Prologue

A variety of data can be represented in strings of binary scores. In general, the binary scores reflect either the presence or absence of certain attributes of a certain object. For example, in psychology binary data may indicate if people do or do not posses a certain psychological trait; in ecology, the objects could be regions or districts in which certain species do or do not occur (or vice versa, the objects are two species that coexist in a number of locations); in archeology, binary data may reflect that particular artifact types were or were not found in a specific grave; finally, in chemical similarity searching, the objects may be target structures or queries and the attributes certain compounds in a database.

A vast amount of measures has been proposed that indicate how similar binary sequences are. A so-called similarity coefficient reflects in one way or another the association or resemblance of two or more binary variables. In various methods of data analysis, for example, multidimensional scaling or cluster analysis, the full information in the recorded binary variables is not required to perform the analysis. Often, the binary data are first summarized by a few coefficients or a coefficient matrix of pairwise resemblance measures. The information in the similarity coefficients is then used as input for the method of data analysis at hand.

Although the full information in comparing two binary variables is often not required, there are many different similarity coefficients that may be used to summarize the bivariate information. Preferring one coefficient over another may determine what information is summarized or what information is discarded. In order to choose the right coefficient, the different coefficients and their properties need to be better understood. Some properties of similarity coefficients for binary data are studied in this thesis. However, no attempt is made to be complete in the sense that all possible data-analytic applications of coefficients for binary data are covered. Instead, the thesis is centered around two theoretical issues.

The first issue is captured in the question, can the task of choosing the right coefficient be simplified? It may turn out that a coefficient may be placed in a group of coefficients all sharing a certain property. With respect to the property any coefficient in the group or family of coefficients can be used: one is as good as

the other. On the other hand, the property may also divide coefficients in different groups, coefficients that do posses the property and those that do not. For example, when comparing two binary variables it is not uncommon to be interested in the similarity between the variables corrected for possible similarity due to chance. It may turn out that some coefficients become equivalent after correction. The choice of coefficient can then be limited to coefficients that are not equivalent after correction for chance. As a second example, in cluster analysis several algorithms only make use of the ordinal information between the different coefficients, ignoring the numerical values. Coefficients can be grouped on the basis of what information they preserve with respect to an ordinal data analysis. The choice of coefficient can then be limited to coefficients that summarize different ordinal information.

As a second issue, a similarity coefficient must sometimes be considered in the context of the data-analytic study of which it is a part. Some method of data analysis may have certain prerequisites. If a coefficient possesses a specific property, it may be preferred over a coefficient which does not share this characteristic. For example, the outcome of metric data analysis methods like classical scaling, is better understood if the coefficient used in the analysis is metric, that is, satisfies the triangle inequality. As a bonus, the study of various properties of similarity coefficients provides a better understanding of the coefficients themselves. The insight obtained from how different coefficients are related, for example, one coefficient is the product of a transformation applied to a second coefficient, provides new ways of interpreting both coefficients.

The dissertation contains a mathematical approach to the analysis of resemblance measures for binary data. A variety of data-analytic properties are considered and for various coefficients it is established whether they possess the property or not. Counterexamples are sometimes used to show that a coefficient lacks a property. All mathematics are on the level of high school algebra and to read the thesis no 'higher' mathematical training is required. A statement is referred to as a proposition if it is believed to be a new result; a statement is called a theorem if the result is already known.

The first half of the dissertation (Part I and II) is devoted to what is basically two-way information. In the literature on data-analytic methods like, for example, cluster analysis, factor analysis, or multidimensional scaling, a distinction is made between two types of two-way information. Two-way similarity may be the bivariate information between two binary or dichotomous variables, that is, variables with two responses. Two-way similarity may also be the dyadic information between cases, persons, or objects. For the reader who is accustomed to this terminology it is important to note that in the present dissertation this (historical) distinction is largely ignored.

Some of the coefficients that are studied in the thesis have been proposed for comparing variables over cases, whereas others are primarily used to compare objects or cases over variables or attributes. Perhaps only a few coefficients are actually used in both the bivariate and dyadic case. Basically, similarity of two sequences of binary scores is referred to as two-way or bivariate information. The two terms are

considered interchangeable. To simplify the reading the sequences are referred to as variables. When considering a case by variable data matrix, the variables correspond to the columns. The latter notion is important in Part II on similarity matrices. A similarity matrix is obtained by calculating all two-way or pairwise coefficients between the columns of the case by variable data table. Finally, when two or more sequences are compared the words multi-way and multivariate are used.

This thesis consists of nineteen chapters divided into four parts. Part I and II are devoted to the bivariate case: a coefficient reflects the similarity of two variables at a time. Properties of individual coefficients are considered in Part I, whereas Part II focuses on properties that are studied in terms of coefficient matrices. Part III and IV are concerned with definitions and generalizations of various concepts from Part I and II to the multi-way case: a coefficient measures the resemblance of two or more binary variables. Part III is somewhat different from the other parts because no similarity coefficients are encountered in its chapters. Instead, various generalizations of the triangle inequality and other multi-way possibilities are studied in Part III. Some of the properties derived in Part III are used in Chapter 18 on metric properties of multi-way coefficients.

