Feature network models for proximity data : statistical inference, model selection, network representations and links with related models

Frank, L.E.

Chapter 1

Introducing Feature Network Models

Feature Network Models (FNM) are graphical models that represent dissimilarity data in a discrete space with the use of features. Features are used to construct a distance measure that approximates observed dissimilarity values as closely as possible. Figure 1.1 shows a feature network representation of all presidents of the United States of America, based on 14 features, which are binary variables that indicate

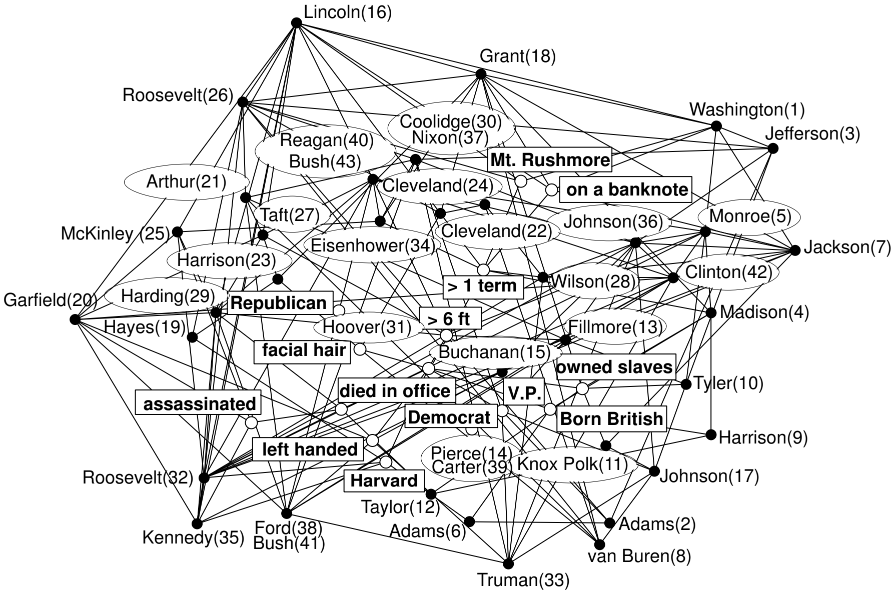

Figure 1.1: Feature network of all presidents of the USA based on 14 features from Schott (2003, pp. 14-15). The presidents are represented as vertices (black dots) and labeled with their names and chronological number. The features are represented as internal nodes (white dots).

whether a president has the characteristic or not (the characteristics were adapted from Schott, 2003, pp. 14-15). The features are: political party, whether the president served more than 1 term, was assassinated or died in office, was taller than 6 ft, served as vice-president (V.P.), had facial hair, owned slaves, was born British, appeared on a banknote, is represented at Mount Rushmore, went to Harvard and is left handed.

In the network the presidents are represented as vertices (black dots) labeled with their name and chronological number. The features are represented as internal nodes (white dots). The general idea of a network representation is that an edge between objects gives an indication of the relation between the objects. For example, there is an edge present between president Kennedy and the feature *assassinated*, which in turn has a direct link with the feature *died in office*. As a result of the embedding of this 14-dimensional structure in 2-dimensional space, objects that are close to each other in terms of distances are more related to each other than objects that are further apart. The network representation has a rather complex structure with a large number of edges and is not easily interpretable. One of the objectives of Feature Network Models is to obtain a parsimonious network graph that adequately represents the data. The three components of FNM, the features, the network representation and the model, will be explained successively in this introduction, using a small data set with16 objects and 8 features, that can be adequately represented in 2 dimensions. While explaining the different components, the topics of the chapters of this monograph will be introduced.

1.1 Features

A feature is, in a dictionary sense, a prominent characteristic of a person or an object. In the context of FNM, a feature is a binary (0,1) vector that indicates for each object or stimulus in an experimental design whether a particular characteristic is present or absent. Features are not restricted to nominal variables, like eye color, or binary variables as voiced versus unvoiced consonants. Ordinal and interval variables, if categorized, can be transformed into a set of binary vectors (features) using dummy coding. Table 1.1 shows an example of features deriving from an experimental design created by Tversky and Gati (1982). The stimuli are 16 types of plants that vary depending on the combination of two qualitative variables, the form of the ceramic pot (4 types) and the elongation of the leaves of the plants (4 types), see Figure 1.2. The two variables can be represented as features using dummy coding for the levels of each variable and Table 1.1 shows the resulting feature matrix. In the original experiment, all possible pairs of stimuli were presented to 29 subjects who were asked to rate the dissimilarity between each pair of stimuli on a 20-point scale. The data used for the analyses are the average dissimilarity values over the 29 subjects as presented in Gati and Tversky (1982, Table 1, p. 333).

In psychology, the concept of feature as basis for a model has been introduced by Tversky (1977) who proposed the Contrast Model, which is a set-theoretical approach where objects are characterized by subsets of discrete features and similarity between objects is described as a comparison of features. The Contrast Model

Table 1.1: Feature matrix of 16 plants (Figure 1.2) varying in form of the pot (features: a, b, c) and elongation of the leaves (features: p, q, r), see Tversky and Gati (1982).

		Features					
Plants		a	b	c	p	q	r
1	ap	1	0	0	1	0	0
2	aq	1	0	0	0	1	0
3	ar	1	0	0	0	0	1
4	as	1	0	0	0	0	0
5	bp	0	1	0	1	0	0
6	bq	0	1	0	0	1	0
7	br	0	1	0	0	0	1
8	bs	0	1	0	0	0	0
9	cp	0	0	1	1	0	0
10	cq	0	0	1	0	1	0
11	cr	0	0	1	0	0	1
12	cs	0	0	1	0	0	0
13	dp	0	0	0	1	0	0
14	dq	0	0	0	0	1	0
15	dr	0	0	0	0	0	1
16	ds	0	0	0	0	0	0

was intended as an alternative to the dimensional and metric methods like multidimensional scaling, because Tversky questioned the assumptions that objects can be adequately represented as points in some coordinate space and that dissimilarity behaves like a metric distance function. Believing that it is more appropriate to represent stimuli in terms of many qualitative features than in terms of a few quantitative dimensions, Tverksy proposed a set-theoretical approach where objects are characterized by subsets of discrete features and similarity between objects is described as a comparison of features. According to Tversky, the representation of an object as a collection of features parallels the mental process of participants faced with a comparison task: participants extract and compile from their data base of features a limited list of relevant features on the basis of which they perform the required task by feature matching. This might lead to a psychologically more meaningful model since it is testing some possible underlying processes of similarity judgments.

