test's p-values between studies. If the clustering is valid, the linking of objects in clusters should have a strong correlation with distances between objects in the original data matrix (Kassambara 2017). This method is applicable to validate hierarchical clustering structures and compare linkages for the same data set. The linkage algorithm with the highest cophenetic correlation is the best representation of the data.

2.5.3 Bootstrap Uncertainties

To understand the effect of bootstrapping with a non-parametric test in hierarchical clustering, 100 dendrograms are simulated for each non-parametric test with 10,000;

1,000 and 100 bootstrap samples. Firstly the variation within each sample size is visually inspected by a tanglegram in terms of how a dendrogram, repeated x number of times is differs to a dendrogram repeated y number of times. Tanglegrams will show whether the pairings have altered, or small/big clusters have changed orders. When clusters have changed orders, this can be seen as a variation of a dendrogram which is regarded as a small uncertainty as agglomerative algorithms tend to change the position of big clusters. If the pairings have changed then further analysis is required to quantify how the studies changed effect the clustering.

The groups (number of similar dendrograms) in each bootstrap sample is compared to assess whether dendrograms are replicated in different sample sizes. If the two identical dendrograms in different bootstrap sample sizes are replicated, then how many times are each one replicated within their respective sample size? If the number of each dendrogram in the same then there is little uncertainty between sample size. Another method to assess the differences is to compare how many different dendrograms are produced in each sample size. If the number is roughly the same then there is little uncertainty.

2.5.4 Assessing for Uncertainty within Studies

This is the most complicated analysis as it requires both visual inspection and very subjective analysis. When comparing for uncertainties within bootstrap samples, the differences are analysed in terms of which studies change pairings. If a study changes pairings then most likely it has some effect on the clustering. Thus by removing the study in question, the dendrogram can be reproduced to assess the impact of this study. If the uncertainty behind the individual study is removed and the dendrograms produced result in a higher cophenetic correlation, then the clustering algorithms are affected by this study either positively or negatively depending on the change. Finally, the study with the highest sample size will be removed for each test in the complete-linkage algorithm to assess how sample size impacts the hierarchical clustering.

Chapter 3

Results & Discussion

The results section is split in to 3 subsections. Firstly, the effect of the bootstrap sample size for each non-parametric test is visualized. The second subsection assesses the effect of removing studies which effected the dendrograms in the bootstrapping. Finally, the validity of each linkage algorithm for 10,000 bootstrap samples of each test is assessed.

3.1 Bootstrap Uncertainty

For each test and number of bootstrap samples, the number of similar dendrograms from 100 simulations were recorded and assigned into groups where the first group (Group 1) contains the highest number of dendrograms and the last group, the least. The number of different dendrograms were also recorded. The results are summarised in Table 3.1.

Table 3.1: *Uncertainty of Bootstrap samples for non-parametric tests*

	Number of Groups	Number of dendrograms in Group 1	Number of different dendrograms
Kolmogorov-Smirnov			
10,000	4	63	5
1,000	14	13	38
100	8	3	83
Anderson-Darling			
10,000	4	56	3
1,000	19	13	26
100	7	4	82
Cramer-von			
10,000	13	22	7
1,000	20	8	40
100	4	2	92

3.1.1 Uncertainty within 10,000 Bootstrap Samples

For each non-parametric test, the group analysis in Table 3.1 is further analysed in terms of how many dendrograms are in each group which is given below:

- Kolmogorov-Smirnov test: 63 (Group 1), 15 (Group 2) and 6 (Group 3 & 4).
- Anderson-Darling test: 56 (Group 1), 25 (Group 2),7 (Group 3) and 4 (Group 4).
- Cramer-von test: 22 (Group 1 & 2), 10 (Group 3), 7 (Group 4), 6 (Group 5 & 6), 5 (Group 7), 4 (Group 8), 3 (Group 9), 2 (Groups 10-13).

Comparing the number of similar dendrograms in group 1, the Kolmogorov-Smirnov

contains the highest, closely followed by the Anderson-Darling and then the Cramer-von. The difference between the Kolmogorov-Smirnov and Anderson-Darling is only 7, whereas the Cramer-von test contains much less dendrograms in Group 1. The number of groups within the Kolmogorov-Smirnov and Anderson-Darling are the same, but for the Cramer-von, there are more than 3 times as many groups. For the number of dendrograms left which are not identically ordered, all tests perform similarly. This suggests the Kolmogorov-Smirnov and Anderson-Darling have similar power levels to replicate the most common dendrogram, whereas the Cramer-von test performs poorly with much higher variation in all aspects apart from the number of different dendrograms. The uncertainty between groups for each test can be explained by visually inspecting the differences between dendrograms in different groups for each test. The variations between each group is shown in Table 3.2 by comparing a dendrogram in Group 1 to all other groups for each test (refer to Figure 2.1)

Table 3.2: *Visual inspection of pairing variation for 10,000 bootstrap samples*

Kolmogorov-Smirnov	
Group 3	$53 \cap 68$ & $25b \vert 4 \cap 9 \longrightarrow 53 \cap 25b \vert 4 \cap 9$ & 68
Anderson-Darling	
Groups 3 & 4	$14 \vert 25a \cap 40$ & $37 \vert 51 \cap 29 \longrightarrow 37 \cap 25a \vert 51 \cap 29$ & $40 \cap 14$
Cramer-von	
Groups 3,6,7,11 & 12	$56 \cap 25b$ & $34 \vert (36 \cap 32) \longrightarrow 25b \vert 4 \cap 9$ & $56 \vert (34 \vert (32 \cap 36))$

For the variation of pairings in the: Kolmogorov-Smirnov test, the studies which affect the clustering are studies 25b, 53 and study 68; Anderson-Darling test, study 14, 25a, 37 and 40 affect the clustering; Cramer-von test, study 25b and 56 are split and merged into separate clusters. For each type of variation, there is low impact to the clustering structure, only small changes of pairings. The dendrograms for each type of variation is shown in Figures B.1, B.2, B.3 for the Kolmogorov-Smirnov, Anderson Darling and Cramer tests respectively and the tanglegrams in Figure B.4 in the appendix.

3.1.2 Uncertainty within 1,000 Bootstrap Samples

For each non-parametric test, the group analysis in Table 3.1 is further analysed in terms of how many dendrograms are in each group which is given below:

- Kolmogorov-Smirnov test: 13 (Group 1), 12 (Group 2), 5 (Group 3), 4 (Groups 4-6), 3 (Groups 7-10), 2 (Groups 11-14).
- Anderson-Darling test: 13 (Group 1), 12 (Group 2), 7 (Group 3), 5 (Group 4), 4 (Groups 5 & 6), 3 (Groups 7-9) and 2 (Groups 10-19).
- Cramer-von test: 8 (Group 1), 5 (Group 2), 4 (Group 3 & 4), 3 (Groups 5-11), 2 (Groups 12-20).

Comparing the number of similar dendrograms in group 1, the Kolmogorov-Smirnov and Anderson-Darling tests produce 13 dendrograms whereas the Cramer-von test produces only 8 dendrograms. This difference is seen again for the number of dendrograms in group 2 for each test. In terms of number of different dendrograms, the Anderson-Darling performs best with similar results for the Kolmogorov-Smirnov and Cramer-von tests. This suggests for 1,000 bootstrap samples, the Anderson-Darling test performs slightly better than the Kolmogorov-Smirnov test and the Cramer-von test is again, the test with the highest uncertainty. The visual inspections of differences between dendrograms in different groups for each test are shown in Table 3.3.