Part I consists of five chapters. Notation and some basic concepts concerning similarity coefficients are introduced in Chapter 1. We consider axioms for both similarity and dissimilarity coefficients. A first distinction is made between coefficients that do and coefficients that do not include the number of negative matches. A second distinction is made between coefficients that have zero value if the two variables are statistically independent and coefficients that have not. Also, some attention is paid to the problem of indeterminate values for coefficients that are fractions.

Chapter 2 is used to put the similarity coefficients for binary data into a broader perspective. The formulas considered in this thesis are often special cases that are obtained when more general formulas from various domains of data analysis are applied to dichotomous data. Furthermore, the same formulas may be encountered when two nominal variables are compared. For example, when comparing partitions from two cluster analysis algorithms or when measuring response agreement between two judges, a general approach is to count the four different types of pairs that can be obtained. The formulas defined on the four types of pairs may be equivalent to formulas defined on the four quantities obtained when comparing two binary variables.

In Chapter 3 it is shown that some resemblance measures belong to some sort of family of coefficients. Various relations between coefficients become apparent from studying their membership to a family. For most properties studied in Part I, greater generality is obtained if one works with (various types of) coefficient families. Linearity, another topic of this chapter, and metric properties (Chapter 10) are studied for families in which each coefficient is linear in both numerator and denominator.

Correction for chance agreement is the theme of Chapter 4. The chapter focuses on a coefficient family for which the study of correction for chance is relatively

simple. Several new properties on equivalences of coefficients after correction for chance irrespective of the choice of expectation are presented. In addition, a variety of properties of corrected coefficients are considered. Special interest is taken in a certain class of coefficients that become equivalent after correction. Also discussed is the relationship between the actual formula (coefficient) obtained after correction for chance and the particular choice of expectation.

The maximum value of various similarity coefficients is the topic of Chapter 5. Maximum values are studied in relationship to coefficient families that are power means. It is shown that different members of a specific family all have the same maximum value. New formulas are obtained if a coefficient is divided by its maximum value. Several results are presented that show what formulas are obtained after division by the maximum value. Two classes of coefficients are considered that become either a coefficient by Simpson (1943) or a coefficient by Loevinger (1947, 1948). Also, it is shown that Loevinger's coefficient is obtained if a general family of coefficients is corrected for both similarity due to chance and maximum value.

Part II consists of five chapters. In many applications of data analysis the data consist of more than two binary variables. In Part II various concepts and properties are considered that can only be studied when multiple variables (more than two) are considered. For example, multiple column vectors can be positioned next to each other to form a so-called data matrix. Given a binary data matrix, one may obtain a coefficient matrix by calculating all pairwise coefficients for any two columns of the data matrix. Different coefficient matrices are obtained, depending on the choice of similarity coefficient.

Chapter 6 focuses on how the 1s and 0s of the various column vectors of the data matrix may be related. For example, the 1s and 0s may be related in such a way that the data matrix exhibits certain patterns, possibly after a certain re-ordering or permutation of the columns, or after permuting both columns and rows of the data matrix. The 1s and 0s of the various column vectors may also be related in more complicated ways, not immediately clear from visual inspection. For example, some sort of probabilistic model can supposedly underlie the patterns of 1s and 0s of the various variables. Chapter 6 is used to describe some one-dimensional models and data structures that imply a certain ordering of the column vectors. These data structures are later on used in the remaining chapters of Part II for the study of various ordering properties of similarity matrices.

Chapter 7 is devoted to Robinson matrices. A square similarity matrix is called a Robinson matrix if the highest entries within each row and column are on the main diagonal and moving away from this diagonal, the entries never increase. A similarity matrix may or may not exhibit the Robinson property depending on the choice of resemblance measure. However, it seems to be a common notion in the classification literature that Robinson matrices arise naturally in problems where there is essentially a one-dimensional structure in the data. It is shown in Chapter 7 that the occurrence of a Robinson matrix is a combination of the choice of the similarity coefficient, and the specific one-dimensional structure in the data. Important coefficients in this chapter are the coefficient by Braun-Blanquet (1932) and Russel

and Rao (1940).

Eigendecompositions of several coefficient matrices are studied in Chapter 8. It is shown what information on the order of the model probabilities can be obtained from the eigenvector elements corresponding to the largest eigenvalues of various similarity matrices. It is therefore possible to uncover the correct ordering of several latent variable models considered in Chapter 6 using eigenvectors. The point to be made here is that the eigendecomposition of some similarity matrices, especially matrices corresponding to asymmetric coefficients, are more interesting compared to the eigendecomposition of other matrices. The important coefficients in this chapter have corresponding similarity matrices that are non-symmetrical. Also, the diverse matrix methodology of an eigenvalue method called homogeneity analysis is studied.