Distinctive features versus common features

The Contrast Model describes the similarity between two objects in terms of a linear combination of the features they share (the *common features*) and the features that distinguish between them (*distinctive features*). The idea is that the similarity between two objects increases with addition of common features and/or deletion of distinctive features. In set-theoretical terms, a *common feature* is equal to the intersection of the feature sets that belong to each pair of objects and a *distinctive feature* is equal to the union minus the intersection of the feature sets (= the symmetric set difference).

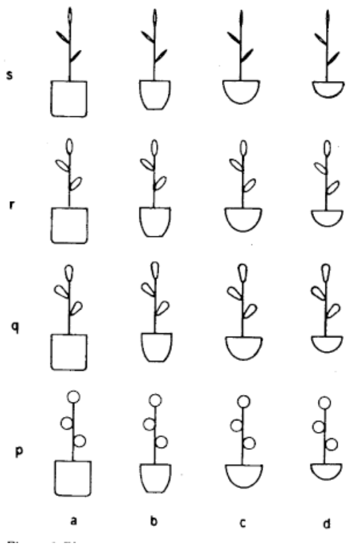

Figure 1.2: Experimental conditions *plants* data. The 16 plants vary in the form of the pot and in elongation of the leaves. (Adapted with permission from: Tversky and Gati (1982), Similarity, separability, and the triangle inequality. *Psychological Review, 89*, 123-154, published by APA.)

For example plant 1 (Table 1.1) is characterized by the feature set $\mathcal{S}_1 = \{a, p\}$ and plant 2 is characterized by the feature set $\mathcal{S}_2 = \{a, q\}$. The plants 1 en 2 have one common feature $\{a\}$ and two distinctive features $\{p, q\}$. The mathematical representation of the similarity between the plants 1 and 2 following the Contrast Model is equal to:

$$\varsigma(\mathcal{S}_1, \mathcal{S}_2) = \theta f(\mathcal{S}_1 \cap \mathcal{S}_2) - \alpha f(\mathcal{S}_1 - \mathcal{S}_2) - \beta f(\mathcal{S}_2 - \mathcal{S}_1), \tag{1.1}$$

where the first part after the equal sign represents a function f of the common features part with the corresponding weight θ and the remaining two parts express functions of the distinctive features part with corresponding weights α and β, and the total similarity value is expressed as a linear combination of the common features part and the distinctive features part. Tversky (1977) and Gati and Tversky (1984) observed that the relative weight of distinctive features and common features varies with the nature of the task: in conceptual comparisons, the relative weight of common to distinctive features was higher in judgments of similarity than in judgments of dissimilarity. The relative weight of common to distinctive features also changed depending on the task instructions: when subjects were instructed to rate the amount in which to objects differ, the relative weight of the distinctive features to common features increases.

Table 1.2: Overview of graphical and non-graphical models based on common features (CF) and distinctive features (DF)

Model	Author(s)	CF, DF	Graphical representation
Contrast Model (CM)	Tversky (1977)	CF + DF	not available
Additive similarity trees	Sattath and Tversky (1977)	DF	additive tree
ADCLUS	Shepard and Arabie (1979)	CF	clusters with contour lines
MAPCLUS	Arabie and Carroll (1980)	CF	clusters with contour lines
EXTREE	Corter and Tversky (1986)	DF	additive tree + marked segments
CLUSTREES	Carroll and Corter (1995)	CF	trees like MAPCLUS and EXTREE
Feature Network Models (FNM)	Heiser (1998)	DF	network (trees)
Modified Contrast Model (MCM)	Navarro and Lee (2004)	CF + DF	clusters

The Contrast Model in its most general form has been used in practice with a priori features only (Gati & Tversky, 1984; Keren & Baggen, 1981; Takane & Sergent, 1983), but many models have been developed since, which search for either the common features part or the distinctive features part of the model, or a combination of both. The models that are based uniquely on common features, the *common features models* are several versions of additive clustering: ADCLUS (Shepard & Arabie, 1979), MAPCLUS (Arabie & Carroll, 1980) and CLUSTREES (Carroll & Corter, 1995). It should be noted that the CLUSTREES model differs from the other common features models because it finds distinctive feature representations of common features models. The additive similarity trees (Sattath & Tversky, 1977)) and the extended similarity trees (EXTREE, Corter & Tversky, 1986) both use distinctive features and are *distinctive features models*. A model that has the closest relation to the Contrast Model is the Modified Contrast Model developed by Navarro and Lee (2004) that aims at finding a set of both common and distinctive a priori unknown features that best describes the data. Table 1.2 gives an overview of the models with the corresponding graphical representation, which will be explained in Section 1.2.

Feature Network Models (FNM) use the set-theoretical approach proposed by Tversky, but are restricted to distinctive features. The definition of distinctive features used in FNM states that features are not inherently distinctive, but become distinctive after application of the set-theoretic transformation, in this case, the symmetric set difference. This definition means that it is not possible to classify, for example, the features describing the plants in Table 1.1 as distinctive or common because the set-theoretic transformations have not taken place yet. Chapter 4 makes the definition of distinctive feature more concrete by defining the complete set of distinctive features and by showing how to generate the complete set in an efficient way using a special binary code, the Gray code.