Table 3.3: *Visual inspection of pairing variation for 1,000 bootstrap samples*

Kolmogorov-Smirnov	
Groups 3,4,7,11	$53 \cap 68$ & $25b\|4 \cap 9 \longrightarrow 53 \cap 25b\|4 \cap 9$ & 68
Anderson-Darling	
Groups 2,5-8,10,13,15	$37 \cap 25a$ & $40 \cap 14 \longrightarrow 37\|29 \cap 51$ & $14\|40 \cap 25a$
Group 11	$37 \cap 25a$ & $40 \cap 14 \longrightarrow 37\|12\|30 \cap 18$ & $14\|40 \cap 25a$
Groups 12,19	68 & $67\|53 \cap 4\|25b \cap 9$ & $37 \cap 25a$ & $40 \cap 14 \longrightarrow$ $68 \cap 53$ & $25b\|9 \cap 4\|\|67$ & $37\|\|12\|30 \cap 18$ & $14\|40 \cap 25a$
Groups 16,17,18	$67\|53 \cap 4\|\|25b \cap 9 \longrightarrow 53 \cap 68$ & $67\|\|25b\|4 \cap 9$

Cramer-von	
Groups 2,5,7,8,14,15,18,19	$68\ \&\ 67 \cap 53 \longrightarrow 53 \cap 68\ \&\ 67 \vert 4 \cap 9$
Groups 3,6,11,16,20	$56 \cap 25b \longrightarrow 53\ \&\ 25b \vert 9 \cap 4$

For the variation of pairings in the: Kolmogorov-Smirnov test, the studies which affect the clustering are studies 25b, 53 and study 68; Anderson-Darling test, studies 14, 25a, 25b 37, 40, 68, 67 affect the clustering; Cramer-von test, studies 53,67, 25b and 56 affect the clustering. For the Kolmogorov-Smirnov test, the same variation occurs for the 10,000 sample size case therefore no further analysis is required. For the Anderson-Darling test there are several studies which change pairings for different groups. This required further analysis by removing each study and assessing the uncertainty between the original dendrogram and resulting dendrogram (with one less study). For the Cramer-von test, the variation of pairings is similar to the 10,000 case but with the extra uncertainty of study 68 and 53. The dendrograms for each type of variation is shown in Figures B.5, B.6, B.7 for the Kolmogorov-Smirnov, Anderson Darling and Cramer tests respectively and the tanglegrams in Figure B.7 in the appendix.

3.1.3 Uncertainty within 100 Bootstrap Samples

For each non-parametric test, the group analysis in Table 3.1 is further analysed in terms of how many dendrograms are in each group which is given below:

- Kolmogorov-Smirnov test: 3 (Group 1), 2 (Groups 2-8).
- Anderson-Darling test: 4 (Group 1), 3 (Group 2 & 3) and 2 (Groups 4-7).
- Cramer-von test: 2 (Groups 1-4).

As expected, the number of similar dendrograms within each group has dramatically decreased for each test and the number of groups has decreased too. In comparing tests for 100 bootstrap samples, the Anderson-Darling and Kolmogorov-Smirnov tests seem to perform similarly because of the high uncertainty, the number of groups can assess how well this number of bootstrap samples performs within each test. The Cramer-von test performs the worst with only 4 groups and no similar dendrograms

in a group above 2. The visual inspections of differences between dendrograms in different groups for each test are shown in Table 3.4.

Table 3.4: *Visual inspection of pairing variation for 100 bootstrap samples*

Kolmogorov-Smirnov	
Group 3	$37 \cap 25a \longrightarrow 37 \vert 30 \cap 18$ & $25a \vert 29 \cap 51$
Anderson-Darling	
Group 2	$68 \cap 53$ & $67 \vert\vert 25b \vert 9 \cap 4 \longrightarrow 68$ & $67 \cap 53 \vert\vert 25b \vert 9 \cap 4$
Group 3	$14 \cap 40$ & $37 \cap 25a \longrightarrow 14 \vert 25a \cap 40$ & $37 \vert\vert 12 \vert 30 \cap 18$
Cramer-von	Too many pairing changes.

For the variation of pairings in the: Kolmogorov-Smirnov test, the studies which affect the clustering are studies 37 and 25a; Anderson-Darling test, studies 53,67; and for the Cramer-von test, there are too many discrepancies of pairings, therefore the studies may not be responsible for this but the actual uncertainties within the Cramer test and the number of bootstrap samples could be. The dendrograms for each type of variation is shown in Figures B.8, B.9, B.10, for the Kolmgorov-Smirnov, Anderson-Darling and Cramer tests respectively and the tanglegrams in Figure B.11 in the appendix.

3.2 Study Sample Size Uncertainty

From the bootstrap uncertainty results it can be seen that there are certain studies for each test effecting the ordering of dendrograms which are:

- Kolmogorov-Smirnov: 25a, 25b, 37, 53 and 68
- Anderson-Darling: 14, 25a, 25b, 37, 40 and 53
- Cramer-von: 25b, 53 and 56

Generally the studies are consistent between the tests and the variations of pairings

changing are similar. By removing each study individually from the data and reapplying hierarchical clustering, the impact of the study can be measured in terms of how the clustering structure of dendrograms change.

In terms of the size of the studies detected, they are relatively small compared to larger studies, excluding study 56, therefore by removing each study individually it was seen that there was minimal impact to the clustering. Also, the variations of pairings were all removed for the respective study and the study to which it was paired with now had zero variation in pairings. This result was consistent for each test. By removing a study of large sample size, such as study 56 in the Cramer-von test, several pairings of studies are affected but generally most clusters are alike. Although, when removing study 56, which previously impacted study 25b in Table 3.2, the uncertainty of study 25b changing pairings was not removed. When removing study 56 for the Anderson-Darling test, the impact of pairings was significantly higher than compared to the Kolmogorov-Smirnov and Cramer-von.

3.3 Comparison of Linkage algorithms

The three linkage algorithms are applied to each test with 10,000 bootstrap samples. They were compared by correlation measures shown in Table 3.5. Validating each algorithms representation of the original dissimilarity matrix by the cophenetic distance correlation showed the best representation of the original data was consistently the average-link algorithm and the worst was single-link. The Kolmogorov-Smirnov and Anderson-Darling tests had similarly high correlations = 0.79732 and 0.76194 whereas the Cramer-von tests correlation = 0.62481 for average link and similar differences for all other links were seen.

Table 3.5: *Comparison of Linkage algorithms by correlation measures*

	Merge	Height	Order	CC	BG
Kolmogorov-Smirnov					
Complete-Single	1.06217	1.44304	0.68384	0.35497	0.30830
Complete-Average	0.93030	0.29576	0.60722	0.78399	0.58523
Single-Average	1.09090	0.46996	0.62795	0.63737	0.68463
Anderson-Darling					
Complete-Single	0.92046	0.60887	0.70642	0.36124	0.43489
Complete-Average	1.00481	0.19986	0.73548	0.81585	0.75168
Single-Average	1.12336	1.05429	0.76303	0.63266	0.79536
Cramer-von					
Complete-Single	1.19148	0.51810	0.69247	0.36772	0.55445
Complete-Average	0.97688	0.18767	0.82985	0.74697	0.82797
Single-Average	1.01699	0.70124	0.72019	0.77808	0.68762

Chapter 4

Conclusions

The results in chapter 3 are discussed and concluded in terms of the effect of bootstrapping for each test and which test was least prone to the uncertainties in bootstrapping; the effect of study sample size and which agglomerative algorithm provided the most accurate representation of the original dissimilarity distance structure.

For 10,000 bootstrap samples, the Kolmogorov-Sminrov and Anderson-Darling performed significantly better than the Cramer-von test in terms of the number of small groups with similar dendrograms and the group with the highest number of similar dendrograms (Group 1). The level of variation between groups for each test was seen to be similar but no significant changes to the structure of each dendrogram. For 1,000 bootstrap samples, the Kolmogorov-Smirnov compared significantly better than the Anderson-Darling and Cramer-von test for uncertainties between groups. The Anderson-Darling test and Cramer-von tests consisted of more groups than the Kolmogorov-Smirnov test but the Anderson-Darling test in particular caused the highest number of uncertainties between pairings changing. Although the Cramer-von test had a similar number of groups with the Anderson-Darling test, the number of uncertainties was fairly low. In the 100 bootstrap sample case, the number of groups were lower and there were less uncertainties shown due to their being a lower number of similar dendrograms. Therefore in general, the Kolmogorov-Smirnov test performed the best in all three cases, the Anderson-Darling test performed similarly in the 10,000

case but not for lower number of samples. The Cramer-von test consistently possessed the highest number of uncertainties.

In terms of the effect of sample size of studies, when removing small sample size studies the pairings that were previously unaffected were not anymore. By removing a sample size which is relative large, the structure of the dendrograms had a small impact and the uncertainties behind the study removed and studies affected was not removed. Therefore a sample size which is large has an impact on uncertainty of pairings and does not remove the uncertainties in which the structure has with the study in the clustering. Comparing linkage-algorithms and there representation of incubation period data, the average algorithm consistently performed the best with diverse results for complete and single dendrograms.

Overall, classifying incubation period data with hierarchical clustering is a valid method for subgroup analysis. Although there are many uncertainties, by applying a sensible number of bootstrap samples, above 10,000 the bootstrapping uncertainty would be low but whichever measure of distance is selected, studies with large sample size can significantly effect the clustering. The Kolmogorov-Smirnov test is the preferred test statistic and performs slightly better than the Anderson-Darling which both perform better than the Cramer-von test.