In Chapter 9, a systematic comparison of a one-dimensional homogeneity analysis and the item response theory approach is presented. It is shown how various item statistics from classical item analysis are related to the parameters of the 2-parameter logistic model from item response theory. Using these results, and the assumption that the homogeneity person score is a reasonable approximation of the latent variable, the functional relationships between the discrimination and location parameter of the 2-parameter logistic model and the two category weights of a homogeneity analysis applied to binary data are derived.

The study of metric properties is begun in Chapter 10, where metric properties of coefficients that are linear in both numerator and denominator are discussed. The chapter starts with an introduction of the concept of dissimilarity. Some tools are introduced here for the two-way case. Metric properties for multi-way coefficients are studied in Part IV. Because these tools are technically if not conceptually simpler for the two-way case, they are first presented here and later on generalized to the multi-way case in Chapters 15 and 18.

Part III consists of five chapters. Measures of resemblance play an important role in many domains of data analysis. However, similarity coefficients often only allow pairwise or bivariate comparison of variables or entities. An alternative to two-way resemblance measures is to formulate multivariate or multi-way coefficients. Before considering multi-way formulations of coefficients for binary data in Part IV, Part III is used to explore and extend some concepts from Chapter 10 and the literature on three-way data analysis to the multi-way case. Part III is devoted to possible generalizations and other related multi-way extensions of the triangle inequality, including the perimeter distance function, the maximum distance function, and multi-way ultrametrics.

Before extending the metric axioms, Chapter 11 is used to formulate more basic axioms for multi-way dissimilarities. Axiom systems for two-way and three-way dissimilarities are studied first. The dependencies between various axioms are reviewed to obtain axiom systems with a minimum number of axioms. The consistency and independence of several axiom systems is established by means of simple models. The remainder of Chapter 11 is used to explore how basic axioms for multi-way dissimilarities, like nonnegativity, minimality and symmetry, may be defined.

Chapter 12 explores how the two-way metric may be generalized to multi-way

metrics. A family of k-way metrics is formulated that generalize the two-way metric and the three-way metrics from the literature. Each inequality that defines a metric is linear in the sense that we have a single, possibly weighted, dissimilarity, which is equal to or smaller than an unweighted sum of dissimilarities. The family of inequalities gives an indication of the many possible extensions for introducing k-way metricity. It is shown how k-way metrics and k-way dissimilarities are related to their $(k-1)$-way counterparts.

Multi-way ultrametrics are explored in Chapter 13. In the literature two generalizations of the ultrametric inequality have been proposed for the three-way case. Continuing this line of reasoning three inequalities may be formulated for the four-way case. For the multi-way case $k-1$ inequalities may be defined. Some ideas on the three-way ultrametrics presented in the literature are explored in this chapter for multi-way dissimilarities. The multi-way ultrametrics as defined in this chapter imply a particular class of multi-way metrics.

In Chapter 14 it is explored how two particular three-way distance functions may be formulated for the multi-way case. The chapter is mostly about extensions of the three-way perimeter model. One section covers the maximum function, its multi-way extension, and a metric property of the generalization. The chapter contains both results on decompositions and on metric properties of two multi-way perimeter models. Chapter 15 is completely devoted to two generalizations of a particular theorem from Chapter 10. This result states that if d satisfies the triangle inequality, then so does the function $d/(c+d)$, where c is a positive real value. The result is extended to one family of multi-way metrics. An attempt is made to generalize the result to a class of stronger multi-way metrics.

Part IV consists of four chapters. In this final part, multivariate formulations of similarity coefficients are considered. Multivariate coefficients may for example be used if one wants to determine the degree of agreement of three or more raters in psychological assessment, if one wants to know how similar the partitions obtained from three different cluster algorithms are, or if one is interested in the degree of similarity of three or more areas where certain types of animals may or not may be encountered.

In Chapter 16 and 17 multivariate formulations (for groups of objects of size k) of various bivariate similarity coefficients (for pairs of objects) for binary data are presented. The multivariate coefficients in Chapter 16 are not functions of the bivariate similarity coefficients themselves. Instead, an attempt is made to present multivariate coefficients that reflect certain basic characteristics of, and have a similar interpretation as, their bivariate versions. The multivariate measures presented in Chapter 17 preserve the relations between various coefficients that were derived in Chapter 4 on correction for chance agreement. This chapter is also used to show how the multi-way formulations from the two chapters are related. In Chapter 18 metric properties of various multivariate coefficients with respect to the strong polyhedral generalization of the triangle inequality are studied. Finally, the Robinson matrices studied in Chapter 7 are extended to Robinson cubes in Chapter 19.