Although the two types of feature models, the common features model (CF) and the distinctive features model (DF), are in a sense opposed to each other and can function as separate models, there is a clear relation between the two. Sattath and Tversky (1987), and later Carroll and Corter (1995), have demonstrated that the CF model can be translated into the DF model and vice versa. However, these theoretical results have not been applied in the practice of data analysis, where one fits either one of the two models, or the combination of both. Chapter 5 adds an important result to the theoretical translation between the CF model and the DF model, and shows the consequences for the practice of data analysis. It will become clear that for any fitted CF model it is possible to find an equally well fitting DF model with the same shared features (common features) and feature weights, and with the same number of independent parameters. Following the same results, a model that combines the CF and DF models can be expressed as a combination of two separate DF models.

Where do features come from?

The features in Table 1.1 are a direct result of the experimental design and represent the physical characteristics of the objects (the plants). Most of the feature methods mentioned in the previous sections use a priori features that derive from the experimental design or a psychological theory. In the literature, examples where features are estimated from the data are rare. There is, however, no necessary relation between the physical characteristics that are used to specify the objects and the psychological attributes that subjects might use when they perceive the objects. It is therefore useful to estimate the features from the data as well. An example of a data analysis with theoretic features and with features estimated from the data will be given for the *plants* data. Chapter 4 is entirely devoted to the subject of selecting adequate subsets of features resulting from theory or estimated from the data.

It should be noted that a well known set of theoretic features plays an important role in phonetics as part of the Distinctive Feature Theory. The distinctive features form a binary system to uniquely classify the sounds of a language, the phonemes. The term distinctive used here is not the set-theoretic term used for the distinctive features of the FNM. Various sets of distinctive features have been proposed in phonetics and the first set consisting of 14 features has been proposed by Jakobson, Fant, and Halle (1965): the distinctive features are the ultimate distinctive entities of language since none of them can be broken down into smaller linguistic units (p. 3). A subset of these distinctive features will be used to illustrate the Feature Network Models in Chapter 2 and Chapter 4.

Feature distance and feature discriminability

FNM aim at estimating distance measures that approximate observed dissimilarity values as closely as possible. The symmetric set difference can be used as a distance measure between each pair of objects O_i and O_j that are characterized by the corresponding feature sets $\mathcal{S}_i$ and $\mathcal{S}_j$. Following Goodman (1951, 1977) and Restle (1959, 1961), a distance measure that satisfies the metric axioms can be expressed as a simple count μ of the elements of the symmetric set difference, a count of the non common elements between each pair of objects O_i and O_j and becomes the *feature distance*:

$$d(O_i, O_j) = \mu[(\mathcal{S}_i - \mathcal{S}_j) + (\mathcal{S}_j - \mathcal{S}_i)] = \mu[(\mathcal{S}_i \cup \mathcal{S}_j) - (\mathcal{S}_i \cap \mathcal{S}_j)]. \tag{1.2}$$

Heiser (1998) demonstrated that the feature distance in terms of set operations can be re-expressed in terms of coordinates and as such, is equal to a city-block metric on a space with binary coordinates, a metric also known as the *Hamming distance*. If $\mathbf{E}$ is a binary matrix of order $m \times T$ that indicates which of the T features describe the m objects, as in Table 1.1, the re-expression of the feature distance in terms of coordinates is as follows:

$$\begin{aligned} d(O_i, O_j) &= \mu[(\mathcal{S}_i \cup \mathcal{S}_j) - (\mathcal{S}_i \cap \mathcal{S}_j)] \\ &= \sum_t |e_{it} - e_{jt}|, \end{aligned} \tag{1.3}$$

where $e_{it} = 1$ if feature t applies to object i, and $e_{it} = 0$ otherwise. In the example of the plants 1 and 2 the feature distance is equal to the sum of the distinctive features $\{p, q\}$, in this case 2. The properties of the feature distance and especially the relation between the feature distance and the city-block metric are discussed in Chapter 5.

For fitting purposes, it is useful to generalize the distance in Equation 1.3 to a weighted count, i.e., the weighted feature distance:

$$d(O_i, O_j) = \sum_t \eta_t |e_{it} - e_{jt}|, \tag{1.4}$$

where the weights η_t express the relative contribution of each feature. Each feature splits the objects into two classes, and η_t measures how far these classes are apart. For this reason, Heiser (1998) called the feature weight a *discriminability parameter*. The feature discriminability parameters are estimated by minimizing the following least squares loss function:

$$\min_{\boldsymbol{\eta}} = \|\mathbf{X}\boldsymbol{\eta} - \boldsymbol{\delta}\|^2, \tag{1.5}$$

where $\mathbf{X}$ is of size $n \times T$ and $\boldsymbol{\delta}$ is a $n \times 1$ vector of dissimilarities, with n equal to all possible pairs of m objects: $\frac{1}{2}m(m-1)$. The problem in Equation 1.5 is expressed in a more convenient multiple linear regression problem, where the matrix $\mathbf{X}$ is obtained by applying the following transformation on the rows of matrix $\mathbf{E}$ for each pair of objects, where the elements of $\mathbf{X}$ are defined by:

$$x_{lt} = |e_{it} - e_{jt}|, \tag{1.6}$$

where the index $l = 1, \cdots, n$ varies over all pairs (i, j). The result is the binary $(0, 1)$ matrix $\mathbf{X}$, where each row represents the distinctive features for each pair of objects, with 1 meaning that the feature is distinctive for a pair of objects. It is important to notice that features become truly distinctive features only after this transformation, while the features in the matrix $\mathbf{E}$ are not inherently common or distinctive. The weighted sum of these distinctive features is the feature distance for each pair of objects and is equal to $\mathbf{d} = \mathbf{X}\boldsymbol{\eta}$. The feature distances serve as starting point for the construction of the network, as will become clear in the next section.

1.2 Feature Network

In general, there are two types of graphical representations of proximity data: spatial models and network models. The spatial models - multidimensional scaling - represent each object as a point in a coordinate space (usually Euclidean space) in such a way that the metric distances between the points approximate the observed proximities between the objects as closely as possible. In network models, the objects are represented as vertices in a connected graph, so that the spatial distances along the edges between the vertices in the graph approximate the observed proximities among the objects. In MDS, the primary objective is to find optimal coordinate values that lead to distances that approximate the observed proximities between the objects, whereas in network models, the primary objective is to find the correct set of *relations* between the objects that describe the observed proximities.