4.1 Improvements

Since this methodology is fairly new and this dissertation is the first attempt to analyse the validity, the methods could be improved with further theoretical analysis. The first improvement would be to use simulated data of the incubation period. The data used in this dissertation only contains 30 studies which vary with sample size. Therefore the credibility of such methods are not as reliable as they would be with more data. Secondly only non-parametric tests which measure the whole distributions were used as a measure of dissimilarity. Tests such as Mann-Whitney which measure the central tendency of a data set and do not consider the distribution itself can be applied with

higher moments to look at the shape, then this would not lose information from the data set.

Another approach of assessing the uncertainty of bootstrapping and sample size is to consider the sequence of clustering by using p-values but with weighted comparison, such as the sample size or an estimate of mean along with distance. This analysis would require careful construction but would be the next step to this dissertations findings.

References

Anderberg, M. R. (1973), *Cluster analysis for applications*, Probability and mathematical statistics, 19, Academic Press, New York.

Arshad, M., Rasool, M. & Ahmad, M. (2003), 'Anderson Darling and Modified Anderson Darling Tests for Generalized Pareto Distribution', *Journal of Applied Sciences* **3**(2), 85–88.

Awofisayo-Okuyelu, A., Hall, I., Adak, G., Hawker, J. I., Abbott, S. & Mccarthy, N. (2017), 'A systematic review and meta-analysis on the incubation period of Campylobacteriosis', *Epidemiology and Infection* **145**(11), 2241–2253.

Bacchetti, P. & Jewell, N. P. (1991), 'Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times', *Biometrics* **47**(3), 947–60.

Baker, F. B. (1972), 'Numerical Taxonomy for Educational Researchers', *Review of Educational Research* **42**(3), 345–358.

Baker, F. B. (1974), 'Stability of Two Hierarchical Grouping Techniques Case I: Sensitivity to Data Errors', *Journal of the American Statistical Association* **69**(346), 440–445.

Baringhaus, L. & Franz, C. (2004), 'On a new multivariate two-sample test', *Journal of Multivariate Analysis* **88**(1), 190–206.

Bobisud, H. M. & Bobisud, L. E. (1972), 'A Metric for Classifications', *Taxon* **21**(5/6), 607–613.

Bock, H. H. (1989), Probabilistic Aspects in Cluster Analysis, *in* O. Optiz, ed., 'Conceptual and Numerical Analysis of Data', Springer Berlin Heidelberg, pp. 12–44.

Brookmeyer, R. (2014), Incubation Period of Infectious Diseases, *in* 'Wiley StatsRef: Statistics Reference Online', American Cancer Society.

Carmichael, J. W., George, J. A. & Julius, R. S. (1968), 'Finding Natural Clusters', *Systematic Zoology* **17**(2), 144–150.

Conover, W. J. (1999), *Practical nonparametric statistics*, Wiley series in probability and statistics. Applied probability and statistics section, 3rd ed. edn, John Wiley, New York ;.

Day, W. (1985), 'Optimal algorithms for comparing trees with labeled leaves', *Journal of Classification* **2**(1), 7–28.

Dwivedi, A. K., Mallawaarachchi, I. & Alvarado, L. A. (2017), 'Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method', *Statistics in Medicine* **36**(14), 2187–2205.

Efron, B. (1993), *An introduction to the bootstrap*, Monographs on statistics and applied probability ; 57, Chapman & Hall, New York ;.

Eichner, M. & Dietz, K. (2003), 'Transmission Potential of Smallpox: Estimates Based on Detailed Data from an Outbreak', *American Journal of Epidemiology* **158**(2), 110–117.

Everitt, B. (1980), 'Cluster analysis', *Quality and Quantity* **14**(1), 75–100.

Farrell, P. J. & Rogers-Stewart, K. (2006), 'Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test', *Journal of Statistical Computation and Simulation* **76**(9), 803–816.

Fowlkes, E. B. & Mallows, C. L. (1983), 'A Method for Comparing Two Hierarchical Clusterings', *Journal of the American Statistical Association* **78**(383), 553–569.

G. N. Lance & W. T. Williams (1966), 'A Generalized Sorting Strategy for Computer Classifications', *Nature* **212**(5058), 218–218.

Gail, M. & Brookmeyer, R. (1988), 'Methods for projecting course of acquired

immunodeficiency syndrome epidemic', *Journal of the National Cancer Institute* **80**(12), 900–911.

Galili, T. (2015), 'dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering', *Bioinformatics* **31**(22), 3718–3720.

Gan, G. (2007), *Data clustering theory, algorithms, and applications*, ASA-SIAM series on statistics and applied probability ; 20, Society for Industrial and Applied Mathematics SIAM Market Street, Floor 6, Philadelphia, PA 19104, Philadelphia, Pa.

Gibbons, J. D. (2003), *Nonparametric statistical inference*, Statistics, textbooks and monographs ; 168, 4th ed., rev. and expanded. edn, Marcel Dekker, New York.

Gordon, A. D. (1987), 'A Review of Hierarchical Classification', *Journal of the Royal Statistical Society. Series A (General)* **150**(2), 119–137.

Gordon, A. D. (1999), *Classification, 2nd Edition*, CRC Press. Google-Books-ID: _w5AJtbfEz4C.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002), 'Cluster validity methods: part I', *ACM SIGMOD Record* **31**(2), 40–45.

Jain, A. K. (1988), *Algorithms for clustering data*, Prentice-Hall, Englewood Cliffs.

Kassambara, A. (2017), *Practical guide to cluster analysis in R: unsupervised machine learning*, edition 1. edn, STHDA, United States].

Kiefer, J. (1959), 'K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-V. Mises Tests', *The Annals of Mathematical Statistics* **30**(2), 420–447.

Mcquitty, L. L. (1960), 'Hierarchical Linkage Analysis for the Isolation of Types', *Educational and Psychological Measurement* **20**(1), 55–67.

Mcquitty, L. L. (1967), 'Expansion of Similarity Analysis By Reciprocal Pairs for Discrete and Continuous Data', *Educational and Psychological Measurement* **27**(2), 253–255.

Miner, J. R. (1922), 'The Incubation Period of Typhoid Fever', *The Journal of Infectious Diseases* **31**(3), 296–301.

Morlini, I. & Zani, S. (2012), 'Dissimilarity and similarity measures for comparing dendrograms and their applications', *Advances in Data Analysis and Classification* **6**(2), 85–105.

Murdoch, D. J., Tsai, Y.-L. & Adcock, J. (2008), 'P -Values are Random Variables', *The American Statistician* **62**(3), 242–245.

N. Lance, G. & T. A. Williams, W. (1967), 'A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems', *The Computer Journal* **9**.

Nishiura Hiroshi (2007), 'Early efforts in modeling the incubation period of infectious diseases with an acute course of illness', *Emerging Themes in Epidemiology* **4**(1), 2.

Odilia Yim & Kylee T. Ramdeen (2015), 'Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data', *Tutorials in Quantitative Methods for Psychology* **11**(1), 8–21.

Reilly, C., Wang, C. & Rutherford, M. (n.d.), 'A rapid method for the comparison of cluster analyses', p. 15.

Sartwell, P. E. (1949), 'The distribution of incubation periods of infectious disease', *American journal of epidemiology* **141**(5), 386–94.

Scholz, F. W. & Stephens, M. A. (1987), 'K-Sample Anderson-Darling Tests', *Journal of the American Statistical Association* **82**(399), 918–924.

Sneath, P. H. (1957), 'The application of computers to taxonomy', *Journal of general microbiology* **17**(1), 201–26.

Sneath, P. H. A. (1973), *Numerical taxonomy: the principles and practice of numerical classification*, WHFreeman, San Francisco.

Sokal, R. & Michener, C. (1958), 'A statistical method for evaluating systematic relationships', *University of Kansas Scientific Bulletin* **28**, 1409–1438.

Sokal, R. R. & Rohlf, F. J. (1962), 'The Comparison of Dendrograms by Objective Methods', *Taxon* **11**(2), 33–40.

Tango, T. (1998), 'Maximum likelihood estimation of date of infection in an outbreak

of diarrhea due to contaminated foods assuming lognormal distribution for the incubation period', *[Nihon Koshu Eisei Zasshi] Japanese Journal of Public Health* **45**(2), 129–141.

Waterman, M. S. & Smith, T. F. (1978), 'On the similarity of dendrograms', *Journal of Theoretical Biology* **73**(4), 789–800.

Zaït, M. & Messatfa, H. (1997), 'A comparative study of clustering methods', *Future Generation Computer Systems* **13**(2), 149–159.

Appendix A

Proofs

Proof of equation (2.18) taken from (Gan 2007):

Let C_1, C_2 and C_3 be three nonempty, mutually non-overlapping clusters and assume:

$$D(C_i, C_j) = \frac{1}{n_i n_j} \sum (C_i, C_j), \quad 1 \leq i < j \leq 3$$

where $n_i = |C_i|, n_j = |C_j|$, and $\sum(C_i, C_j)$ is the total between-clusters distance of C_i and C_j, that is,

$$\sum (C_i, C_j) = \sum_{x \in C, y \in C'} d(x, y).$$

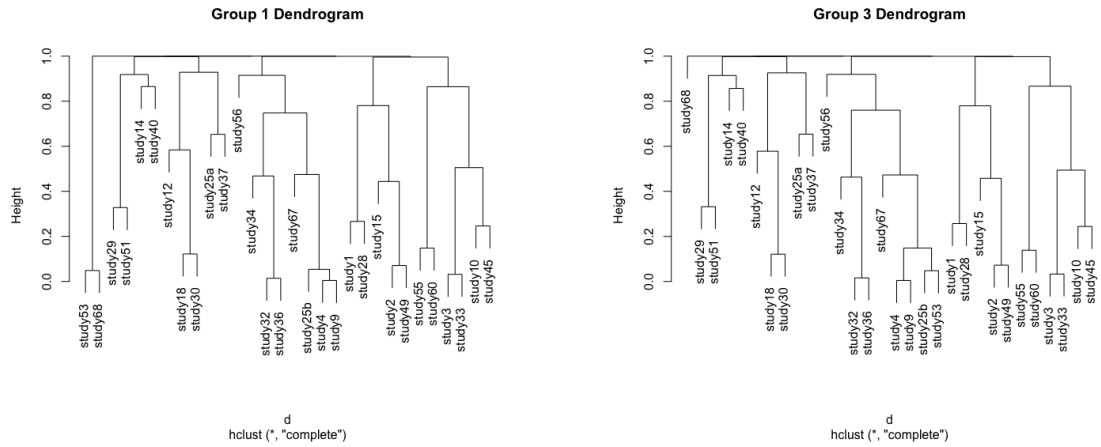
From equations (2.17) and (2.18)

$$\begin{aligned} & D(C_1, C_2 \cup C_3) \\ & \qquad = \frac{n_2}{n_2 + n_3} D(C_1, C_2) + \frac{n_3}{n_2 + n_3} D(C_1, C_3) \\ & \qquad = \frac{n_2}{n_2 + n_3} \cdot \frac{1}{n_2 n_1} \sum (C_1, C_2) + \frac{n_3}{n_2 + n_3} \cdot \frac{1}{n_1 n_3} \sum (C_1, C_3) \\ & \qquad = \frac{1}{n_1 (n_2 + n_3)} \sum (C_1, C_2 \cup C_3), \end{aligned}$$

since $\sum(C_1, C_2) \sum(C_1, C_3) = \sum(C_1, C_2 \cup c_3)$ which verifies equation (2.18)

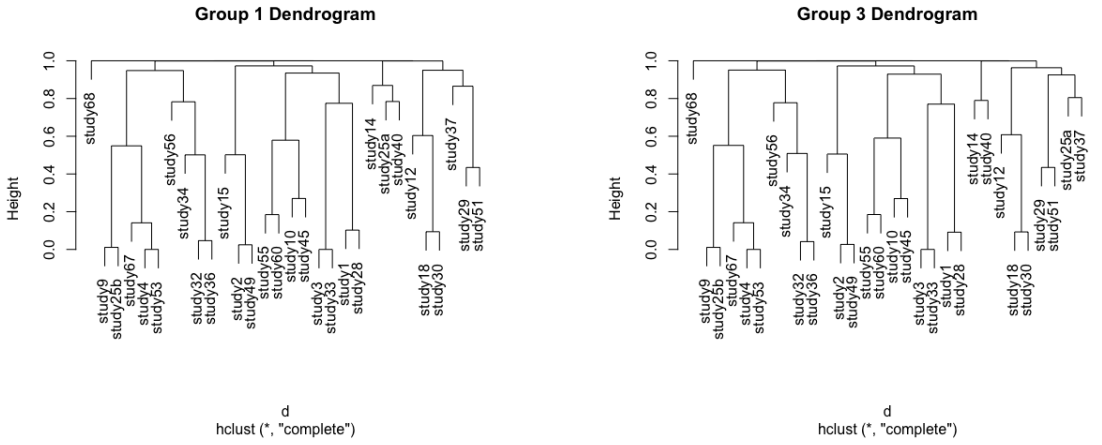
Appendix B

Figures



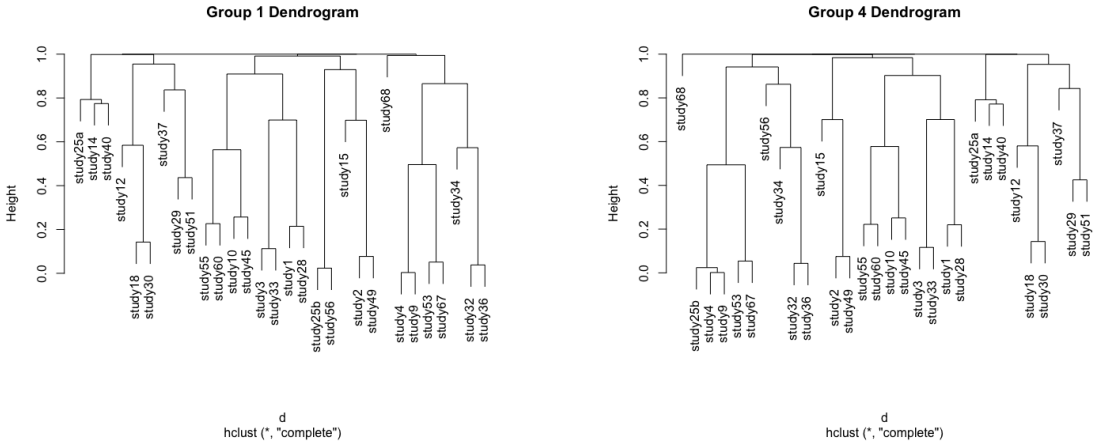
(a) *Group 1* **(b)** *Group 3*

Figure B.1: *Group 1 dendrogram and Group 3 dendrogram for the Kolmogorov-Smirnov test with 10,000 bootstrap samples*



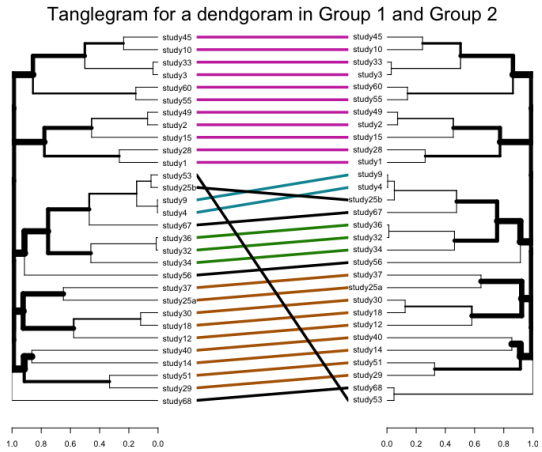
(a) *Group 1* **(b)** *Group 3*

Figure B.2: *A dendrogram in group 1* ***(a)*** *and a dendrogram in group 3* ***(b)*** *for the Anderson-Darling test with 10,000 bootstrap samples*