Parsimonious feature graphs

The symmetric set difference, which is the basis of FNM, describes the relations between the object pairs in terms of distinctive features and permits a representation of the stimuli as vertices in a network using the feature distance. In the network, called a *feature graph*, the structural relations between the objects in terms of distinctive features is expressed by edges connecting adjacent objects and the way in which the objects are connected depends on the fitted feature distances. Distance in a network is the path travelled along the edges; the distance that best approximates the dissimilarity value between two objects is the shortest path between the two corresponding vertices in the network.

The feature distance has some special properties resulting from its set-theoretical basis that allows for a representation in terms of shortest paths, which also considerably reduces the number of edges in the network. A complete network, i.e. a network where all pairs of vertices (representing the m objects) are connected has $n = \frac{1}{2}m(m-1)$ edges. Figure 1.3 shows a complete network of the *plants* data where all pairs of plants are connected with an edge. Such a network is obviously not adequate in explaining the relations between the objects, due to lack of parsimony.

The feature distance parallels the path-length distance in a valued graph when one of the metric axioms, the triangle inequality, becomes an equality: $d_{ik} = d_{ij} + d_{jk}$ (Flament, 1963; Heiser, 1998). In a network graph, each time that the distance d_{ik} is exactly equal to the sum $d_{ij} + d_{jk}$ the edge between the objects i and k can be

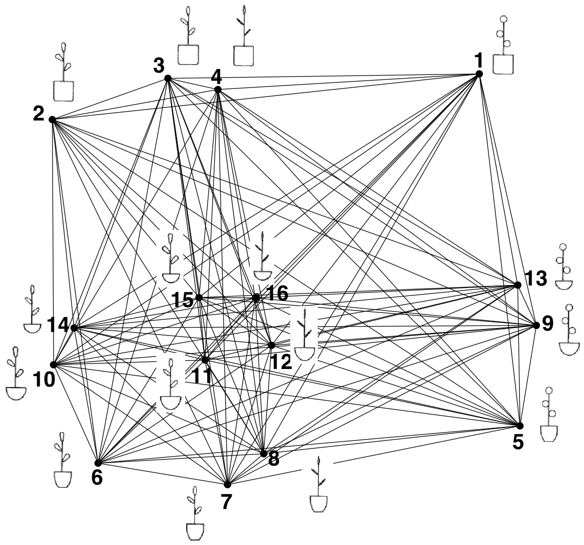

Figure 1.3: Complete network *plants* data.

excluded, resulting in a parsimonious subgraph of the complete graph. In terms of features the condition $d_{ik} = d_{ij} + d_{jk}$ is reached when object j is *between* objects i and k. The objects can be viewed as sets of features: $\mathcal{S}_i$, $\mathcal{S}_j$, and $\mathcal{S}_k$. Betweenness of $\mathcal{S}_j$ depends on the following conditions (Restle, 1959):

1. $\mathcal{S}_i$ and $\mathcal{S}_k$ have no common members which are not also in $\mathcal{S}_j$;

2. $\mathcal{S}_j$ has no unique members which are in neither $\mathcal{S}_i$ nor $\mathcal{S}_k$.

Apart from the experimental objects, we can also identify hypothetical objects called *internal nodes*. These are new feature sets defined in terms of the intersection of available feature sets. As an example, Figure 1.4 shows two of the plants (numbers 13 and 14, with feature sets $\{d, p\}$ and $\{d, q\}$, respectively) and an internal node defined by a feature set containing the single feature $\{d\}$. It is clear that betweenness holds with respect to the internal node, because its feature set is exactly equal to the intersection of the sets belonging to plants 13 and 14, as can be seen in the right part of Figure 1.4. For the network representation in terms of edges, the betweenness condition implies that the feature distances between the three objects reach the triangle equality condition. Calling the internal node *'dpot'*, we have $d_{14,13} = d_{14,dpot} + d_{dpot,13} = 1 + 1 = 2$. For the ease of explanation the feature distances are represented as unweighted counts of the number of distinctive features. Consequently, the edge between the plants 13 and 14 can be excluded (see the left part of Figure 1.4).

Hence, sorting out the additivities in the fitted feature distances, with the possible inclusion on internal nodes, and excluding edges that are sums of other edges

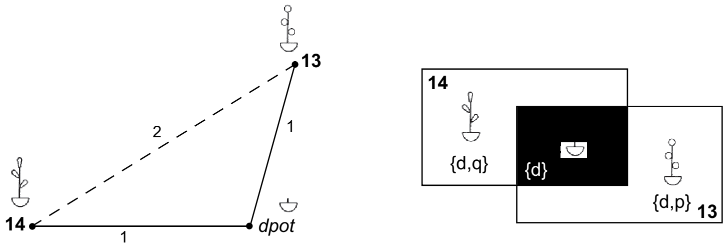

Figure 1.4: Triangle equality and betweenness.

results in a parsimonious subgraph of the complete graph, expressed in a binary adjacency matrix with ones indicating the presence of an edge. It should be noted that the approach of sorting out the additivities is different from the network models of Klauer (1989, 1994) and Klauer and Carroll (1989), who sort out the additivities on the observed dissimilarities. Using the fitted distances instead leads to better networks because the distances are model quantities whereas dissimilarities are subject to error. The network approach used in FNM is also different from the social network models (*cf.* Wasserman & Faust, 1994) that use the adjacency matrix as the starting point of the analyses, whereas for the FNM it is the endpoint.