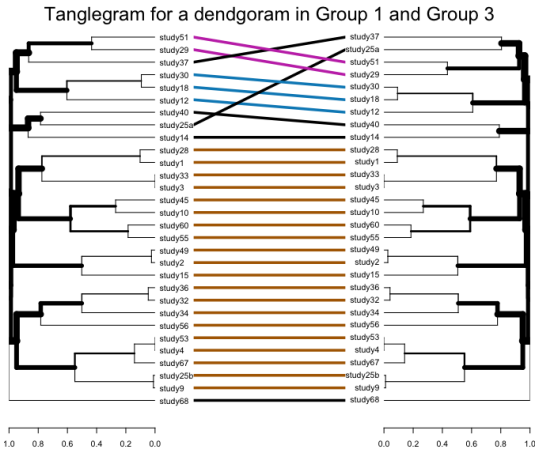


(a) *Group 1* **(b)** *Group 3*

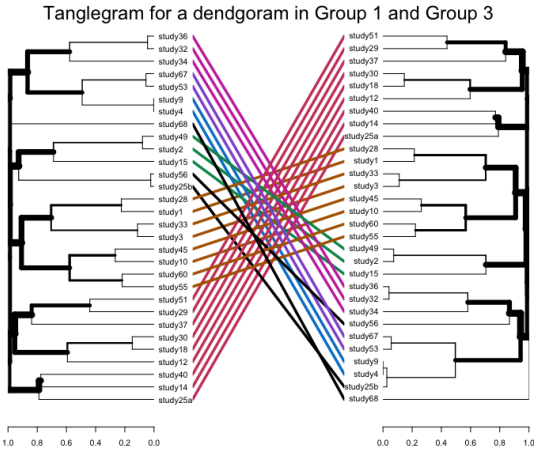
Figure B.3: *A dendrogram in group 1* ***(a)*** *and a dendrogram in group 4* ***(b)*** *for the Cramer-von test with 10,000 bootstrap samples*



(a) *Kolmogorov-Smirnov*

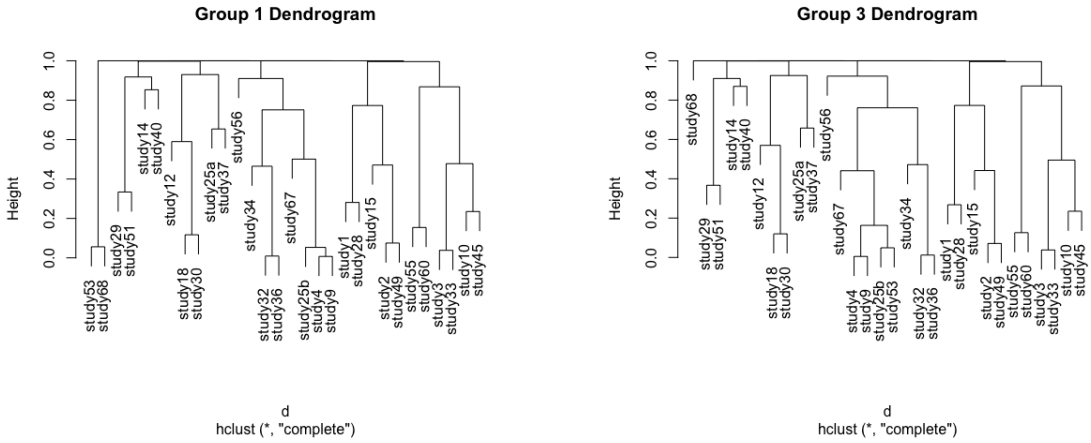


(b) *Anderson-Darling*



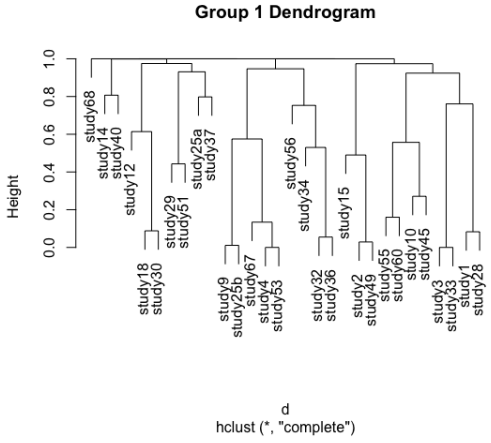
(c) *Cramer-von*

Figure B.4: *Tanglegrams for the Kolmogorov-Smirnov test, groups 1 & 2* ***(a)****; Anderson-Darling test, groups 1 & 3* ***(b)****; Cramer-von test, groups 1 & 3,* ***(c)***

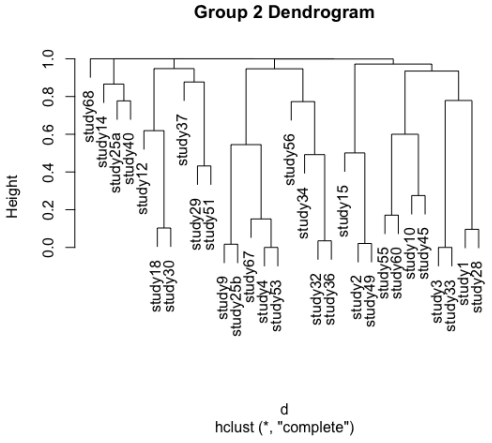


(a) *Group 1* **(b)** *Group 3*

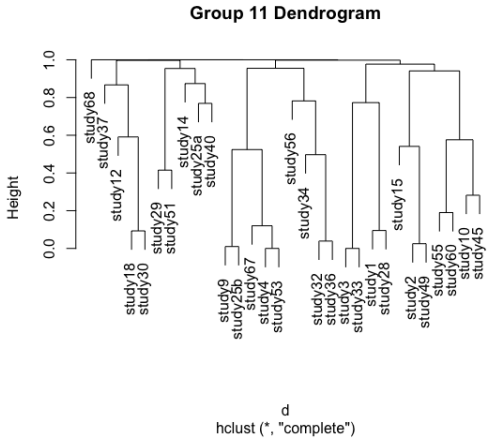
Figure B.5: *Group 1 dendrogram and Group 3 dendrogram for the Kolmogorov-Smirnov test with 1,000 bootstrap samples*



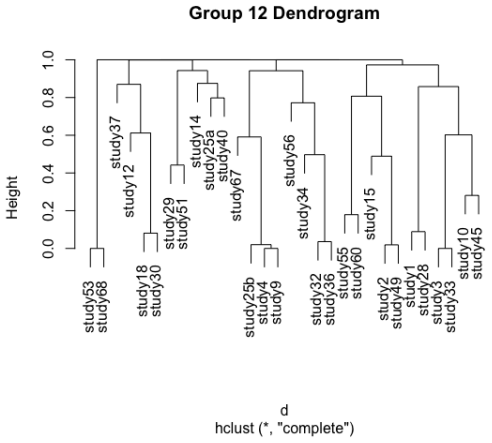
(a) *Group 1*



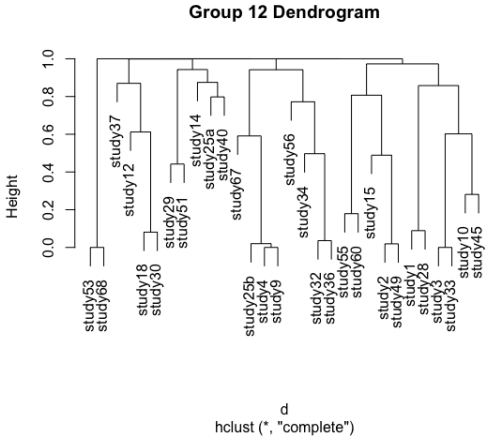
(b) *Group 2*



(c) *Group 11*

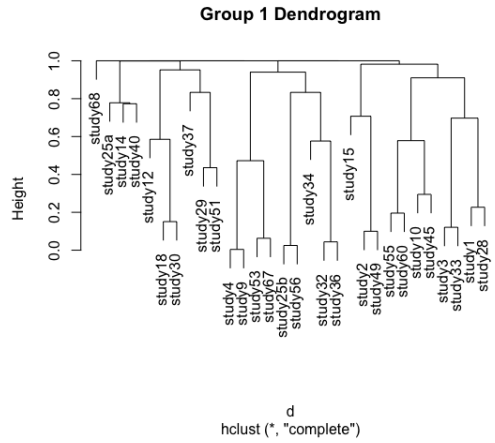


(d) *Group 12*

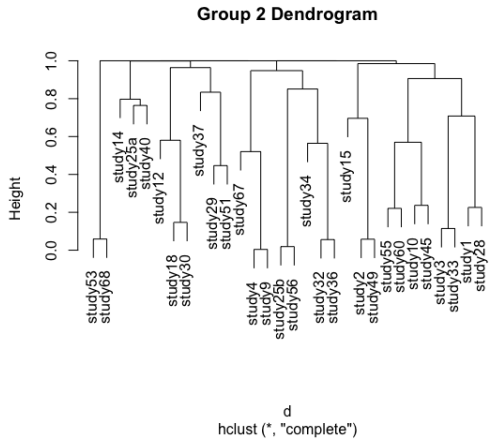


(e) *Group 16*

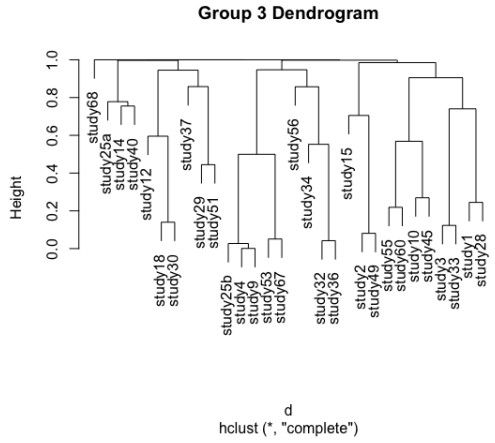
Figure B.6: *A dendrogram in group 1* ***(a)****, dendrogram in group 2* ***(b)****, dendrogram in group 11* ***(c)****, dendrogram in group 12* ***(d)*** *and a dendrogram in group 16* ***(E)*** *for the Anderson-Darling test with 1,000 bootstrap samples*



(a) *Group 1*

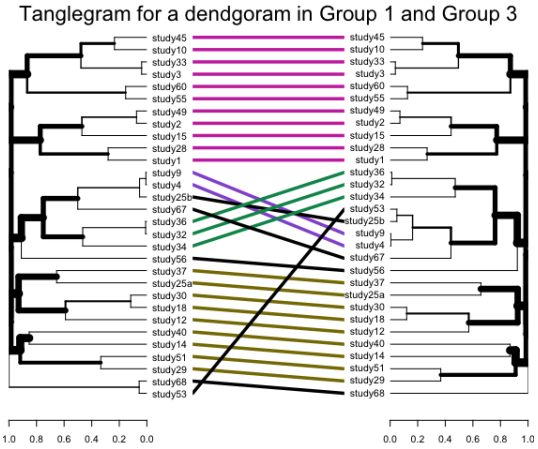


(b) *Group 2*

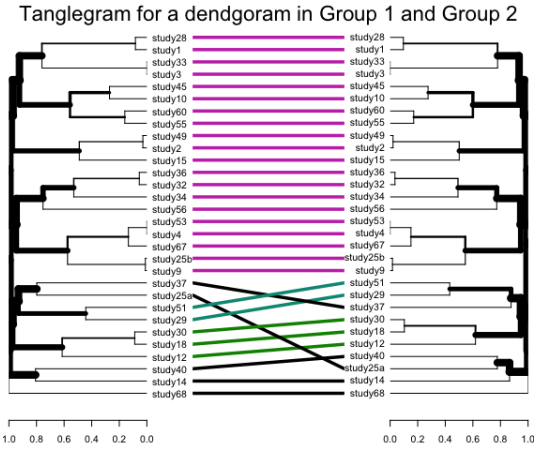


(c) *Group 3*

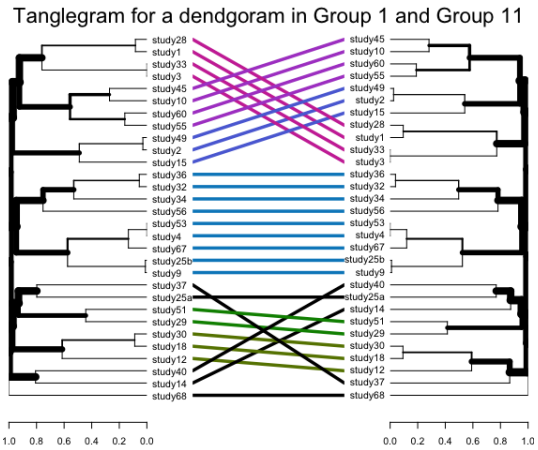
Figure B.7: *A dendrogram in group 1* ***(a)****, dendrogram in group 2* ***(b)*** *and dendrogram in group 3* ***(c)*** *for the Cramer-von test with 1,000 bootstrap samples*



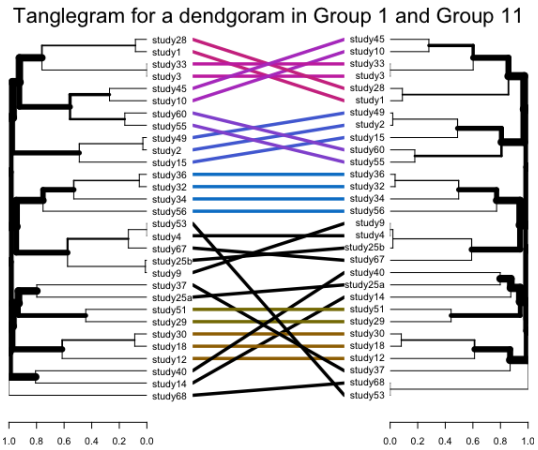
(a) *Kolmogorov-Smirnov (Group 1-3)*



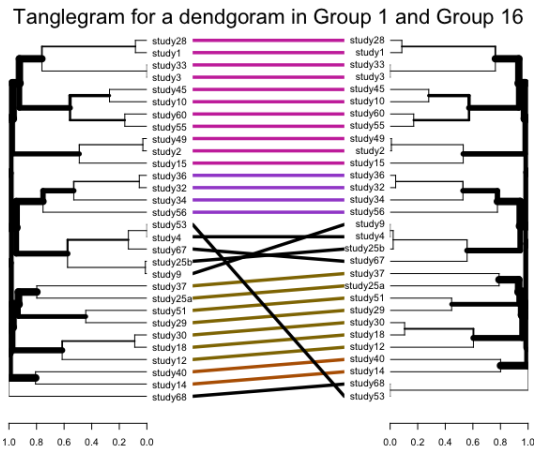
(b) *Anderson-Darling (Group 1-2)*



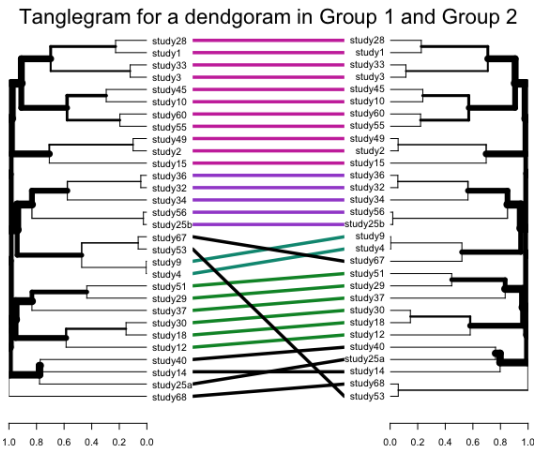
(c) *Anderson-Darling (Group 1-11)*



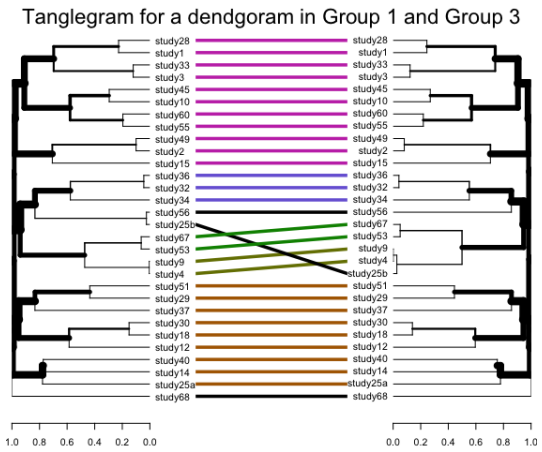
(d) *Anderson-Darling (Group 1-12)*



(e) *Anderson-Darling (Group 1-16)*

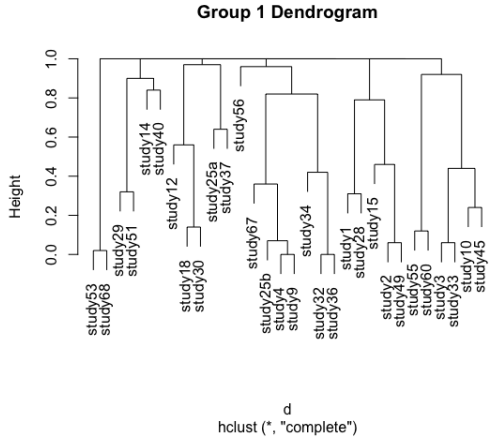


(f) *Cramer-von (Group 1-2)*

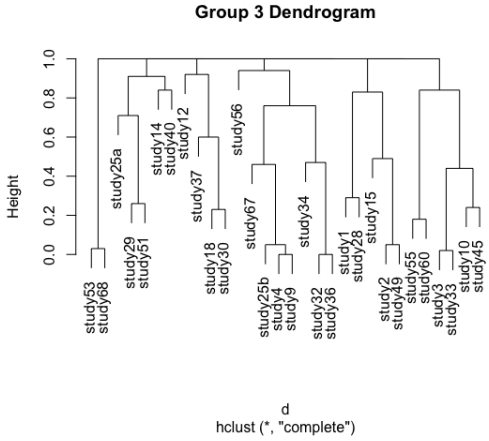


(g) *Cramer-von (Group 1-3)*

Figure B.7: *Tanglegrams for the Kolmogorov-Smirnov test, groups 1 & 3* ***(a)*** *; Anderson-Darling test, groups 1 & 3* ***(b)****, groups 1 & 11* ***(c)****, groups 1 & 12* ***(d)****,groups 1 & 16* ***(e)****,; Cramer-von test, groups 1 & 2,* ***(f)****,groups 1 & 3* ***(h)***