Figure 1.5 shows the result of sorting out the triangle equalities on the fitted feature distances for the *plants* data, where features d and s have been omitted. Note that each of the first four features a, b, c, and d is redundant, since it is a linear combination of the other three. The same is true for p, q, r, and s. To avoid problems with multicollinearity in the estimation, one feature in each set has to be dropped. Hence from now on, we continue the example with a reduced set of 6 features. The *feature graph* clearly has gained in parsimony compared to the complete network in Figure 1.3. The network has been embedded in 2-dimensional Euclidean space, with the help of PROXSCAL (a multidimensional scaling program distributed as part of the Categories package by SPSS, Meulman & Heiser, 1999), allowing ratio proximity transformation. The edges in the network express the relations between pairs of plants based on the weighted sum of their distinctive features. More details on the interpretation of the network model will be given in section 1.3. Chapter 5 discusses in detail the betweenness condition as well as the algorithm used for sorting out the triangle equalities and also introduces the internal node as a way to simplify the network representation.

Embedding in low-dimensional space

A network is by definition coordinate-free because it is entirely determined by the presence or absence of edges between vertices and by the lengths of these edges. The distances between the objects in a network do not serve the same interpretational purpose as in multidimensional scaling, where distances are a direct expression of the strength of the relation between the objects. In FNM the relation between the

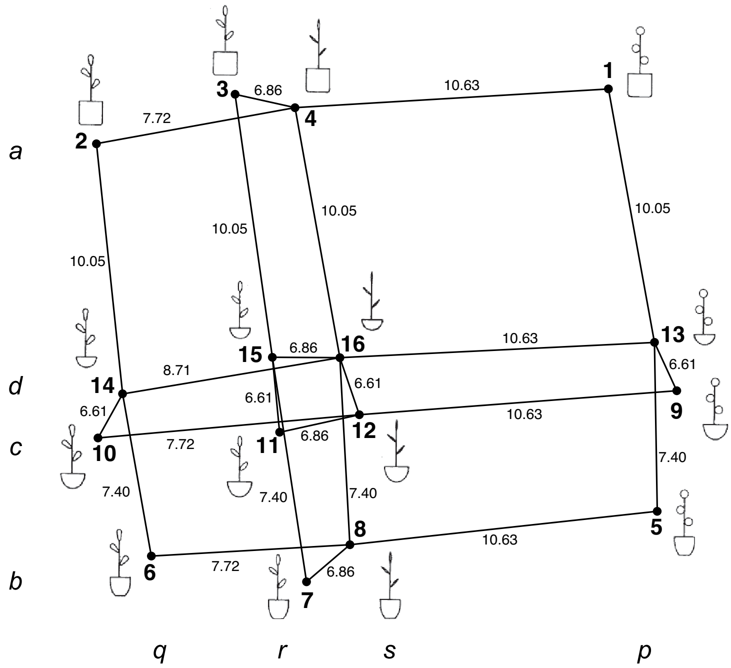

Figure 1.5: Feature graph of the *plants* data using the features resulting from the experimental design with varying elongation of leaves and form of the pot (with 6 of the 8 features).

objects is primarily expressed by the presence of edges between the vertices. The embedding of the network in a lower dimensional space is therefore of secondary importance. In this monograph, the embedding chosen for the feature graphs results from analysis with PROXSCAL (Meulman & Heiser, 1999) of the Euclidean distances computed on the weighted feature matrix. Most of the representations are in 2 dimensions, sometimes other options are chosen to obtain a representation that is better in terms of visual interpretability.

The embedding discussed so far concerns the vertices of the network representing the objects, while the features themselves can also be visualized. There are several possibilities to represent features graphically in the network plot. Most of the network plots in this monograph show the features as vectors. This representation is obtained using PROXSCAL (Meulman & Heiser, 1999) with the option of restricting the solution of the common space by a linear combination of the feature variables. Another version is to represent the features as internal nodes, as in the presidents network (Figure 1.1). The representation of features as internal nodes which will be discussed and illustrated in Chapter 5.

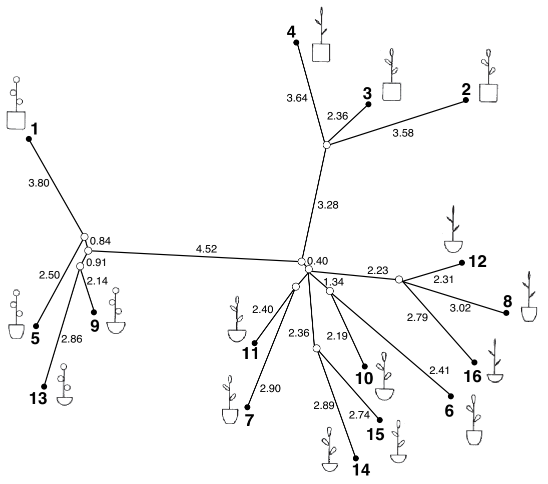

Figure 1.6: Additive tree representation of the *plants* data.

Feature structure and related graphical representation

The feature structure typically represented by FNM is a non-nested structure, or in terms of clusters, an overlapping cluster structure. In contrast, hierarchical trees and additive trees require a strictly nested feature structure. The graphical representation of a non-nested structure is more complex than a tree (Carroll & Corter, 1995). At least three solutions have been proposed in the literature (see the overview in Table 1.2): ADCLUS starts with a cluster representation and adds contour lines around the cluster to reveal the overlapping structure; two other representations start with a basic tree and visualize the overlapping structure by multiple trees (Carroll & Corter, 1995; Carroll & Pruzansky, 1980) or by extended trees (Corter & Tversky, 1986). Extended trees represent non-nested feature structures graphically by a generalization of the additive tree. The basic tree represents the nested structure and the non-nested structure is represented by added marked segments that cut across the clusters of the tree structure (Carroll & Corter, 1995, p. 288). The FNM is the only model that represents this overlapping feature structure by a network representation.

Imposing restrictions on the feature structure in FNM allows for other graphical representations than a network. Chapter 3 shows that an additive tree is a special subgraph of the complete feature graph, where each edge is represented by a separate feature. To obtain an additive tree representation as in Figure 1.6, the feature matrix must satisfy certain conditions. To produce a tree graph with FNM a nested set of features is not sufficient. A set of internal nodes (hypothetical stimulus objects) has to be added to the set of actual stimulus objects, or external nodes. As will become clear in Chapter 3, if the internal nodes have the correct feature structure, they will force the betweenness condition to hold in such a way that a tree graph results.