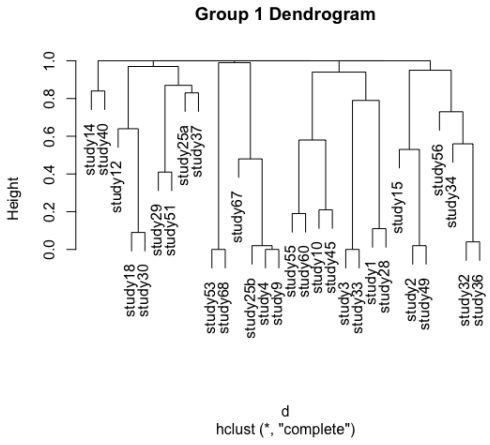


(h) *Group 1*

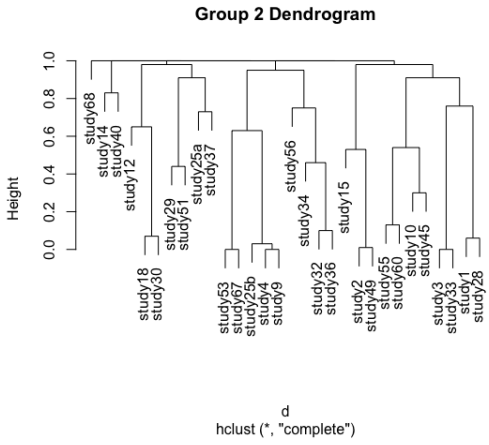


(i) *Group 3*

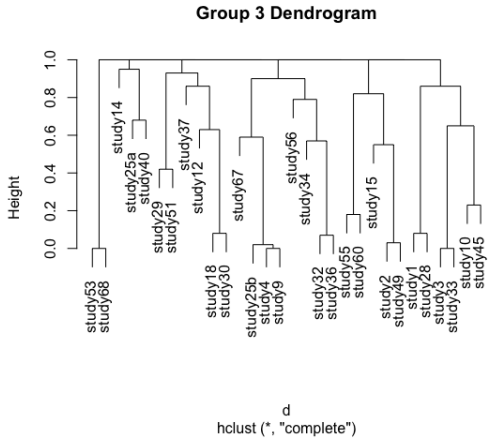
Figure B.8: *Group 1 dendrogram and Group 3 dendrogram for the Kolmogorov-Smirnov test with 100 bootstrap samples*



(a) *Group 1*

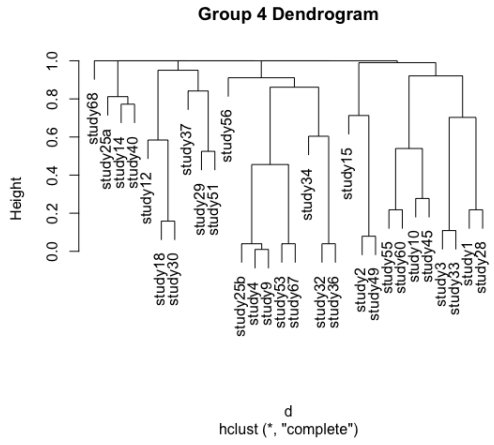


(b) *Group 2*

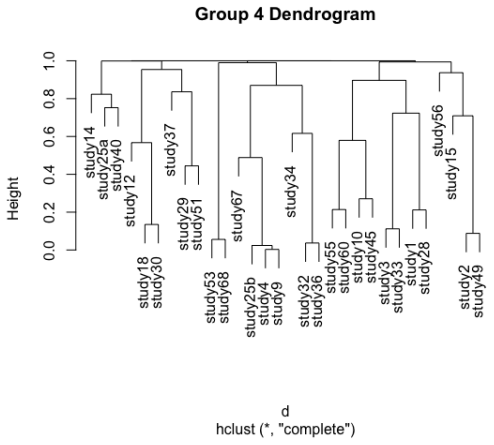


(c) *Group 3*

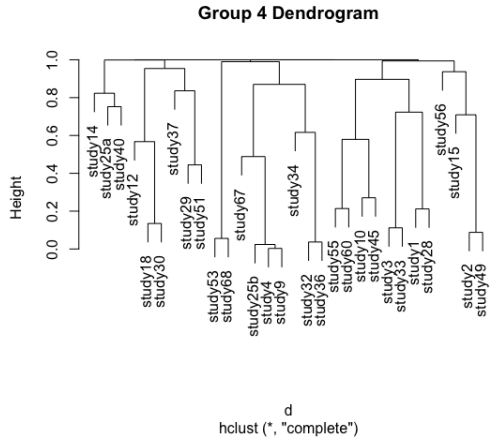
Figure B.9: *A dendrogram in group 1* ***(a)****, dendrogram in group 2* ***(b)****, dendrogram in group 3* ***(c)*** *for the Anderson-Darling test with 100 bootstrap samples*



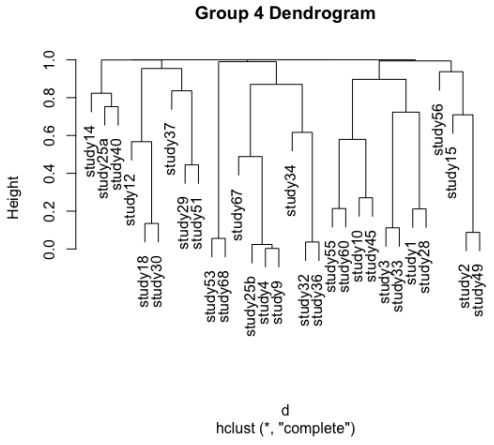
(a) *Group 1*



(b) *Group 2*

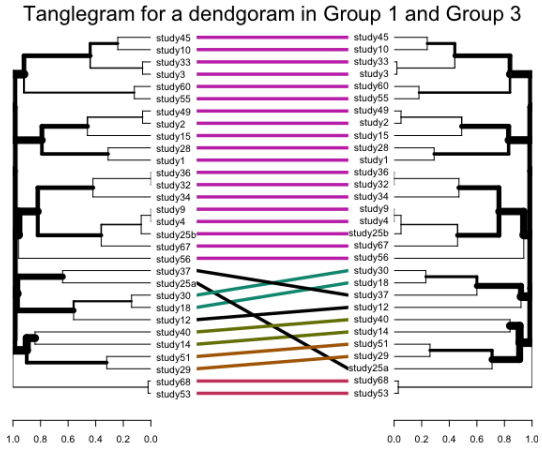


(c) *Group 3*

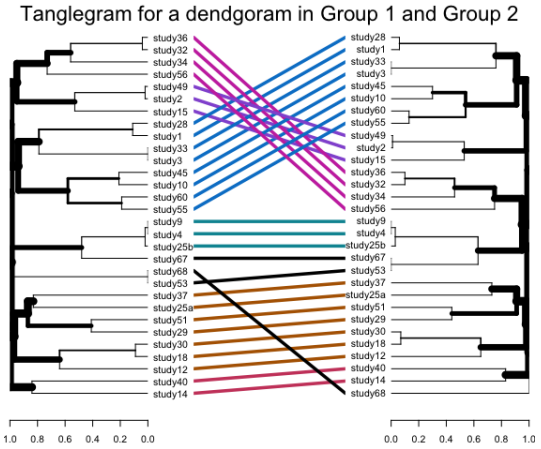


(d) *Group 4*

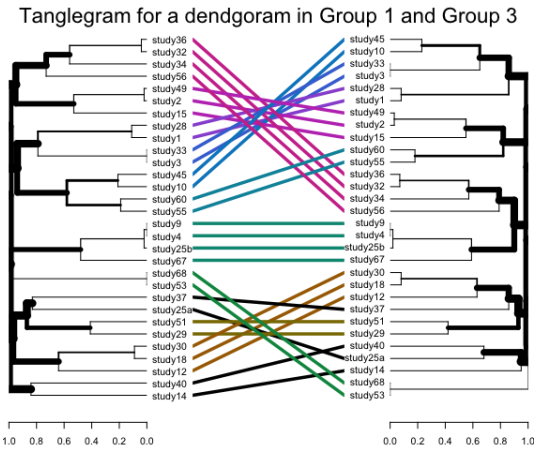
Figure B.10: *A dendrogram in group 1* ***(a)****, dendrogram in group 2* ***(b)****, dendrogram in group 3* ***(c)*** *and dendrogram in group 3* ***(d)*** *for the Cramer-von test with 100 bootstrap samples*



(a) *Kolmogorov-Smirnov (Group 1-3)*



(b) *Anderson-Darling (Group 1-2)*



(c) *Anderson-Darling (Group 1-3)*

Figure B.11: *Tanglegrams for the Kolmogorov-Smirnov test, groups 1 & 3* ***(a)*** *; Anderson-Darling test, groups 1 & 2* ***(b)****, groups 1 & 3* ***(c)***

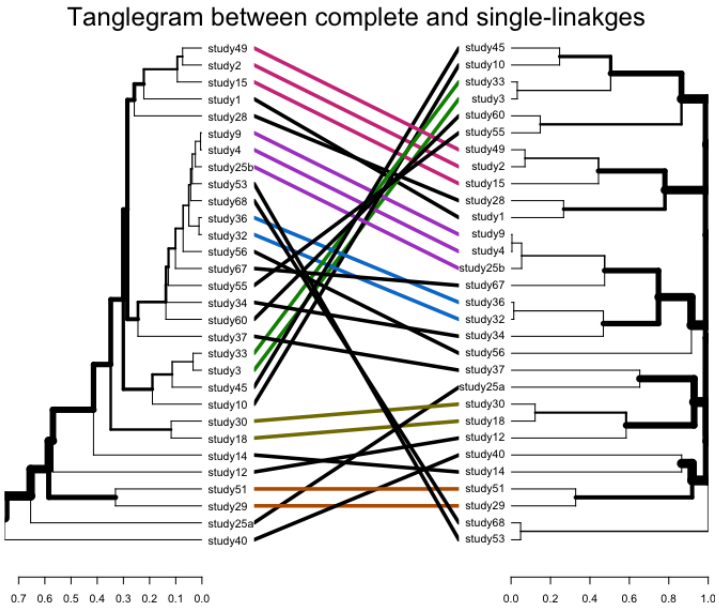


Figure B.12: *Tanglegram of Complete and Single-link algorithms for Kolmogorov-Smirnov test*

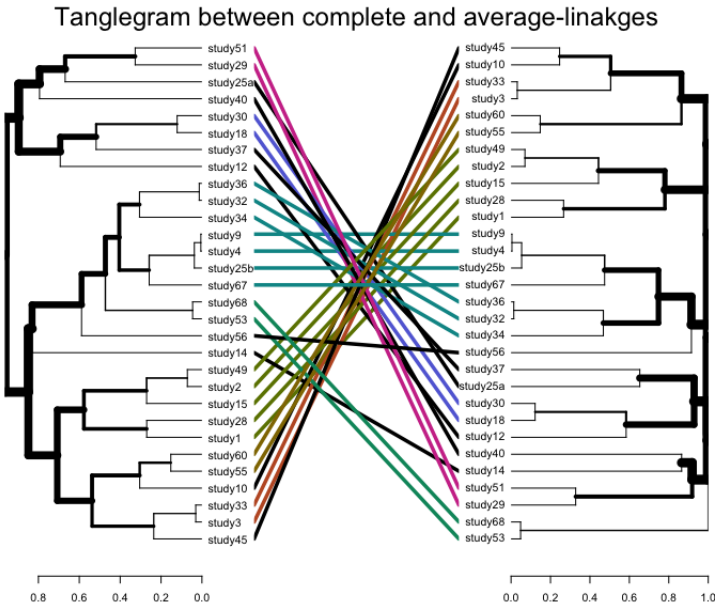


Figure B.13: *Tanglegram of Complete and average-link algorithms for Kolmogorov-Smirnov test*

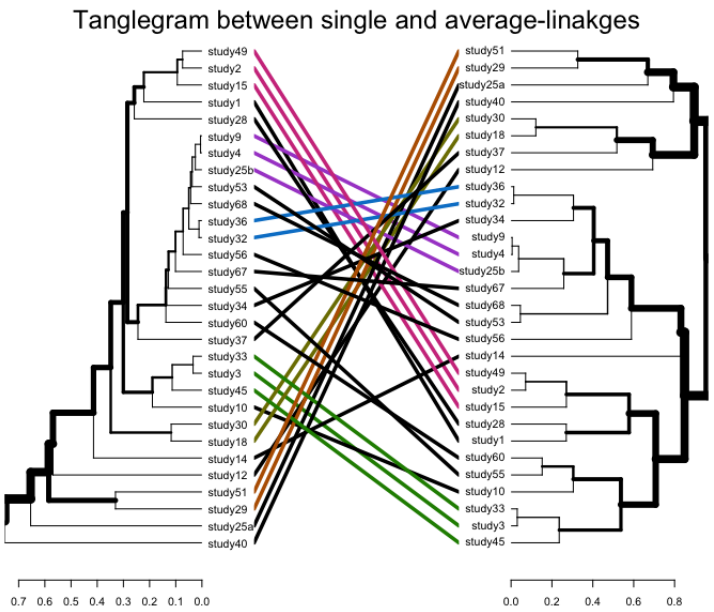
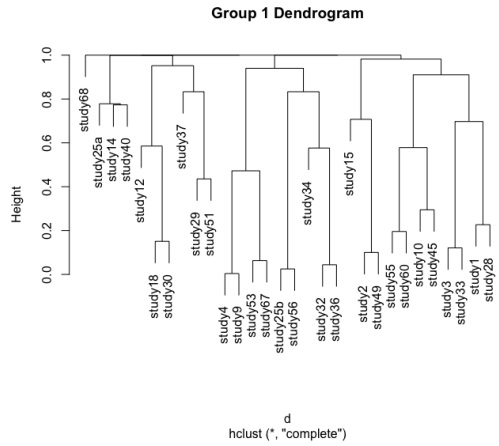
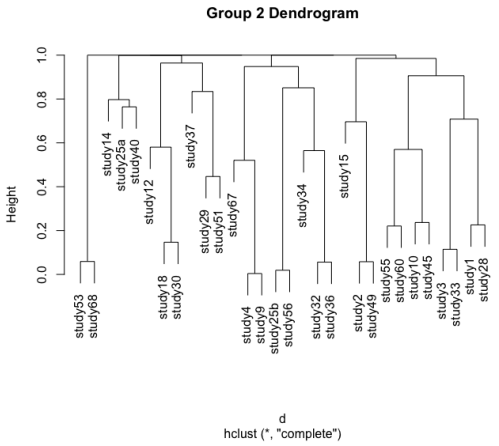


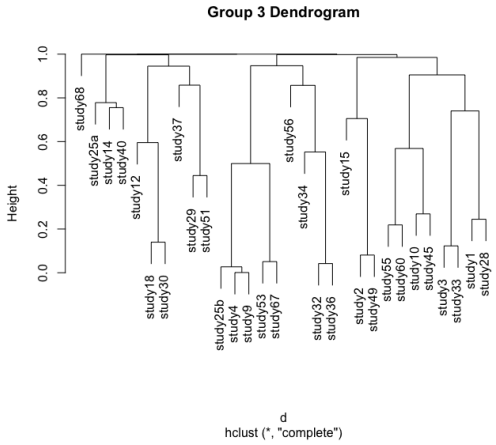
Figure B.14: *Tanglegram of single and average-link algorithms for Kolmogorov-Smirnov test*



(a) *Group 1*



(b) *Group 2*



(c) *Group 3*

Figure B.15: *A dendrogram in group 1* ***(a)****, dendrogram in group 2* ***(b)****, and dendrogram in group 3* ***(c)*** *for 1,000 bootstrap samples in the cramer test*