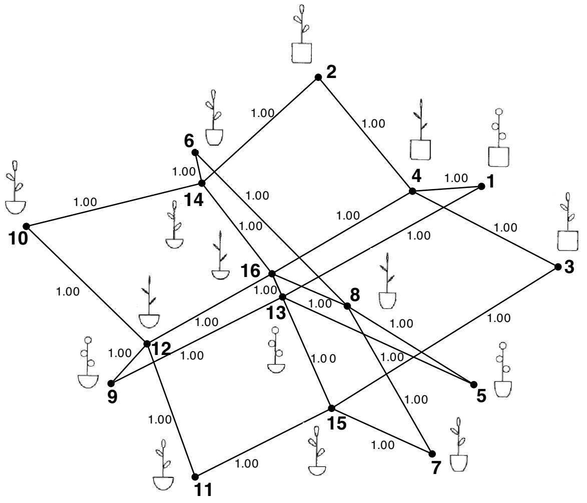

Figure 1.7: Feature network representing a 6-dimensional hypercube based on the unweighted, reduced set of features of the *plants* data. Embedding in 2-dimensional Euclidean space was achieved with PROXSCAL allowing ordinal proximity transformation with ties untied and the Torgerson start option.

Feature networks and the city-block model

The symmetric set difference is a special case of the city-block metric with binary dimensions represented by the distinctive features. Therefore, the network representation lives in city-block space. The dimensionality of this city-block space is defined by the number of features T forming a T-dimensional rectangular hyperblock, or hypercuboid with the points representing the objects located on the corners. In the special case when the symmetric set difference is equal for adjacent objects in the graph, the structure becomes a hypercube. The feature structure of the *plants* data yields a hypercube structure when all feature discriminability parameters are set equal to one. Figure 1.7 shows the resulting 6-dimensional network structure using the theoretical features and after sorting out the triangle equalities. Using the weighted feature distance transforms the lengths of the edges of the 6-dimensional hypercube into a 6-dimensional hypercuboid as in Figure 1.5. (The visual comparison between the network representations in the Figures 1.5 and 1.7 requires some effort because due to the embedding in lower dimensional space, the emplacement of the plants has changed.)

Chapter 5 demonstrates that there exists a universal network representation of city-block models. The results rely on the additivity properties of the city-block distances and the key elements of the network representation consisting of betweenness, metric segment additivity and internal nodes. The universal network construction rule also applies to other models beside the distinctive features model, namely the common features model (additive clustering), hierarchical trees, additive trees and extended trees.

1.3 Feature Network Models: estimation and inference

Figure 1.8 shows an overview of the steps necessary to fit a Feature Network Model on data using the program PROXGRAPH that has been developed in Matlab. Starting with observed dissimilarities and a set of features, the feature discriminability parameters are estimated as well as the feature distances. The estimated feature distances lead to an adjacency matrix after being processed by the triangle equality algorithm and to coordinates obtained with PROXSCAL (Meulman & Heiser, 1999), leading to the final result, a feature network. As explained so far, the network representation of the dissimilarity data provides a convenient way to describe and display the relations between the objects. At the same time the network representation suggests a psychological model that relates mental representation to perceived dissimilarity. The psychological model is not testable with the graphical representation only. In FNM the psychological model can be tested by assessing which feature(s) contributed more than others to the approximation of the dissimilarity values. The statistical inference theory proposed in this monograph derives from the multiple regression framework, as will become clearer in the following. An important topic of this monograph is the estimation of standard errors for the feature discriminability parameters in order to construct 95% confidence intervals . Another way to decide which features are important is to use model selection techniques to obtain a relevant subset of features.

Statistical inference

The use of features, when considered as prediction variables, leads in a natural way to the univariate multiple regression model, which forms the starting point for statistical inference. It is however not the standard regression model because positivity restrictions have to be imposed on the parameters. The feature discriminability parameters represent edge lengths in a network or a tree and, by definition, networks or trees with negative edge lengths have no meaning and cannot adequately represent a psychological theory. The problem becomes more prominent in additive tree representations because each edge is represented by a separate feature. Therefore, the feature discriminability parameters are estimated by adding the following positivity constraint to Equation 1.5:

$$\min_{\boldsymbol{\eta}} = \|\boldsymbol{\delta} - \mathbf{X}\boldsymbol{\eta}\|^2 \qquad \text{subject to } \boldsymbol{\eta} \geq 0. \tag{1.7}$$

The multiple regression approach has been used earlier in models related to FNM. The Contrast Model (Takane & Sergent, 1983), the common features model (Arabie & Carroll, 1980), and the tree models (Corter, 1996) use ordinary least squares to estimate the parameters of the models. The use of nonnegative least squares is sparse in the literature of tree models. Arabie and Carroll (1980) implemented a subroutine in the MAPCLUS algorithm that encourages the weights to become positive,

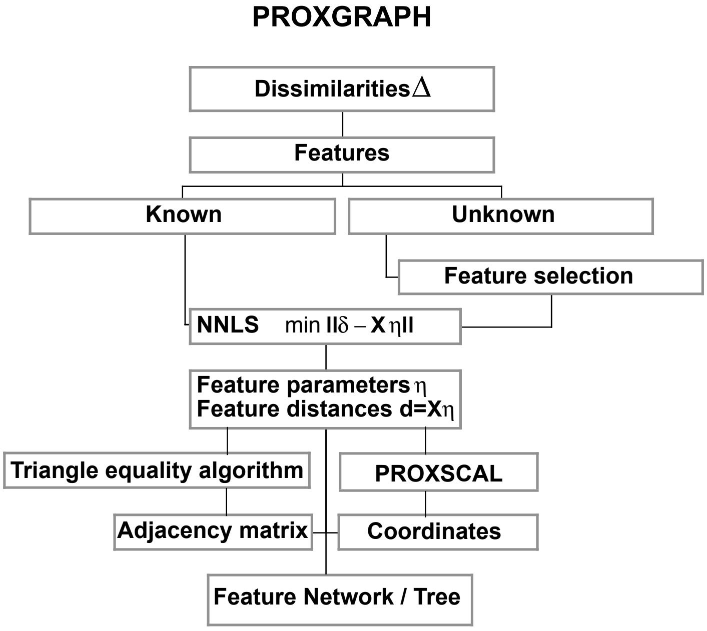

Figure 1.8: An overview of the steps necessary to fit Feature Network Models with PROXGRAPH.

but claim to explicitly avoid the use of nonnegative least squares because in the context of iterative algorithm it would reduce the numbers of clusters in the solution. Hubert, Arabie and Meulman (2001) have successfully implemented nonnegative least squares in their algorithm for the estimation of the edge lengths in additive trees and ultrametric trees. In the domain of phylogenetic trees, nonnegative least squares has been introduced by Gascuel and Levy (1996).

While nonnegative least squares has been used to estimate the parameters in models related to FNM, there are no examples in the literature of the estimation of theoretical standard errors, with or without nonnegativity constraints. The models related to FNM, the extended tree models (Corter & Tversky, 1986), the CLUSTREE models (Carroll & Corter, 1995) and, the Modified Contrast Model (Navarro and Lee, 2004) do not explicitly provide a way to test for significance of the features. It should be noted, however, that Corter and Tversky (1986) use a descriptive version of the *F*-test by permuting residuals to test the significance of the overlapping features added to the additive tree solutions. The other network models (Klauer, 1989, 1994; Klauer & Carroll, 1989) only give the best fitting network and yield no standard errors for the parameter estimates. Krackhardt (1988) provided a way to test the significance of

Table 1.3: Feature discriminability estimates, standard errors and 95% confidence intervals for *plants* data using six features selected from the complete experimental design in Table 1.1 and associated with the network graph in Figure 1.5 ($R^2 = 0.60$).

Features	$\hat{\eta}$	$\hat{\sigma}_{\hat{\eta}}$	95% t-CI	
a	6.29	0.57	5.15	7.42
b	3.64	0.57	2.50	4.77
c	2.85	0.57	1.71	3.98
d	0.00*	0.00	0.00	0.00
p	6.86	0.57	5.73	8.00
q	3.95	0.57	2.82	5.08
r	3.09	0.57	1.96	4.22
s	0.00*	0.00	0.00	0.00

* To avoid multicollinearity, the fourth level of the flowerpots (d) and the plants (s) has been omitted.

regression coefficients in networks for dyadic data that suffer from various degrees of autocorrelation by using quadratic assignment procedures. Unfortunately, his results do not apply to FNM because of the presence of constraints on the feature parameters.

Statistical inference in inequality constrained least squares problems is far from straightforward. A recent review by Sen and Silvapulle (2002) showed that topics on statistical inference problems when the associated parameters are subject to possible inequality constraints abound in the literature, but solutions are sparse. In the context of the inequality constrained least squares problem, only one author (Liew, 1976) has produced a way to compute theoretical standard errors for the parameter estimates. Liew (1976), however, did not evaluate the sampling properties of the theoretical standard errors. Chapters 2 of this monograph shows an application of the theoretical standard errors and associated 95% confidence intervals for feature networks with a priori known features. The performance of the theoretical standard errors is compared to empirical standard errors using Monte Carlo simulation techniques. Chapter 3 evaluates the performance of the theoretical standard errors for features structures in additive trees and the results are extended to the case where the feature structure (i.e., the tree topology) is not known in advance.

Table 1.3 shows the feature discriminability parameters and the associated theoretical standard errors and 95% t-confidence intervals for the theoretic features of the *plants* data. To avoid multicollinearity, the feature discriminability parameters and the associated standard errors have been estimated with a smaller feature set, than the set of theoretical features presented in Table 1.1. Two features have been omitted, namely the fourth level of the flowerpots (feature d) and the fourth level of plants (feature s). As a result these two features have zero values in Table 1.3. The overall fit of the model is reasonable with an R^2 equal to 0.60. The estimates of the feature discriminability parameters indicate that the features a, representing the square formed pot, and p, representing the round shaped leaves, are the most important in

Table 1.4: Feature matrix resulting from feature subset selection with the Positive Lasso on the *plants* data.

		Features					
Plants		F_1	F_2	F_3	F_4	F_5	F_6
1	ap	1	1	0	1	1	1
2	aq	1	1	1	0	1	1
3	ar	1	1	1	0	0	1
4	as	1	1	1	0	0	0
5	bp	0	1	0	1	1	1
6	bq	0	1	1	0	1	1
7	br	0	1	1	0	0	1
8	bs	0	1	1	0	0	0
9	cp	0	0	0	1	1	1
10	cq	0	0	1	0	1	1
11	cr	0	0	1	0	0	1
12	cs	0	0	1	0	0	0
13	dp	0	0	0	1	1	1
14	dq	0	0	0	0	1	1
15	dr	0	0	0	0	0	1
16	ds	0	0	0	0	0	0

distinguishing the plants. The network representation in Figure 1.5 reflects the importance of the features a and p by larger distances between plants that possess these features and plants that do not possess these features. The edges in the network (Figure 1.5) are labeled with the feature distances, which can be reconstructed from the feature discriminability parameters in Table 1.3. For example, the distance between the plants 2 and 4 is equal to 3.95, which is the sum of the feature discriminability parameters corresponding to their distinctive features: $q(= 3.95) + s(= 0.00)$.

Finding predictive subsets of features

In many research settings the features are not known a priori and the main objective is to find a relevant set of features that explain the dissimilarities between the objects as accurately as possible. Chapter 4 proposes a method to find adequate sets of features that is closely related to the predictor selection problem in the multiple regression framework. The basic idea is to generate a very large set of features (or, if possible, the complete set of features) using Gray codes. Since features are binary variables, they can efficiently be generated with binary coding. Next, a subset of features is selected with the Lasso option of the Least Angle Regression (LARS) algorithm (Efron, Hastie, Johnstone, & Tibshirani, 2004), a recently developed efficient model selection algorithm that is less greedy than the traditional forward selection methods used in the multiple linear regression context. To meet the positivity constraints necessary in FNM, the Lasso has been modified into a Positive Lasso. The resulting strategy incorporates model selection criteria during the search process, leading to a set of features that is not necessarily optimal in the current data, but that constitutes a good compromise between model fit and model complexity. This approach of finding a balanced trade-off between goodness-of-fit and prediction accuracy has not been used in the psychometric models related to FNM, except for the independently developed Modified Contrast Model (Navarro & Lee, 2004) that uses a forward feature selection method and a model selection criterion related to the BIC criterion.

Table 1.4 displays the results of the Positive Lasso subset selection method on the *plants* data. The 6 selected features differ in several aspects from the theoretical features derived from the experimental design. Only the two most important features from the experimental design, features a and p were selected by the Positive Lasso (the features F_1 and F_4) in Table 1.4. Figure 1.9 represents the corresponding feature graph, which is clearly different from the feature graph based on the theoretic features (Figure 1.5): it is more parsimonious and has better overall fit ($R^2 = 0.81$). The plants have the same order in the network as in the experimental design and form a grid where each edge represents exactly one feature. For example plant number 6 and plant number 2 are connected with an edge representing the square shaped pot.

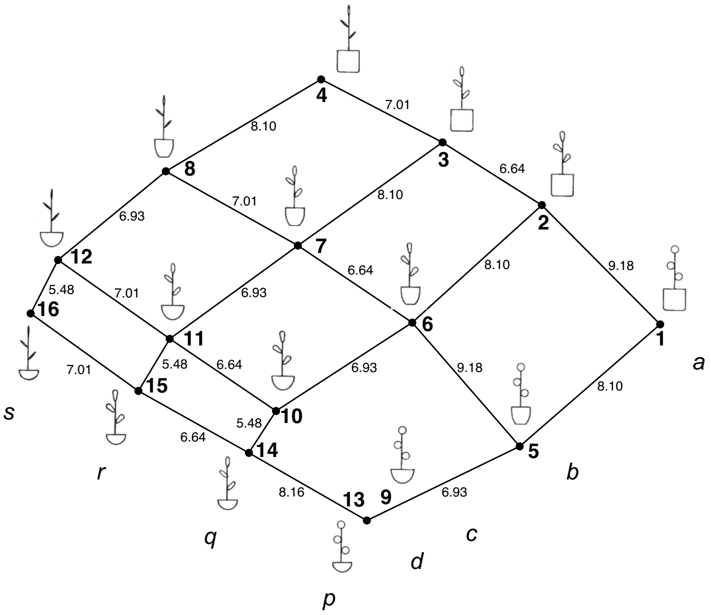

Figure 1.9: Feature graph for the *plants* data, resulting from the Positive Lasso feature subset selection algorithm on the complete set of distinctive features. The original experimental design is the cross classification of the form of the pot (a,b,c,d) and the elongation of the leaves (p,q,r,s). Embedding in 2-dimensional space was done with PROXSCAL using ratio transformation and the simplex start option. ($R^2 = 0.81$)

The edge lengths show that the pots of the form c and d are perceived as more similar (the plant numbers 9 and 13 even coincide on the same vertex in the network) than the pots with form a and b. Moreover, the network representation shows that it can perfectly represent the three types of triples of stimuli mentioned by Tversky and Gati (1982), where the geometric models based on the Euclidean metric fail. The three types of triples are:

1. *Unidimensional triple*: all stimuli differ on 1 dimension, e.g. the plants 1, 5 and 9 representing the combinations (ap, bp, cp) in Figure 1.9;

2. *2-dimensional triple*: all pairs of stimuli differ on both dimensions, e.g. the plants 1, 6 and 11 with feature combinations (ap, bq, cr);

3. *Corner triple*: two pairs differ on one dimension and one pair differs on 2 dimensions, e.g. the plants 1, 5 and 6, having feature combinations (ap, bp, bq).

Only the city-block metric or the feature network is able to correctly display the relations between these three types of triples because the triangle inequality reduces to the triangle equality in all cases. The Euclidean model and other power metrics than the city-block are able to represent unidimensional triples, and into some extent 2-dimensional triples but fail in representing the corner triples. According to the power metrics other than the city-block model, the distance between plants 1 and 6 (differing on two dimensions) is shorter than the sum of the distances between the plant pairs (1,5) and (5,6). The network representation shows that the shortest path between the plants 1 and 6 is the path from 1 to 5 to 6.

1.4 Outline of the monograph

This monograph is organized as follows. Chapter 2 explains how to obtain theoretical standard errors for the constrained feature discriminability parameters in FNM with a priori known features. The performance of the theoretical standard errors is compared to empirical standard errors (resulting from the bootstrap method) using Monte Carlo simulation techniques. Chapter 3 shows that additive trees can be considered as a special case of FNM if the objects are described by features that form a special structure. The statistical inference theory based on the multiple regression framework is further developed by extending the theory from FNM to additive trees, but also by extending the use of theoretical standard errors and associated 95% confidence intervals to a priori unknown feature structures (in this case, tree topologies). Chapter 4 proposes a new method to find predictive subsets of features, especially for the situation where the features are not known in advance. Using the multiple regression framework of the FNM, a new version of Least Angle Regression is developed that restricts the feature discriminability parameters to be nonnegative and is called the Positive Lasso. While the Chapters 2, 3 and 4 all extend the statistical inference properties of the FNM, Chapter 5 is not directly concerned with statistical inference and focuses on the properties of feature graphs. It shows that there exists a universal network representation of city-block models that can be extended to a large class of discrete models for similarity data, including the distinctive features model, the common features model (additive clustering), hierarchical trees, additive trees, and extended trees. Chapter 6 concludes this monograph with a general conclusion and discussion.

Since the chapters 2 - 5 all represent separate papers, a certain amount of overlap is inevitable, especially in the sections describing the Feature Network Models.